Doctoral Thesis

Developing Big Data Analytics for Smart Manufacturing

スマートマニュファクチャリング用ビッグデータアナリティクスの開発

March 2023

Saman Fattahi

Doctoral Thesis

Developing Big Data Analytics for Smart Manufacturing

スマートマニュファクチャリング用ビッグデータアナリティクスの開発

Saman Fattahi

Abstract

Big data means horizontally networked yet independent data systems containing a vast number of structured and unstructured datasets. Statistical and logical computational arrangements (referred to as big data analytics) must be installed to make sense of big data. Like human-cyber-physical systems, digital twins, artificial intelligence, the internet of things, and sustainability, big data and big data analytics are essential elements of the fourth industrial revolution or smart manufacturing. In this study, big data of manufacturing processes are categorized into three main issues: 1) datasets for digital twin; 2) control variable-evaluation variable-centric datasets; 3) graphical datasets. This thesis considers the problem of developing big data and big data analytics for smart manufacturing, focusing on three related issues.

Issue 1: digital twins of manufacturing phenomena are supposed to machine-learn the required knowledge using relevant datasets available in big data. Therefore, a research question is how to preprocess manufacturing phenomena-relevant datasets for using them directly in digital twins.

Issue 2: Big data and analytics require expensive resources and sophisticated computation arrangements. Thus, big data hardly benefits small and medium-sized manufacturing organizations, resulting in "big data inequality." Consequently, a research question is how to eliminate big data inequality.

Issue 3: big data is often visualized using several two-dimensional plots (graphical dataset). These plots are then used to make a decision informally. Consequently, a research question is how to make formal decisions by computing two-dimensional plots, not numerical data.

Thus, this thesis is organized as follows.

Chapter 1 presents this thesis's background, objective, scope, and limitations. It also presents a comprehensive literature review on big data relevant to smart manufacturing.

Chapter 2 describes the proposed big data analytics framework showing all the subsystems. In this chapter, the functional requirements of the subsystems are explained. big data analytics is developed for manufacturing process-relevant decision-making. The proposed analytics consists of five integrated systems: 1) big data preparation system, 2) big data analytics exploration system, 3) data visualization system, 4) data analysis system, and 5) knowledge extraction system. The big data analytics preparation system prepares contents that exhibit the characteristics of digital manufacturing commons.

Chapter 3 deals with Issue 1. A digital twin consists of five modules (input, modeling, simulation, validation, and output modules), and big data must supply datasets for building these modules. This chapter presents a manufacturing phenomenon-related datasets preprocessing method considering the four modules of digital twins (input, modeling, simulation, and validation modules). As an example, the preprocessing of surface roughness-relevant datasets is considered.

Chapter 4 deals with Issue 2. This chapter described the developed big data analytics framework for the control variable-evaluation variable-centric dataset. This system can support user-defined ontology and automatically produces Extensible Markup Language-based datasets. The big data exploration system can extract relevant datasets prepared by the first system. The system uses keywords derived from the names of manufacturing processes, materials, and analyses- or experiments-relevant phrases (e.g., design of experiment). The third system can help visualize relevant datasets extracted by the second system using suitable methods (e.g., scatter plots and possibility distribution). The fourth system establishes relationships among the relevant control variables (variables that can be adjusted as needed) and evaluation variables (variables that measure the performance) combinations for a given situation. In addition, it quantifies the uncertainty in the relationships. The last system can extract knowledge from the

outcomes of the fourth system using user-defined criteria (e.g., minimize surface roughness and maximize material removal rate). The efficacy of the proposed big data analytics is demonstrated using a case study where the goal is to determine the right states of control variables of dry electrical discharge machining for maximizing material removal rate. It is found that the proposed big data analytics is transparent and free from big data inequality.

Chapter 5 deals with Issue 3. Big data analytics is developed to compute two-dimensional plots (graphical datasets) generated from big data. The efficacy of the tool is demonstrated by applying it to assess sustainability in terms of Sustainable Development Goal 12 (responsible consumption and production). Regarding that, engineering materials' functional, economic, and environmental issues play a vital role. Accordingly, three two-dimensional plots generated from big data of engineering materials are computed using the proposed analytics. The plots refer to six criteria (strength, modulus of elasticity, cost, density, CO2 footprint, and water usage). The proposed analytics correctly rank the given materials (mild steel, aluminum alloys, and magnesium alloys).

Chapter 6 describes future research directions and discusses the implication of this study from the viewpoint of smart manufacturing.

Finally, Chapter 7 provides the concluding remarks of this thesis.

This page is kept blank intentionally.

Table of Contents

This page is kept blank intentionally.

Chapter 1: Introduction

Chapter 1 presents this thesis's background, objective, scope, and limitations. It also presents a comprehensive literature review on big data relevant to smart manufacturing.

1.1 Big data and smart manufacturing

Smart manufacturing (or Industry 4.0) [1] embarks on the Human-Cyber-Physical System (HCPS) [2], as shown in Figure 1-1. HCPS consists of Internet of Things (IoT)-based manufacturing enablers, Digital Twins (DT), Big Data (BD), and documentation of past research and operational activities [3].

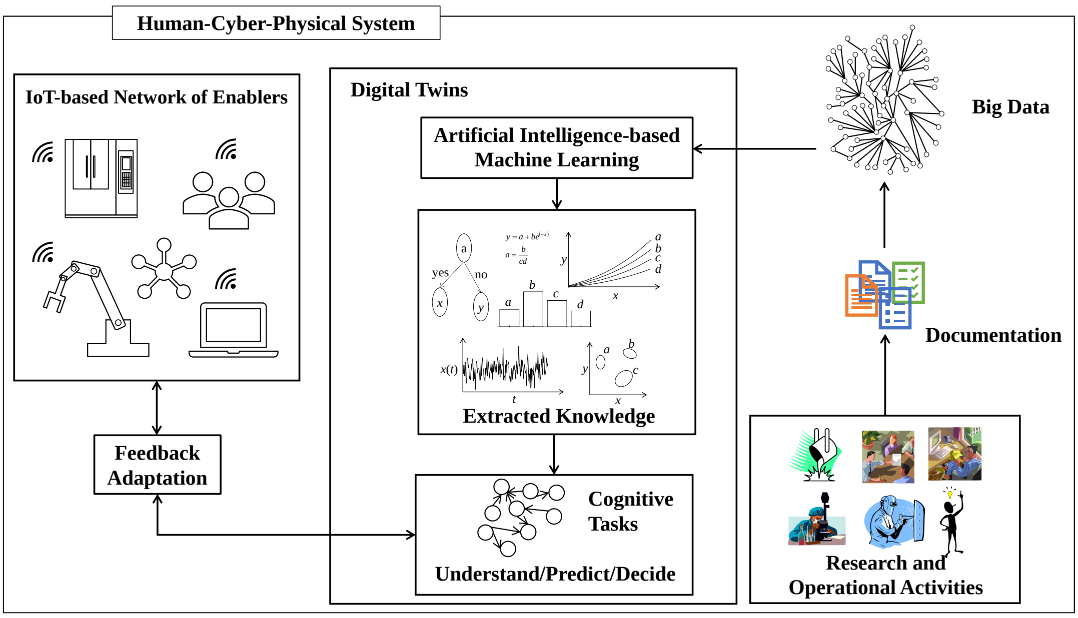


Figure 1-1. Constituents of smart manufacturing

In HCPS, datasets are collected from embedded systems. Moreover, geographically distributed sources play a vital role. This vast array of datasets results in an information silo called big data [3,4]. This silo evolves with time and consists of unstructured, semistructured, and structured datasets [2,3]. The concept of BD was first introduced in the 1990s [6]. This concept enriches many sectors, including healthcare, banking [3], media and entertainment [4], education [5], and transportation [6]. The same argument is valid for the manufacturing sector, as described in Section 2. Specific research shows that approximately 3 Exabytes (EB) of data existed globally in 1986. By 2011, over 300 EB of data were stored in a financial econometric context. Nowadays, BD is characterized by a set of Vs. One of the comprehensive definitions was characterized by 6Vs including volume, velocity, variety, value, veracity, and variability (See Figure 1-2).

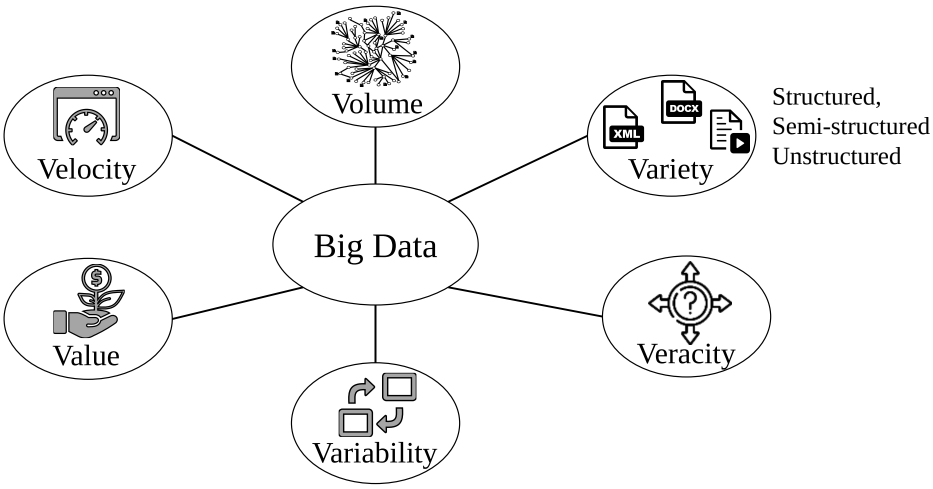


Figure 1-2. Six Vs of BD

Regarding its volume, interestingly, it reached more than 1000 EB annually in 2015, and it is expected that the world will produce and consume 94 zettabytes (94000 EB) of data in 2022 [4-6]. The rapid speed by which datasets are accumulated in BD determines their velocity. The multiplicity in the contents (text, video, and graphics) and structures (structured, unstructured, and semi-structured) means the variety of BD. As far as variety is concerned, traditional BD-relevant database systems effectively manage only the structured datasets. Handling unstructured datasets is still somewhat challenging since those (unstructured datasets) do not fit inside a rigid data model [10]. The accuracy of the datasets determines the veracity of BD. Finally, the economic or social wealth generated from BD is referred to as its value. It is expected that up to the end of 2022, the big data market will grow to $274.3 billion [11].

1.2 Big data analytics

While making decisions using BD, computational arrangements are required. These arrangements are called Big Data Analytics (BDA). The value of BD can be ensured by developing BDA. It (BDA) formally computes the relevant datasets available in BD using statistical or logical approaches [12]. BDA offers a wide range of data visualization facilities in most cases. Users thus rely on the visualized information to make decisions. In addition to visualization facilities, machine learning, and computational intelligence-driven arrangements are often added to BDA. This makes the decision-making process more formal. However, adding these computational arrangements makes the analytics computationally heavy and highly resource-depended. As a result, only large organizations can sustain BDA, and medium and small organizations fall behind. Thus, BDA results in an inequality referred to as BD inequality or digital divide [13]. BD inequality is conceptualized in three dimensions: data access, data representation, and data control inequalities [10-12]. Here, data access-relevant inequality means inaccessibility of the data stored in any data storage infrastructures (e.g., data storage infrastructures of any national or private body) and unavailability of accurate statistics. On the other hand, data representation and control-relevant inequalities mean the lack of infrastructure, human capital, economic resources, and institutional frameworks in developing countries and small firms and organizations, compared to developed countries and big firms and organizations. This generates new dimensions of the digital divide in BDA, knowledge underlying the data, and consequent decision-making abilities. For this purpose, some measures are needed to mitigate BD inequality and the digital divide. For this purpose, it is better to have a BDA system that is not computationally heavy and highly resource-depended. It means that it is better to don’t use the enormous and complex analytical approach in our BDA system, which makes our BDA system a black box where the inputs and outputs are visible, but the inner processes remain unknown [16].

1.3 Big data analytics and smart manufacturing

As it is shown in Figure 1-1, smart manufacturing needs manufacturing enablers such as CAD/CAM systems, process planning systems, CNC machine tools, measuring devices, actuators, robots, and human resources. The difference is that the enablers create an IoT-based network [17], allowing both vertical and horizontal integrations. These enablers get the required knowledge from digital manufacturing commons (DMC). The aspect of DMC was first introduced by Beckmann et al. [18] as a national initiative for US manufacturing innovation. This concept was developed to give access to the tools for manufacturing innovation across small and medium-sized enterprises (SMEs) and large companies, universities, institutes, and entrepreneurs. This way, the digital divide and BD inequality for accessing the data could be mitigated. Accordingly, for developing the BDA framework in our research, we consider the aspect of DMC as the major framework of smart manufacturing. As it is shown in Figure 1-3, these commons create a network that exhibits the characteristics of BD. BD of the manufacturing process in this study are categorized into three main issues: 1) Dataset for DT, 2) CV-EV-centric dataset 3) Graphical dataset. BDA must be incorporated into the smart manufacturing framework to make the connection between IoT-based manufacturing enablers and BD. These commons are usually prepared from heterogenous documentation of past manufacturing activities. The main question is how to prepare DMC and present a comprehensive framework for BDA for smart manufacturing?

In the next part, the literature review, we are going to review the current and past research activities regarding BDA for manufacturing and then consider the aspects of BD inequality and DMC to present a comprehensive scheme for manufacturing processes.

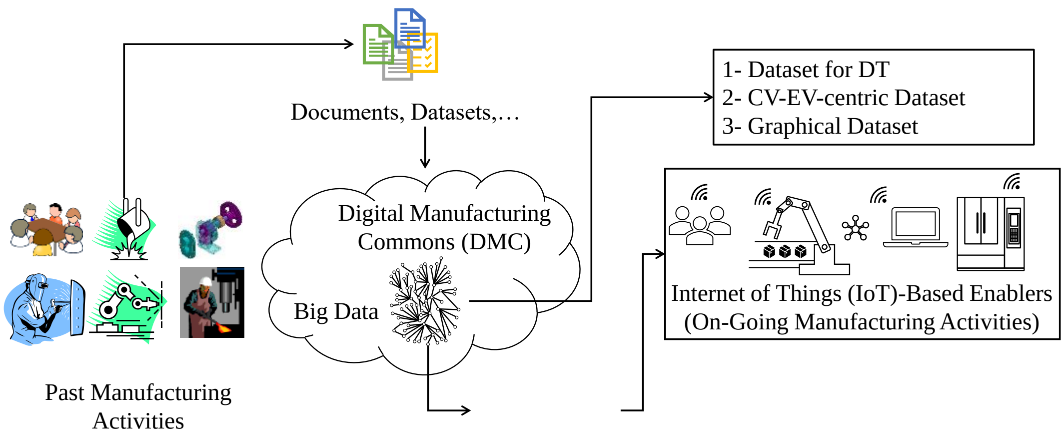


Figure 1-3. Digital manufacturing commons incorporated with BD and BDA in smart manufacturing

1.4 Literature review

This part presents a comprehensive review regarding the main thesis subjects: dataset preparation of manufacturing processes, decision-making using graphical data, and BDA of manufacturing processes.

1.4.1 Developing big data analytics for manufacturing processes

Wang et al. [46] showed that data collected from different elements of manufacturing cyber-physical systems (RFID, sensor, AGV, and alike) results in BD. It (BD) consists of structured, semi-structured, and unstructured datasets. They proposed an analytic to take full advantage of BD. The proposed BDA consists of data collection, storage, cleaning, integration, analysis, mining, and visualization modules. However, they did not articulate the details of the above-mentioned constituents of BDA.

Kahveci et al. [47] presented an end-to-end, IoT-based BDA platform consisting of five interconnected layers and several components for data acquisition, integration, storage, analysis, and visualization. The layers are as follows: control and sensing layer, data collection layer, data integration layer, data storage/analysis layer, and data presentation layer. In addition, they emphasized that data retention policies, as well as the data down-sampling methods, must be incorporated to get benefitted from BDA.

Woo et al. [48] developed a BDA platform for manufacturing systems (Holonic Manufacturing Systems). First, they have identified three implementation challenges: object virtualization, data control, and model control. Next, they introduced analytics that consists of eight modules: process data attribute identification, data acquisition, data pre-processing, context synchronization, training dataset preparation, component model computation, model validation, uncertainty quantification, and model composition and use. As far as machine learning is concerned, the proposed analytics can use the Bayesian network, artificial neural network, or statistical analysis, whatever is appropriate.

Jun et al. [49] presented a cloud-based BDA framework for manufacturing. It uses a user-defined data analytics algorithm template in the form of XML for performing the analysis tasks. The template describes the data characteristics according to a specific manufacturing problem (e.g., failure symptom analysis, RUL prediction, anomaly detection, condition diagnostics, etc.). Based on this, the platform adapts an appropriate algorithm (e.g., similarity-based prognostics, C4.5 rule, ensemble tree, Ada C.X, relative entropy, RMS, peak-to-peak, and alike) and visualization technique (e.g., time series, rule table, prediction table, and alike) for making sense out of the received data.

Lu and Xu [50] presented a BDA-driven cloud-based manufacturing equipment architecture. The manufacturing equipment embedded with sensors, an intelligent adaptive control module, a machine monitoring module, and a data processing module interacts with its DT residing in the cloud. The twin mirrors the real-time status of its physical counterpart. This interaction generates a tremendous amount of data. The analytics analyzes these data stored in a cloud-based repository, unfolds the machine condition, and offers data visualization techniques for enabling on-demand manufacturing services.

Ji and Wang [51] presented a framework for BDA-based fault prediction. The framework considers both real-time and historical data from the shop floor. The data attributes relevant to different manufacturing enablers (e.g., workpieces, machining time, machine tools, machining results, and human factors) are organized in a specified manner. As far as data analytics is concerned, the framework incorporates a data cleansing procedure (e.g., parsing, data transformation, integrity constraint enforcement, duplicate elimination, statistical methods, etc.) and analysis algorithm (e.g., cluster analysis, factor analysis, correlation, and dependence analysis, regression analysis, data mining, A/B testing, and alike) for making sense out of data.

1.4.2 Preparing datasets for digital twin

Moktadir et al. [19] studied the barriers to BD implementation in real-life manufacturing organizations located in a developing economy. They found that collecting reliable datasets from relevant sources is the most significant barrier. The second most significant barrier is related to technology and resource, lack of IT infrastructure, data privacy assurance, complexity in data integration, lack of appropriate BDA, and high investment. As far as the developed economy is concerned, similar barriers still exist, as reported in [24,25].

Syafrudin et al. [22] proposed a real-time monitoring framework of manufacturing systems focusing on the automotive industry. The framework utilizes BD collected from IoT-based sensors and processes it by a hybrid prediction model. In particular, unstructured datasets collected from manufacturing processes by the temperature, humidity, accelerometer, and gyroscope sensors were preprocessed using the platforms known as Apache Kafka (message queue), Apache Storm (real-time processing engine), and MongoDB (data storage). Subsequently, density-based spatial clustering with nose-removing capacity was used to detect outliers and data classification for fault detection. The proposed BD-driven system helped prevent unexpected losses caused by faults during the manufacturing of automotive parts and performing assembly.

Wiech et al. [23] considered that BDA and manufacturing execution systems are needed to achieve the objectives of Industry 4.0. They studied the implementation levels of BDA and manufacturing execution systems by conducting a survey in which more than a hundred organizations participated. They found that BDA and manufacturing execution systems are heavily correlated. In addition, these technologies have unexpectedly distinct performance effects that are not likely to depend on the organizational structures.

Escobar et al. [24] conducted a literature review to understand the interplay of BD and process monitoring for quality control from the context of Industry 4.0 (i.e., Quality 4.0). They proposed that manufacturing BD challenges can be tackled by implementing a sev-en-step approach: identity, accessorize, discover, learn, predict, redesign, and relearn. In addition, they found that a vast array of process datasets collected either from plants or from laboratories in the form of pictures, signals, and direct measurements can be analyzed in real-time using simple machine learning algorithms rather than deep learning.

Yu et al. [25] presented a BD ecosystem for predictive maintenance. The data sources were the sensor signals collected from large-scale manufacturing plants. The ecosystem used numerous technologies (data lake, NoSQL database, Apache Spark, Apache Drill, Apache Hive, and OPC Collector) to solve data ingestion, integration, transformation, storage, analytics, and visualization. In addition, the required transformation protocols, authentication, and encryption methods were also addressed to secure the data transfer. Finally, the Map-Reduce decentralized principal component analysis made sense of BD for predictive maintenance. The Map-Reduce decentralized principal component analysis, which is simple and can respond faster on a real-time basis, was used to make sense of sensor signals for predictive maintenance.

Faheem et al. [26] considered that sensor datasets collected by a wireless sensor network from various sources (equipment, machines, assembly lines, material handling devices, and inspection activities) constitute industrial BD. The datasets are subjected to trigger errors and low transmission quality due to high noise, signal fading, multipath effects, heat, and electromagnetic interference. To solve the abovementioned problems, they introduced a multi-channel and multi-radio architecture called CBI4.0. As confirmed by EstiNet 9.0 simulator, the proposed architecture exhibited robust performance compared to other wireless sensor signal networks used to support BD in the automotive industry.

O'Donovan et al. [27] presented data requirements, system requirements, and information system models for utilizing BD in equipment maintenance. The goal was to provide a scalable and fault-tolerant BD pipeline for integrating, processing, and analyzing datasets relevant to industrial equipment. The focus was on the highly automated large-scale manufacturing environments where internet-aware smart sensors play a vital role.

Shah et al. [28] showed that BD of sensor signals collected from IoT-networked manufacturing devices is effective for manufacturing process modeling and monitoring. However, machine learning techniques must be employed to make sense of large datasets. In particular, they have developed an IoT-based testbed capable of handling BD (about 70GB) from a pipe flow system coupled with five IoT-based vibration sensors. Furthermore, they compared the machine learning performances of complex deep learning models with simple statistical learning models in processing the sensor signals. They identified that simple statistical learning could achieve superior results than deep learning because there are still unsolved challenges making deep learning less effective.

On the contrary, Fang et al. [29] showed that deep learning is more effective than other machine learning (e.g., linear regression, back-propagation, and multi-layer and deep belief networks) in making sense of manufacturing BD while predicting job remaining time. They used BD collected from various sensors in a large-scale job shop equipped with 44 machines producing 13 types of parts. The proposed framework needs raw data collection, candidate dataset design and selection, and predictive modeling using a deep learning approach denoted as stacked sparse autoencoder.

Zhang et al. [30] proposed an energy-aware cyber-physical system where energy-related BD and production-related BD play a vital role. The datasets originated from energy monitors (sensors) mounted on machine tools and gas, liquid, and cutting fluid circulation devices. Before making sense of these datasets, they were cleaned by removing the noise and abnormalities. Finally, deep belief networks classified the continuous energy consumption data according to different machining states, which helped ensure low energy production.

Ko and Fujita [31] developed evidential analytics for unearthing the buried information in BD samples focusing on the manufacturing of semiconductors. They found that raw datasets in BD often exhibit undesirable characteristics such as unspecified sampling principles and analytics baselines, a large number of redundant variables or features, a mixture of relevant and irrelevant datasets, indistinguishable noise, and outliers in datasets. BDA must handle these characteristics and identify the causes of damage beforehand. To achieve this, they proposed analytics denoted as evidential analytics for buried information (EABI) that usages the concept of granular information. EABI consists of three phases. The first phase generates baselines expressing the relevance to damages in directions high and low for variable reduction. The second phase unearths the preference and relevance together. The last phase aggregates evidence among variables for evaluating the samples.

1.4.3 Dealing with graphical datasets

This section presents a literature review to understand the interplay of BD and SDGs as elaborately as possible.

For monitoring sustainability, Mihaly et al. [32] studied the interplay of (*a*) national sustainability policy, (*b*) international partnerships, domestic activities, and achievements, (*c*) status of professional education, (*d*) spatial databases and services to support the implementation of the sustainable development, I a case study on the internationally recognized soil geoinformation system, (*f*) national earth observation information system and perspectives of its applications. They found that BD regarding earth observations and geospatial data must be enacted to decide the right set of policies for implementing SDGs at a national level.

Li et al. [33] showed how to benefit from BD in satisfying SDG 16. In particular, they used BD available in social networks (6 million pieces of tweets) on corruption and identified that the bribery of law enforcing authority and corruption in the healthcare sector put obstacles in achieving SDG Target 16.5. Furthermore, they utilized unsupervised machine learning methods to make sense of tweets expressed in natural language.

Ryan et al. [34] studied the ethical issues regarding AI and BD while meeting SDGs. They conducted six empirical case studies to see how smart information systems can meet the challenges of six SDGs (2, 3, 7, 8, 11, and 12). They showed that smart information systems are insufficient and may exacerbate or create new issues for community development.

Ryan and Anya [35] studied the ethical implications of AI and BD for implementing SDG in smart cities with large populations. They found that dealing with privacy, ensuring accurate datasets, reducing costs, and building general stakeholders' trust are the main concerns of implementing BD for achieving SDGs.

McFeely [36] reported that out of 232 SDG indicators, only 93 are classified as Tier 1, i.e., the indicators are clearly defined, and internationally accepted standards compile data from at least 50 percent of the countries. The remaining indicators are Tier 2 (72 indicators) or Tier 3 (62 indicators). Tier 2 indicators are clearly defined, but countries do not regularly produce the data. Tier 3 indicators are conceptually clear, but no internationally accepted standards are yet available. Some data sources are used for estimating multiple SDGs. For example, utility bills are used to estimate economic well-being, and well-being is directly related to SDGs 1, 8, 10, and 11. Mobile phone utilization datasets are used to estimate public health and disaster, and public health and disaster are directly related to SDG 2, 3, 8, 11, 15, and 16. Nevertheless, web scraping, scanners, mobile phones, social media, satellite images, smart meters, credit cards, road sensors, health records, ship identification, and criminal records are now SDG-related BD sources.

Global assessment of institutional readiness for using BD in official statistics is presented in [37]. The following points have been raised: (*a*) From a strategic coordination viewpoint, BD must be exchanged from all regions through the United Nations Global Platform, where the national statistics organizations must play a vital role. (*b*) Legal frameworks must be revised to materialize data sharing between national statistics organizations, private sector data owners, and other stakeholders without valeting the data privacy and data protection laws. (*c*) Human resource initiatives must be taken overarching competency for BD skills development. Thus, partnerships with higher education institutes to allow up-skilling of BD competency must take place. (*d*) Required IT infrastructure (cloud storage facilities) in all countries with necessary prerequisites must be maintained to benefit from BD.

Mwitondi et al. [38] performed data segmentation considering that each SDG is a node of a BD source. They have elucidated the complex overlap of the SDGs by using data from different sources, as described in [56–58], regarding SDG indicators. The complexity of the data (data randomness, variation in sample size, and socio-economic, cultural, and geopolitical factors) necessities new data handling algorithms. BDA help understand the interplay of the SDG indicators and open new paths to interdisciplinary research.

Ferreira et al. [42] showed that BD consisting of satellite images of earth observation could be used to address the SDG indicators. However, innovative methods and tools to process ever-growing earth observation data are needed. In this respect, data analytics techniques can help make sense of the enormous quantity of earth observation data available in various formats and collected from numerous sources.

Kashyap and Verkroost [43] analyzed BD available on social networks (LinkedIn™) to understand gender gaps using regression analysis. They found that LinkedIn Gender Gap Index (GGI) strongly correlates with International Labor Organization (ILO) ground truth professional gender gaps. This study thus shows the efficacy of social networks being a source of BD for sustainability analysis.

Hassani et al. [44] studied the up-to-date connections between the SDGs and BD using the global Google trend. They summarized the impact of BD on SDGs, showing the current state and challenges to overcome in the foreseeable future. They reported that SDG 1 attracted the most attention to BD projects among all

SDGs. In this respect, data silos collected from mobile phones and satellite images and geodata were identified as the top two data sources by which poverty has been combated. For example, satellite image BD processed by AI can help evaluate the building, car counts, road density, pavement, road width, and roof materials. This information helps the stakeholders to match skilled labor with suitable employment or local products with buyers worldwide. As far as SDG 2 is concerned, BD generated from IoT-driven smart sensor networks and geological surveys contribute to smart farming and precision agriculture. BD generated while monitoring infectious diseases and mental health can contribute to SDG 3. SDG 4 needs prediction and estimation of early school dropouts, which can be performed using BD accurately. In terms of SDG 9 and SDG 12, to the best of our knowledge, BD has just started to earn a great deal of attention. The reason is that without BD, it is not possible to ensure a sustainable supply chain, product life cycle management toward greenhouse gas reduction, consumer behavior assessment, and e-commerce.

Allen et al. [45] presented a comprehensive literature review showing which BD sources estimate SDGs. They reported that datasets generated by satellite data acquisition systems, surveys, tracking systems, sensors, administrative practices, and opinions are BD sources for sustainability assessment. However, regarding SDG 12, there were no reported data sources.

The above literature review reveals that the interplay of BD and SDGs is highly complex. BD's utilization in assessing the degree of fulfillment of SDGs or making necessary arrangements at the global or local level is still in its infancy. Therefore, more research should be conducted to develop more pragmatic methods for mitigating the uncertainty and computational complexity associated with the BD of sustainability.

1.4.4 Machining decision making

Ji et al. [52] introduced a BDA-based machining optimization approach for distributed process planning. This approach represents the machining resources (workpiece, machining requirement, machine tool, cutting tool, machine conditions, and process) by relevant data attributes. This approach also deploys a hybrid algorithm composed of a Deep Belief Network (DBN) and Genetic Algorithm (GA) for performing the optimization using machining data. However, the structure of data analytics is not described in detail.

Chen et al. [53] presented a BD processing scheme for electric discharge machining (EDM). The processing scheme adapts two available BDA tools: Hadoop Distributed File System (HDFS) and Spark (an open-source BDA engine), for extracting key features (e.g., average spark frequency, open/short circuit ratio, short circuit duration, average ignition delay time, average short circuit current, average peak discharge current, average discharge energy, average discharge pulse duration, and alike) underlying EDM-relevant machining data. However, the authors did not elucidate the application of the extracted features for predicting the machining quality of EDM but mentioned it as a future research direction.

1.5 Research objectives

According to the literature review, following the main three focused research areas, it can be seen that many authors have proposed BDA for manufacturing processes. Neither of the work considers the issue of DMC nor steadfast procedures for developing BDA of machining operations. Furthermore, the literature review shows that BD's utilization in assessing the degree of fulfillment of SDGs or making necessary arrangements at the global or local level is still in its infancy. Therefore, more research should be conducted to develop more pragmatic methods for mitigating the uncertainty and computational complexity associated with the BD of sustainability.

Considering these research opportunities, As it was mentioned, BD of the manufacturing process in this study are categorized into three main parts: 1) Dataset for DT; 2) Control Variable (CV)-Evaluation Variable (EV)-centric dataset; 3) Graphical dataset. This thesis considers the problem of developing BD and BDA for smart manufacturing, focusing on these categories and the following three related issues.

Issue 1: DTs of manufacturing phenomena are supposed to machine-learn the required knowledge using relevant datasets available in BD. Therefore, a research question is how to preprocess manufacturing phenomena-relevant datasets for using them directly in DTs.

Issue 2: BD is often visualized using several two-dimensional plots. These plots are then used to make a decision informally. Consequently, a research question is how to make formal decisions by computing two-dimensional plots, not numerical data.

Issue 3: BD and analytics require expensive resources and sophisticated computation arrangements. Thus, BD hardly benefits small and medium-sized manufacturing organizations, resulting in "Big Data (BD) inequality." Consequently, a research question is how to eliminate BD inequality. Chapter 2 presents the mathematical settings needed to understand the computational aspects of the proposed BDA.

1.6 Thesis structure

The remainder of this thesis is organized as follows:

Chapter 1 presents the background, objective, scope, and limitations of this thesis. It also presents a comprehensive literature review on BD relevant to smart manufacturing.

Chapter 2 describes the proposed BDA framework showing all the subsystems. In this chapter, the functional requirements of the subsystems are explained. BDA is developed for manufacturing process-relevant decision-making. The proposed analytics consists of five integrated systems: 1) BD preparation system, 2) BD exploration system, 3) data visualization system, 4) data analysis system, and 5) knowledge extraction system. The BD preparation system prepares contents that exhibit the characteristics of digital manufacturing commons.

Chapter 3 deals with Issue 1. A DT consists of five modules (input, modeling, simulation, validation, and output modules), and BD must supply datasets for building these modules. This chapter presents a manufacturing phenomenon-related datasets preprocessing method considering the four modules of DTs (input, modeling, simulation, and validation modules). As an example, the preprocessing of surface roughness-relevant datasets is considered.

Chapter 4 deals with Issue 2. This chapter described the developed BDA framework for the CV-EV-centric dataset. This system can support user-defined ontology and automatically produces Extensible Markup Language (XML)-based datasets. The BD exploration system can extract relevant datasets prepared by the first system. The system uses keywords derived from the names of manufacturing processes, materials, and analyses- or experiments-relevant phrases (e.g., design of experiment). The third system can help visualize relevant datasets extracted by the second system using suitable methods (e.g., scatter plots and possibility distribution). The fourth system establishes relationships among the relevant control variables (variables that can be adjusted as needed) and evaluation variables (variables that measure the performance) combinations for a given situation. In addition, it quantifies the uncertainty in the relationships. The last system can extract knowledge from the outcomes of the fourth system using user-defined criteria (e.g., minimize surface roughness and maximize material removal rate). The efficacy of the proposed BDA is demonstrated using a case study where the goal is to determine the right states of control variables of dry electrical discharge machining for maximizing material removal rate. It is found that the proposed BDA is transparent and free from BD inequality.

Chapter 5 deals with Issue 3. BDA is developed to compute two-dimensional plots generated from BD. The efficacy of the tool is demonstrated by applying it to assess sustainability in terms of Sustainable Development Goal (SDG) 12 (responsible consumption and production). Regarding SDG 12, functional, economic, and environmental issues of engineering materials play a vital role. Accordingly, three two-dimensional plots generated from BD of engineering materials are computed using the proposed analytics. The plots refer to six criteria (strength, modulus of elasticity, cost, density, CO2 footprint, and water usage). The proposed analytics correctly rank the given materials (mild steel, aluminum alloys, and magnesium alloys).

Chapter 6 describes future research directions and discusses the implication of this study from the viewpoint of smart manufacturing.

Finally, Chapter 7 provides the concluding remarks of this thesis.

Chapter 2: Developing Big Data Analytics Framework

Chapter 2 describes the proposed BDA framework showing all the subsystems. In this chapter, the functional requirements of the subsystems are explained. BDA is developed for manufacturing process-relevant decision-making. The proposed analytics consists of five integrated systems: 1) BD preparation system, 2) BD exploration system, 3) data visualization system, 4) data analysis system, and 5) knowledge extraction system. The BD preparation system prepares contents that exhibit the characteristics of digital manufacturing commons.

This BDA is developed for manufacturing process-relevant decision-making. The proposed analytics consists of five integrated systems: 1) BD preparation system, 2) BD exploration system, 3) data visualization system, 4) data analysis system, and 5) knowledge extraction system. The BD preparation system prepares contents that exhibit the characteristics of digital manufacturing commons. Thus, the system can support user-defined ontology and automatically produces Extensible Markup Language (XML)-based datasets. The BD exploration system can extract relevant datasets prepared by the first system. The system uses keywords derived from the names of manufacturing processes, materials, and analyses- or experiments-relevant phrases (e.g., design of experiment). The third system can help visualize relevant datasets extracted by the second system using suitable methods (e.g., scatter plots and possibility distribution). The fourth system establishes relationships among the relevant control variables (variables that can be adjusted as needed) and evaluation variables (variables that measure the performance) combinations for a given situation. In addition, it quantifies the uncertainty in the relationships. The last system can extract knowledge from the outcomes of the fourth system using user-defined criteria (e.g., minimize surface roughness and maximize material removal rate). The efficacy of the proposed BDA is demonstrated using a case study where the goal is to determine the right states of control variables of dry electrical discharge machining for maximizing material removal rate. It is found that the proposed BDA is transparent and free from BD inequality.

2.1 Background

As described in the introduction part, one of the most important aims of this research is to develop a BDA framework that is free from BD inequality and the digital divide. For this purpose, some measures are needed to mitigate these matters. For this purpose, it is better to have a BDA system that is not computationally heavy and highly resource-depended. It means that it is better not to use the enormous and complex analytical approach in our BDA system, which makes it a black box where the inputs and outputs are visible, but the inner processes remain unknown. Furthermore, as was described in the introduction chapter and indicated in the literature review, the digital manufacturing commons (DMCs) are usually prepared from heterogeneous documentation of past manufacturing activities.

It could be really challenging when it is supposed to convert past manufacturing activities documentation to DMC. The main reason is that there is exponentially increasing documentation regarding manufacturing activities. Figure 2-1 depicts this exponential growth of documents regarding machining operations which is one of the most important and practical manufacturing processes in SMEs and large manufacturing organizations. This graph considers the publication published in Scopus® between 1980 and 2021. Then, it predicts the number of publications with the keyword "machining" in the title. The number of publications grew exponentially from nearly 200 in 1980 to more than 2300 in 2021. The number of publications is expected to reach nearly 4000 in 2030. Similar to DMC, preparing relevant analytics is also a challenge. As described in the literature review (chapter 1), many authors have proposed BDA for manufacturing. Neither of the work considers the issue of DMC nor a steadfast procedure for developing BDA of machining operations. Accordingly, this chapter focuses on the development of a BDA framework for machining operations considering the issue of BD inequality and the digital divide.

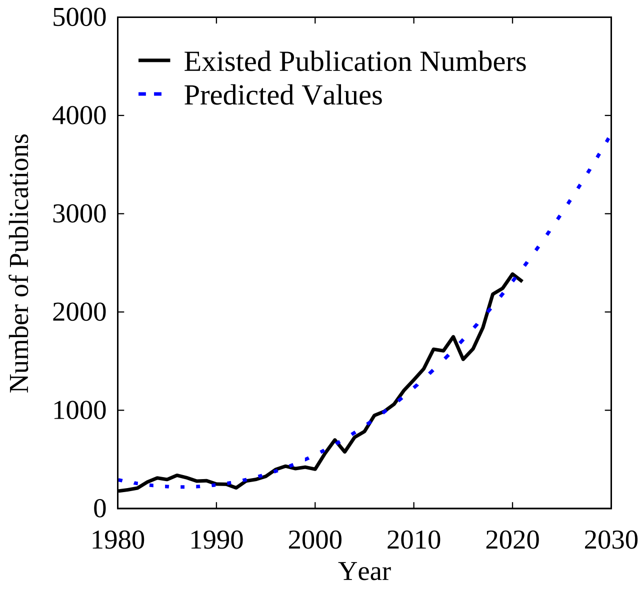


Figure 2-1. Number of Publications with "Machining" in the title (Source: Scopus®)

2.2 Big data analytics for smart manufacturing

As it is explained, the main purpose is to prepare a framework for the BDA of smart manufacturing, which considers the aspect of BD inequality and DMC. Figure 2-2 shows the desired BD analytics conceptual scheme and its parts.

Now, because the purpose is to develop the BDA for manufacturing processes, let us consider those with their attributes. A manufacturing process and, more specifically, a machining operation (turning, milling, electric discharge machining, and alike) is usually controlled by controlling the right set of process variables.

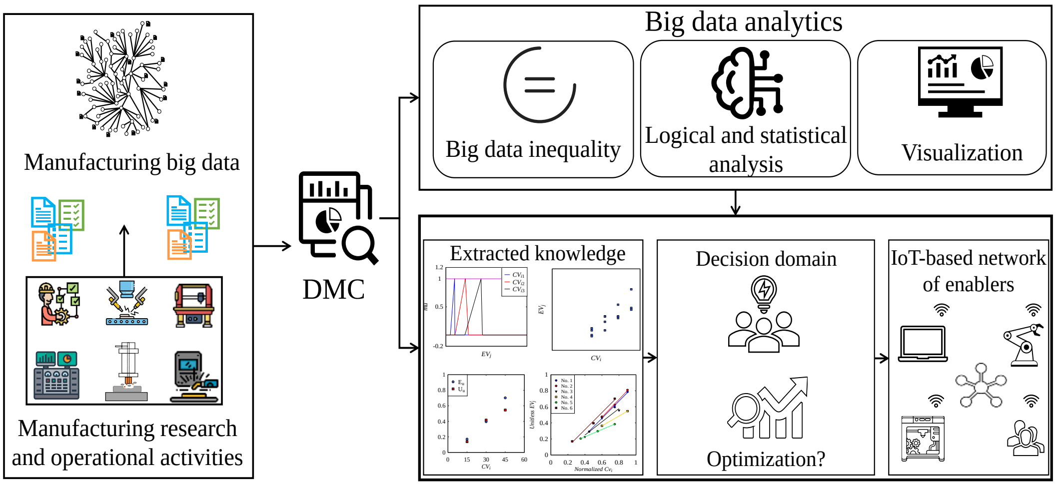


Figure 2-2. BDA concept intended for smart manufacturing and DMC

To run a machining operation (e.g., milling operation), we fix some cutting conditions (see Figure 2-3). Later, we gather the experimental results. For example, experimental results for tool wear, cutting force, chip, and surface roughness. Afterward, we analyze the results and establish the relationship between cutting conditions and experimental results. In this study, we use two terms, control variable (CV) and evaluation variable (EV). CVs are usually the cutting conditions, and EVs are variables that we use to measure the performance of machining operations.

After analyzing these experimental results, some graphs showing the relationship between EVs and CVs can be obtained. For example, let us consider the milling operation and two graphs obtained between CV-EV combinations: feed rate-material removal rate (*MRR*), and feed rate-surface roughness (SR). According to these graphs, it is obvious that, for example, If the purpose is to minimize surface roughness, it is needed to minimize the feed rate, and if the purpose is the maximization of *MRR*, it is needed to maximize the feed rate. This means that there is a conflict in terms of feed rate. Therefore, we need to use sophisticated mathematical procedures to optimize feed rate and get the required knowledge out of that. This matter is true for other cutting conditions, too. So, developing a specific BDA framework can help to solve these sophisticated conflicts in manufacturing processes for decision-making and getting the required knowledge from the BDA framework. Accordingly, we focus on the CV-EV-centric dataset for developing a specific BDA framework in this study. According to this CV-EV-centric dataset and regarding the concept of smart manufacturing, these CV-EV-centric documents of past manufacturing activities entail DMC.

According to Figure 2-2, BDA can obtain knowledge out of DMC for this CV-EV-centric dataset, like maximization and minimization. Based on that, as described in the introduction section, in the next section, the BDA framework is developed that highlights this CV-EV-centric dataset and implies the aspects of BD inequality and the digital divide.

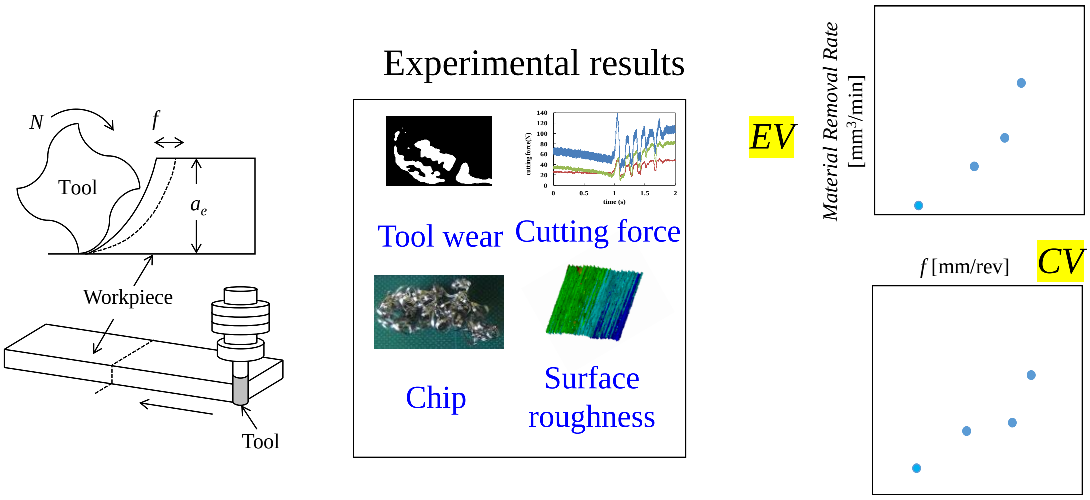


Figure 2-3. Example of a manufacturing process (milling), experimental results, and CV-EV graphs

2.3 Proposed big data analytics framework

The proposed BDA framework (Figure 2-4) consists of five interconnected systems: the BD preparation system, BD exploration system, data visualization system, data analysis system, and knowledge extraction system. Figure 2-4 Shows the proposed BDA framework.

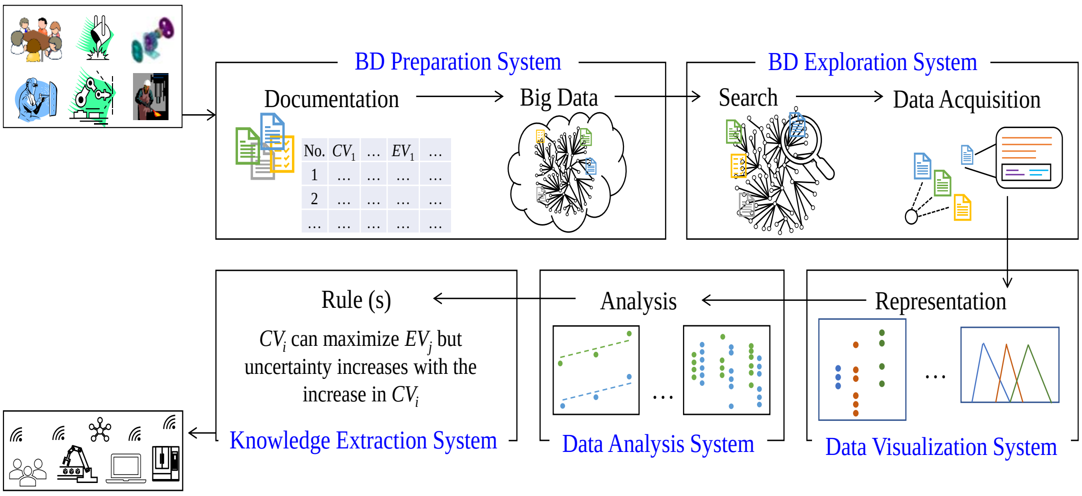


Figure 2-4. Proposed BDA framework

In this study, the framework and five interconnected systems are developed using the JAVA™ platform. It is considered that the framework and systems must be both human- and machine-friendly.

The first system is the BD preparation system. It creates digital manufacturing commons for different purposes for manufacturing processes. It also helps form BD for manufacturing processes. As an example of a CV-EV-centric dataset, the remarkable thing is that it explicitly shows the numerical data relevant to CV-EV combinations.

The second system is the BD exploration system. It helps search the digital manufacturing commons in the BD and acquire the relevant datasets for the particular situation.

The third system is the Data Visualization System. It represents relevant datasets using various visualization schemes like scatter plots and possibility distribution

The fourth system is the data analysis system. For example, the CV-EV-centric dataset establishes relationships among all CV-EV combinations for a given situation. In addition, it quantifies the uncertainty in the relationships.

The last system is the knowledge extraction system. This system extracts rules, equations, and other forms of knowledge that can be used for the smart manufacturing system. Thus, this system directly interacts with IoT-based manufacturing enablers.

As mentioned in the previous slide, we developed a BDA framework for machining operations, now according to that let us consider system development for this BDA scheme for three main dataset categories as it was mentioned before: 1) Dataset for DT, 2) CV-EV-centric dataset, 3) Graphical dataset

Chapter 3: Preparing Datasets for Digital Twin

Chapter 3 deals with Issue 1. A DT consists of five modules (input, modeling, simulation, validation, and output modules), and BD must supply datasets for building these modules. This chapter presents a manufacturing phenomenon-related datasets preprocessing method considering the four modules of DTs (input, modeling, simulation, and validation modules). As an example, the preprocessing of surface roughness-relevant datasets is considered. This chapter presents a manufacturing phenomenon-related datasets preprocessing method considering the four modules of DTs (input, modeling, simulation, and validation modules). As an example, the preprocessing of surface roughness-relevant datasets is considered.

3.1 Background

As it is described in chapter 1, introduction, in the smart manufacturing context, there are manufacturing enablers such as CAD/CAM systems, process planning systems, CNC machine tools, measuring devices, actuators, robots, and human resources. The difference is that the enablers create an IoT-based network, allowing both vertical and horizontal integrations. At the same time, the enablers must perform human-like cognitive tasks, such as understanding current situations, predicting future consequences, deciding the right courses of action, and adapting to new situations as autonomously as possible. The autonomous execution of the abovementioned cognitive tasks requires a great deal of knowledge [54,55] that can be extracted from the relevant segments of BD using artificially intelligent systems.

Machine learning capacities empower these systems [56]. DTs contain knowledge extraction systems, extracted knowledge, and capacities to perform human-like cognitive tasks [54]. DTs also provide (or receive) feedback to (or from) the IoT-based manufacturing enablers to keep the enablers adaptive to new situations [57].

Regarding the aspect of BD, as it was described in the introduction part, BD consists of a vast array of heterogeneous unstructured, semi-structured, and structured datasets. Regarding these datasets, some of them can be accessed through the Internet, and some are not.

However, the HCPS-friendly segments of BD must have the following characteristics. The segments must be readily accessible to all stakeholders, preferably through the Internet. In addition, the segments must be both human- and machine-readable. Moreover, the segments can be effortlessly integrated with the knowledge extraction systems and, thereby, to DT. The segments of BD exhibiting the abovementioned characteristics are not readily available. They need to be prepared using the relevant documentation of past research and operational activities. The documentation is, by nature, messy, and there is no steadfast procedure by which the documentation can be converted into an HCPS-friendly BD. So, this chapter's main focus is on this issue. In particular, this study uses the case of surface roughness to elucidate how to prepare datasets to be included in an HCPS-friendly BD.

Surface roughness [58] is a major concern of all manufacturing processes, and ensuring the right surface roughness requires knowledge. In most cases, the required knowledge is extracted from experimental datasets. In HCPS, a dedicated DT supplies the surface roughness-relevant knowledge to the manufacturing enablers. However, before a DT supplies the required knowledge, it must be constructed. The question is, from where the DT collects the information? The obvious answer is that BD supplies the information. In other words, BD must host datasets useful for building a DT. Consequently, a mere digitized version of the documentation of experimental and operational activities is not enough to build BD;

datasets needed to build DT must be added to BD. Otherwise, BD may not be useful for HCPS.

3.2 Surface roughness data

As it is described in the background, the root of all problems associated with constructing BD and its utilization is the datasets themselves [59]. Thus, before presenting the proposed method to prepare the datasets to be included in BD of surface roughness, it is important to see how and what kind of datasets of surface roughness is often documented after performing experimental and operational activities. This section serves this purpose.

Nowadays, surface roughness [4,5] is measured by laser-based non-contact surface metrology instruments [62], as shown in Figure 3-1(a). This type of instrument moves a laser source on the surface to be measured in a definite trajectory, as shown in Figure 3-1 (b). For this, the instrument first sets an XY-mesh (Figure 3-1 (c)) and obtains the height information of the surface (Figure 3-1 (d)). In addition to height information, the instruments can represent a mesh using a pixel [63]. As such, two types of information are produced. One of the types consists of a 3D surface-based on height information (Figure 3-1 (e)-(f)).

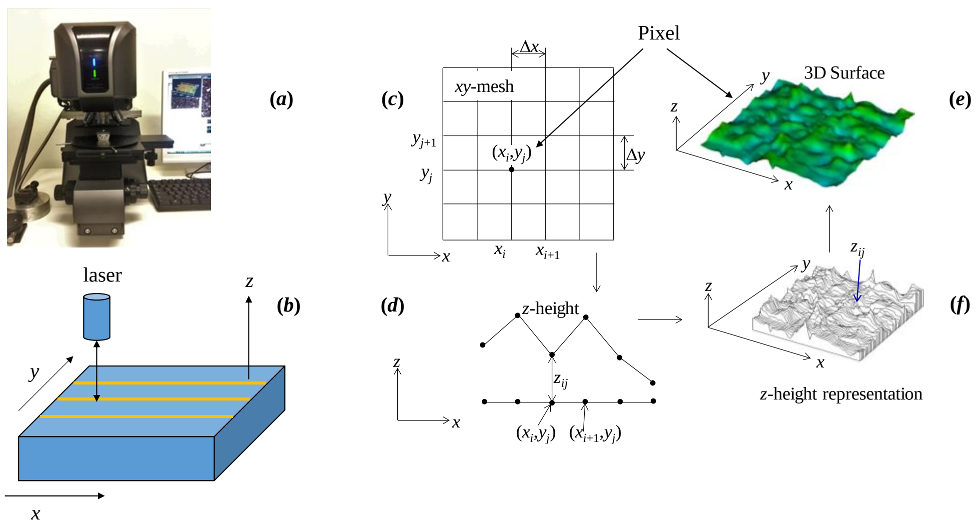


Figure 3-1. Measuring surface roughness

The other is the images of the measured surface, as shown in Figure 3-2(a)-(c). In particular, Figure 3-2(a) shows a raw image of an arbitrary surface. Figure 3(b) shows a binary image [63]. This image is extracted from the image shown in Figure 3-2(a) and carries valuable topographical information about the surface [63]. Figure 3-2(c) shows a color image where different heights are depicted using different colors. This is also useful for surface topography analysis. Figure 3-2(d) shows a 3D surface rendered from the height datasets using a curve-smoothening technique. Figure 3-2(e) shows a height profile of the surface for a given y (or x) along the x (or y) direction [62]. A primary profile (not shown in Figure 3-2) is obtained by removing linear or curved "form error" underlying the surface height profile 5,6]. Figure 3-2(f) shows a surface roughness profile obtained by removing the waviness from a primary profile.

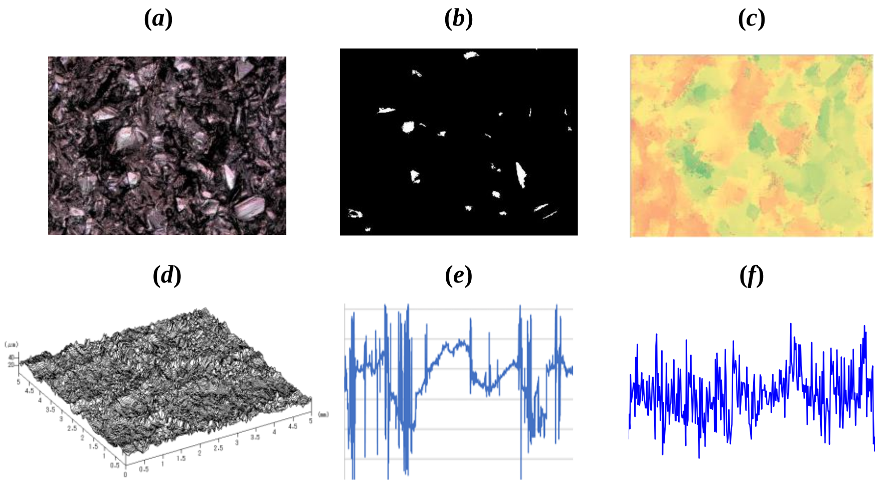


Figure 3-2. Different types of surface roughness data

An online system developed by NIST of the USA (see Figure 3-3) supports the quantification as mentioned above.

The system provides a user interface (Figure 3-3) to upload a surface profile dataset. The system then uses the standard procedures and calculates parameters, including R_a and R_z. Instead of using the standard parameters, advanced parameters such as fractal dimension, surface entropy, and possibility distributions [19] can be used to accurately quantify the complexity of surface roughness.

The surface profile height datasets must be retained in BD to calculate these advanced parameters. Thus, XML-based datasets where the information of Ra and Rz are retained, as shown in [64], require a revisit.

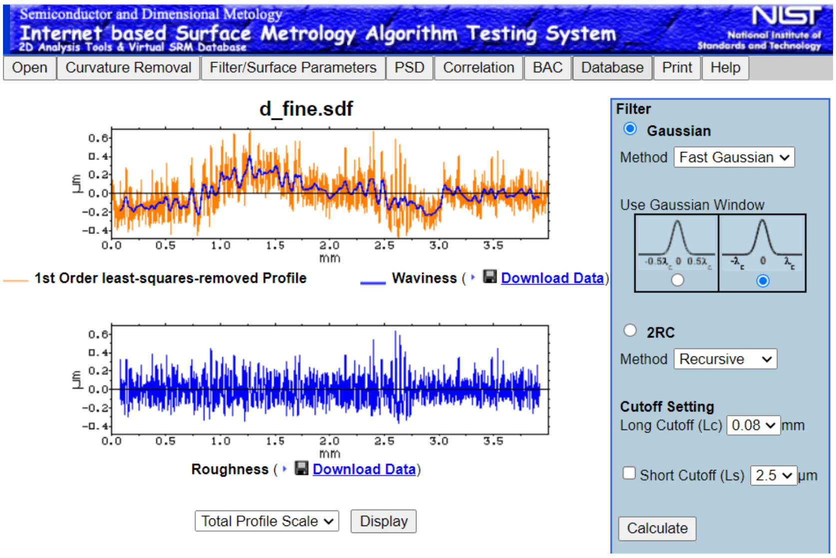


Figure 3-3. Existing Internet-based surface roughness system

3.3 Preparing datasets of surface roughness for big data

In chapter 1, introduction, the scope of limitations of BD from the perspective of Industry 4.0 is described. In synopsis, the following remarks can be made.

(a) The raw or proposed datasets added to BD may not be able to bring benefits to smart manufacturing if the targeted usages of the datasets are not considered in the first place. This means that the datasets regarding surface roughness, as shown in Figure 3-2, should not be added directly to construct BD of surface roughness. Instead, the datasets must be preprocessed based on the targeted use before adding them to the BD of surface roughness.

(b) There is no steadfast approach available to preprocess raw datasets collected from real-time sensor signals or past experimental and operational activities for building BD.

(c) Large organizations with the strength of maintaining sophisticated IT infrastructures (for example, the automotive and chemical industries) can execute sophisticated BD analytics. However, the opposite scenario prevails for small and medium-sized organizations.

(d) The effectiveness of well-known machine learning approaches while making sense of BD is controversial. For example, some authors advocate complex machine learning approaches (artificial neural networks-driven machine learning such as deep learning); others advocate rather simple ones.

Therefore, many open questions regarding constructing and functionalizing BD for smart manufacturing exist. In order to address the abovementioned problems in a befitting manner, this section presents a method that can be used to prepare the datasets of BD of surface roughness from the context of HCPS. Before presenting the proposed method, three salient issues (BD inequality, semantic annotation, and DT) are presented as follows.

First, consider the issue of BD inequality [65]. Recall that BD consists of a vast array of heterogeneous datasets (unstructured, semi-structured, and structured) that evolve with time [66],[67]. Arrangements to extract knowledge from a relevant segment of BD are computationally heavy and highly resource-dependent as well. As a result, BD benefits large organizations. Medium and small organizations fall behind. This results in BD inequality [65]. Unfortunately, studies dealing with BD integration with HCPS, e.g., [14 –17], have not yet addressed BD inequality. For example, consider the work in [68]. The authors formulated BD analytics, where the BD is integrated with machine learning and computational intelligence paradigms. The arrangement requires highly sophisticated computing devices and high-skill human resources. As such, these systems are beyond the affordability of medium and small organizations. One way to minimize the involvement of sophisticated computing devices and high-skill human resources is how the datasets are prepared in the first place. Care should be taken while developing the dataset preparation methods so that the methods help mitigate BD inequality.

Secondly, consider the issue of semantic annotation or metadata. It (semantic annotation or metadata) has become a crucial issue due to the advent of web technology, and many authors have contributed toward semantic annotation or metadata [18–24]. The fact of the matter is that the new web technology called Semantic Web (SW) [79] is in the process of replacing its predecessor. SW-based datasets need both the datasets themselves and “data about datasets.” This “data about datasets” is referred to as semantic annotation or metadata. As a result, all relevant datasets can be gathered quickly, and search engines become more effective, which is not the case now. At the same time, the semantic annotations facilitate the amalgamation of relevant datasets scattered in different information silos. However, for constructing metadata, different types of ontological approaches are proposed in the literature. Most of the approaches depend heavily on the query language and data access protocol (e.g., SPARQL) customized for the resource description framework (RDF) [80]. As a result, the current semantic annotation preparation approaches are unscalable and esoteric. Making the semantic annotation preparation approaches more user-friendly, scalable, and less esoteric is challenging. This challenge can be overcome if natural language-based semantic annotations are used, ensuring the freedom of using any phrases that the users prefer. In this respect, concept mapping is the right approach, as shown in [27, 28].

Lastly, consider the case of DT, as we described in the introduction part. By definition, DTs are the computable virtual abstractions of real-world objects, processes, and phenomena [21,27]. They have real-time response capacity. As mentioned, DTs host knowledge extraction systems, knowledge-base, and human-like cognitive tasks and provide (or receive) feedback to (or from) the IoT-based manufacturing enablers. They keep the enablers adaptive to new situations. Thus, they serve as the brains of IoT-based enablers. Since surface roughness is a manufacturing phenomenon, a DT dealing with surface roughness is a phenomenon twin. A phenomenon twin (of surface roughness) consists of five modules: input module, modeling module, simulation module, validation module, and output module [82]. The input module extracts information from a source (e.g., BD) for building other modules. The modeling module models a phenomenon. The simulation module simulates the expected outcomes of the phenomenon upon request from the respective enablers. The validation model validates the integrity of results produced by the twin. Finally, the output module integrates the twin with the relevant IoT-embedded enablers. See [29,30] for more details. Considering the issues of BD inequality, semantic annotation, and DT, a BD preparation method is proposed, as schematically illustrated in Figure 3-4. As seen in Figure 3-4, a dataset of BD consists of four segments denoted as semantic annotation, roughness model, simulation algorithm, and simulation system. These segments can be downloaded while developing a DT of surface roughness.

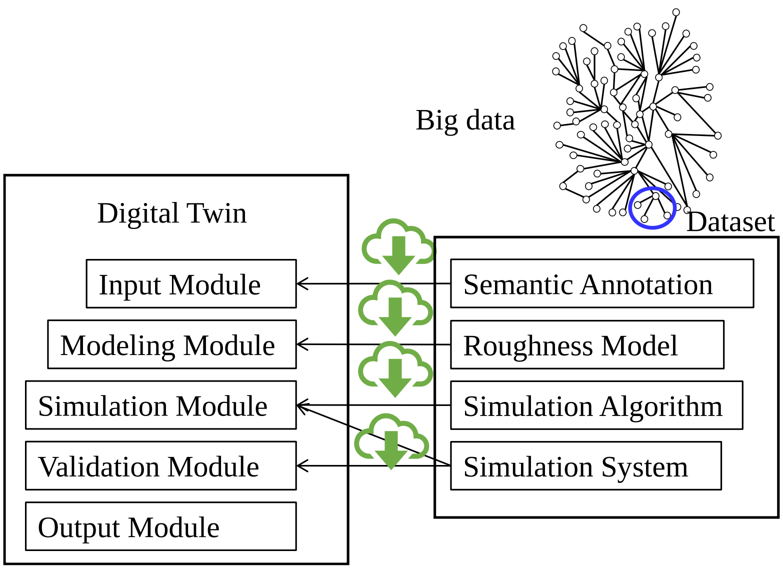


Figure 3-4. Proposed dataset preparation method

The segment denoted as semantic annotation provides information for the input module of the DT. The segment denoted as the roughness model provides information for the modeling module of the DT. The segment denoted as the simulation algorithm provides information for the simulation module of the DT. The segment denoted as the simulation system provides information for the DT's simulation and validation modules. There are no segments in a dataset that provides information for the output module of the DT. Based on the proposed method, a set of datasets of surface roughness is constructed for different types of material processes such as turning, milling, grinding, polishing, and electric discharge machining. The results regarding grinding are shown in the next section.

3.4 Results and discussions

This section presents the noteworthy results obtained using the proposed method and discusses the results' implications. For better understanding, the roughness model segment of the dataset is presented first, as follows. The roughness model segment presents the information of the delay map of a roughness profile. Two pieces of information are stored. The first is the roughness DNA, and the other is two sets of possibility distributions induced from the delay map. For this, the following formulation is considered.

Let $x(i) \in [0,1]$, $i = 0,1,\ldots$, be the normalized heights of a measured surface. A delay map consists of the ordered-pair $(x(i),x(t+d))$, $i = 1,2,\ldots$, where d denotes the delay, a non-zero integer. Let Sj, $i = 1,2,\ldots,M$ be the states of roughness dividing the interval [0,1] into M mutually exclusive intervals. The states of $x(i)$ are represented by a roughness DNA = $(Si|\ i = 0,\ldots,N)$ so that $\forall Si \in \{S_j \mid j = 1,\ldots,M\}$. The abscissa of the delay map is represented by the triangular fuzzy numbers $(a1(Sj),b1(Sj),c1(Sj))$, $j = 1,\ldots,M$, where $[a1(Sj), c1(Sj)]$ is the range of Sj and $b1(S\mathrm{j})$ is the midpoint of Sj. The ordinate of the delay map is represented by the triangular fuzzy number $(a2(Sj),b2(Sj),c2(Sj))$, $j = 1,\ldots,M$ where the support $[a2(Sj), c2(Sj)]$ and core $b2(Sj)$ are determined using the probability-possibility transformation applied to the map, as defined in [74].

Figure 3-5 shows a typical surface roughness model. Note the presence of two sets of possibility distributions and the roughness DNA. The case shown in Figure 3-5 corresponds to five-state modeling, $S1,\ldots, S5$, where $M = 5$. The simulation algorithm is shown by Algorithm 2-1 (Surface Profile Simulation). The calculation processes associated with Algorithm 2-1 are schematically illustrated in Figure 3-6. As seen in Algorithm 2-1, the simulation process acknowledges the roughness model (roughness DNA, possibility distributions, definitions of states). It initializes the roughness height $x(0)$ by a random number $r1 \in [0,1]$. After that, it calculates the maximum degree of belief and sets it as $\mu(0)$ from the fuzzy numbers assigned to the abscissa. It then calculates the values denoted as $x(SLj)$ and $x(SRj)$ for the state Si (this time, i = 1) dictated by the roughness DNA, as schematically illustrated in Figure 3-6 (a). The algorithm then chooses one of the calculated values that is the most far from $x(0)$ compared to the other and assigs it as $x(1)$. This way, the simulation algorithm continues its simulation process for all $i = 0,\ldots,N$. This results in simulated roughness heights $x(i)$, $i = 0,\ldots,N$. The simulated heights are linearly interpolated according to steps 20,…,26, as schematically illustrated in Figure 3-6 (b). As seen in Figure 3-6 (b), two consecutive heights $x(i)$ and $x(i+1)$ are linearly interpolated using a random value $r2 \in [u,v]$, where $u \leq 0$ and $v \geq 1$. This results in a new time series $w(k)$, $k = 0,\ldots,2$N, where $x(k) = x(k/2)$ if k is an even number and $w(k)$ is the interpolated value when k is an odd number. Therefore, the outcomes, $w(k)$, $k = 0,\ldots,2N$, are the simulated surface roughness models.

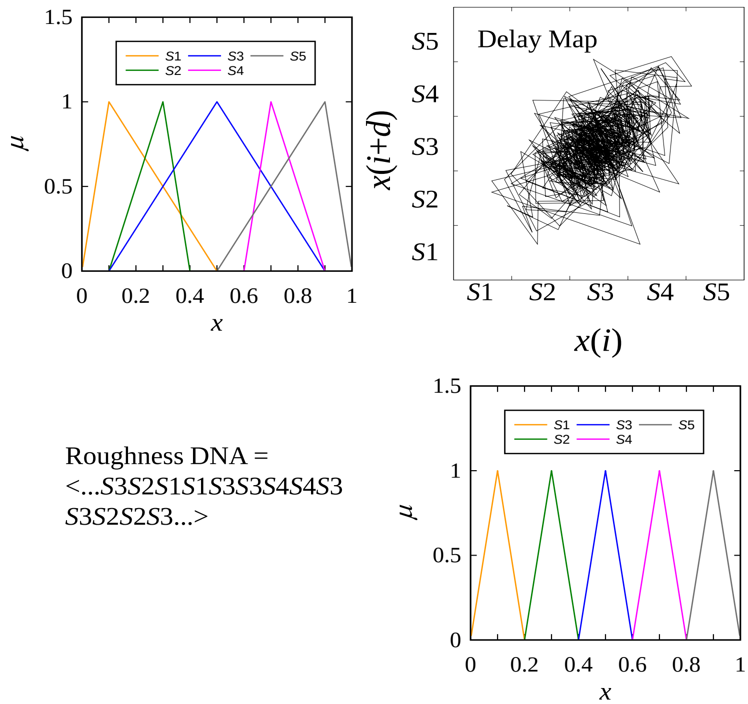


Figure 3-5. Constituents of roughness model

Algorithm 2-1. Surface Profile Simulation

Define

$$M, N, [u, v], (Sj \mid j = 1, \dots, M), DNA = (Si \mid i = 0, \dots, N)$$
$$\{(a1(Sj), b1(Sj), c1(Sj)) \mid j = 1, \dots, M\}$$
$$\{(a2(Sj), b2(Sj), c2(Sj)) \mid j = 1, \dots, M\}$$

Initialization

$x(0) \leftarrow r_1 \in [0,1]$

Calculate

For $i = 0, \dots, N-1$

For $j = 1, \dots, M$

$p(j) = \frac{x(i) - a1(Sj)}{b1(Sj) - a1(Sj)}, q(j) = \frac{c1(Sj) - x(i)}{c1(Sj) - b1(Sj)}$

$\mu(j) = \max\left(0, \min\left(p(j), q(j)\right)\right)$

End For

$\mu(i) = \max_{i=1,\dots,M}\left(\mu(j)\right)$

For $j = 1, \dots, M$

If $S(i+1) = Sj$ Then

$x(SLj) = a2(Sj) + \mu(i)\left(b2(Sj) - a2(Sj)\right)$

$x(SRj) = c2(Sj) - \mu(i)\left(c2(Sj) - b2(Sj)\right)$

If $|x(i) - x(SLj)| \geq |x(i) - x(SRj)|$ Then

$x(i+1) = x(SLj)$

Else

$x(i+1) = x(SRj)$

End For

End For

Output

$\{x(i) \mid i = 0, \dots, N\}$

Calculate

For $k = 1, \dots, 2N$

If $\left(\frac{k}{2}\right) \in \mathbb{N}$

$w(k) = x\left(\frac{k}{2}\right)$

Else

$t \leftarrow r_2 \in [u, v]$

$w(k) = x\left(\frac{k-1}{2}\right) \times (1-t) + x\left(\frac{k-1}{2} + 1\right) \times t$

End For

Output

$\{w(k) \mid k = 0, \dots, 2N\}$

The segment coupled with the simulation algorithm segment is the simulation system because the system is developed using the simulation algorithm. The simulation system must be kept simple to mitigate BD inequality. At the same time, it must be executable by many stakeholders, including those who belong to medium and small organizations.

Based on this consideration, the authors have used a spreadsheet-based computer program to build the simulation system; one of the user interfaces is shown in Figure 3-7. As seen in Algorithm 2-1, the user can set a delay and input the data points of surface roughness heights. The system simulates the surface roughness heights. The similarity between the simulated and real surface roughness heights can be compared in terms of time series, delay maps, and possibility distributions, as shown in Figure 3-7.

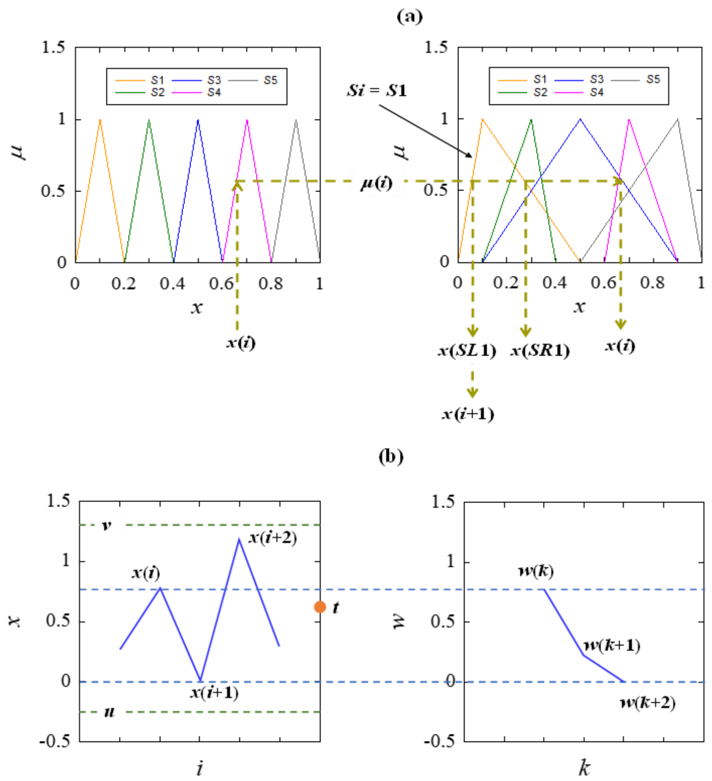


Figure 3-6. Illustrations of steps 3,…,27 of algorithm 2-1

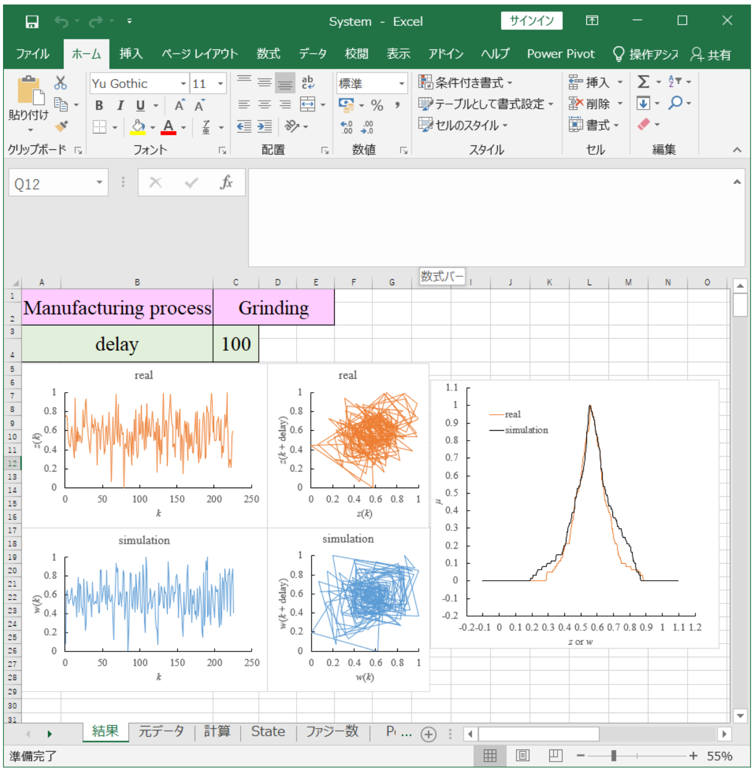


Figure 3-7. A user interface of the simulation system

The dataset is not complete by the segments shown in Figure 3-5, Figure 3-6, and Figure 3-7. The semantic annotation segment must be added to complete the dataset. This segment (semantic annotation) becomes the face of the dataset. It integrates other segments shown in Figure 3-5, Figure 3-6, and Figure 3-7. At the same time, the XML codes generated from the semantic annotation link the BD to a DT and other constituents of HCPS. The remarkable thing is that the semantic annotation segment manifests a concept map (a user-defined ontology of the issue considered). Before constructing the concept map, a set of proposition blocks (PBs) must be considered and expressed by natural language (here, English). The number of propositions depends on the individuals who construct them. Some PBS provides a general description, and others represent the dataset segments. PBs share some common concepts, resulting in a concept map.

Here, seven propositions-based PBs for constructing the semantic annotation segment of the dataset are considered. These PBs are listed in Table 3-1.

Table 3-1. Proposition blocks for concept mapping

Blocks	Propositions
PB1	Surface roughness profile heights of a manufacturing process called <name of the process> produce a delay map
PB2	Abscissa and ordinate of the delay map are divided by some fuzzy numbers
PB3	A Delay map entails a roughness DNA
PB4	Surface roughness profile heights are simulated using a simulation, producing simulate roughness profile heights
PB5	Simulated roughness heights can be further processed by linear interpolation
PB6	Simulation process and linear interpolations underlie a simulation algorithm
PB7	Simulation algorithm manifests a simulation system

The resulting concept map is shown in Figure 3-8. This map is the semantic annotation segment of the surface roughness dataset to be added to BD. The URL of this annotation is <https://cmapspublic2.ihmc.us/rid=1WYZG3ZMM-1SZ8JR8-41S0/surface-roughness%20datasets%20for%20big-data.cmap>. It can be accessed through the Internet. The annotation also carries the roughness DNA, fuzzy numbers of the return map, simulation algorithm, and the simulation system, which can be downloaded for reuse (for building DT of surface roughness). Alternatively, the XML data of the annotation can be generated to using it in the IoT-based enabler networks directly.

Note that other methods can also model chaotic data points like surface roughness heights: Markov chain, DNA-based computing, non-stationary Gaussian process, semantic modeling, and alike [84]. The roughness model, simulation algorithm, and simulation system segments accord with the modeling method. Even though the dataset segments are reconstructed according to the modeling method, the dataset structure remains valid. Thus, the dataset structure (semantic annotation, roughness model, simulation algorithm, and simulation system) serves as the metadata of the surface roughness.

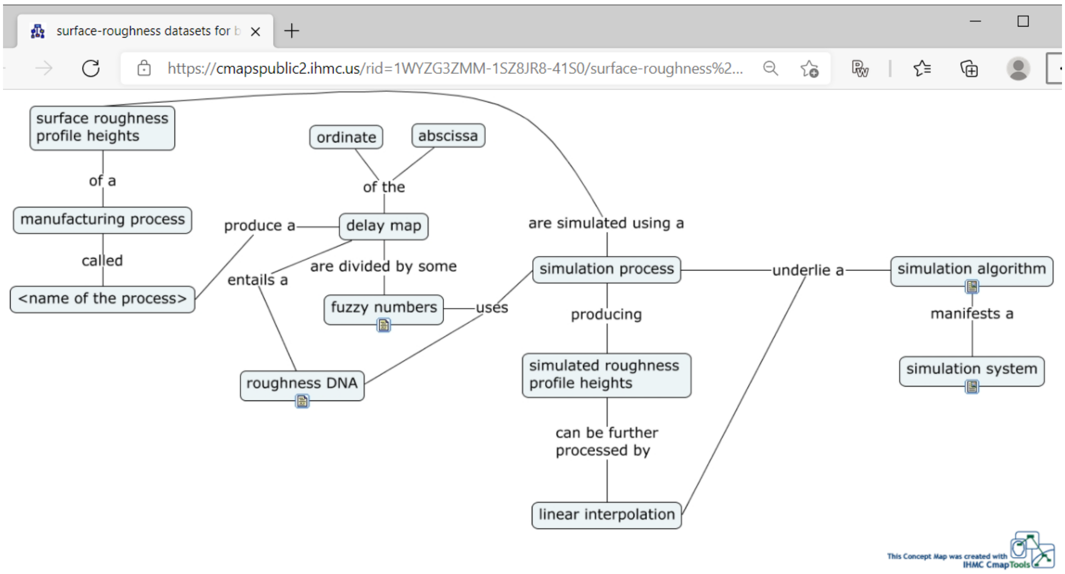


Figure 3-8. Semantic annotation (concept map) segment of the dataset

3.5 Security assurance

Security assurance in Industrial Internet-of-Thing (IIoT) is a critical issue. The datasets prepared by the presented data preparation method help assure security. To understand the interplay of security assurance and the presented dataset preparation method, consider Roughness DNA (one of the elements of the roughness model) and the provenance layer of the Semantic Web (SW), as schematically illustrated in Figure 3-9. The concept of SW has been proposed to smoothly exchange and reuse information among large information silos [84]. The main idea is to extend the potency of the Web with an analogous extension of the human cognitive process [32,33]. SW consists of four layers, namely, the syntax layer (XML, URI, and Unicode), the semantic layer (ontology and RDF), the provenance layer (rule, logic, proof, and trust), and the application layers [85]. The syntax layer encodes datasets to be exchanged. The semantics layer provides the meaning of the datasets. The provenance layer ensures the trustworthiness of the datasets for reuse. Finally, the applications layer hosts the applications by which the users (humans and other systems) exchange and reuse the information. The

XML codes produced from the semantic annotation (Figure 3-8) populate the syntax layer. Similarly, the semantic annotation itself provides the information for the semantics layers. Finally, some segments of the Roughness Model (Figure 3-5) can be used in the provenance layer. In this case, Roughness DNA can be input to the DNA Based Computing (DBC) [34,35] system to see the integrity of the datasets to be transferred to DTs of surface roughness. It has been shown that DBC effectively builds trust in the contents related to the manufacturing process. Consequently, DBC can help achieve a trustworthy exchange of content. This phenomenon is referred to as the "pragmatic adaptation of resources" from one working environment (BD) to another (DT). See [34,35] for more details. It is worth mentioning that as a consequence of the "biological station of manufacturing," representation of manufacturing data and knowledge using DNA-like information has earned a great deal of attention [7,36].

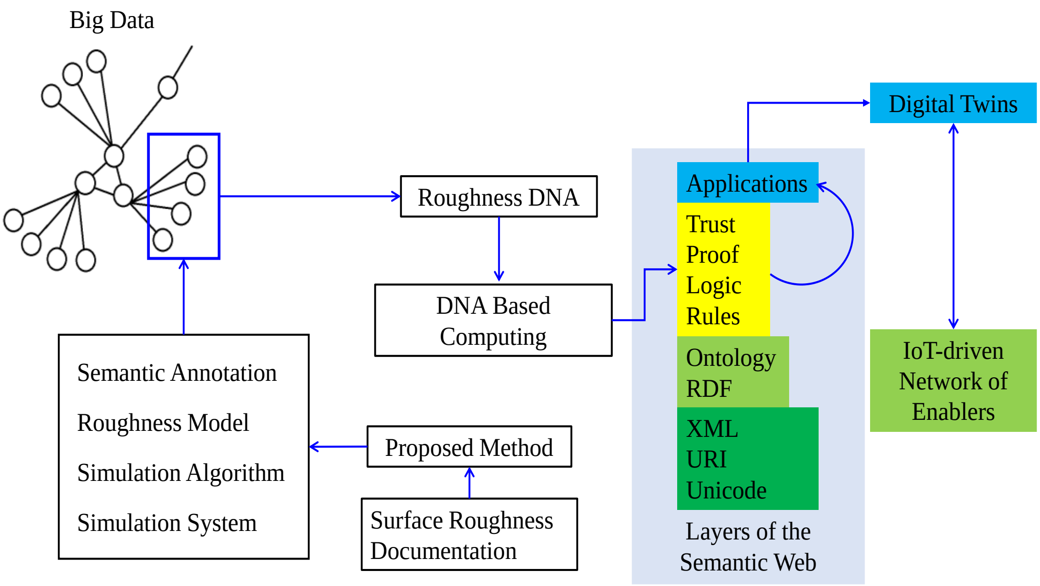


Figure 3-9. Aspects of security assurance.

.

Chapter 4: Machining Decision Making

Chapter 4 deals with Issue 2. This chapter described the developed BDA framework for the CV-EV-centric dataset. This system can support user-defined ontology and automatically produces XML-based datasets. The BD exploration system can extract relevant datasets prepared by the first system. The system uses keywords derived from the names of manufacturing processes, materials, and analyses- or experiments-relevant phrases (e.g., design of experiment). The third system can help visualize relevant datasets extracted by the second system using suitable methods (e.g., scatter plots and possibility distribution). The fourth system establishes relationships among the relevant control variables (variables that can be adjusted as needed) and evaluation variables (variables that measure the performance) combinations for a given situation.

In addition, it quantifies the uncertainty in the relationships. The last system can extract knowledge from the outcomes of the fourth system using user-defined criteria (e.g., minimize surface roughness and maximize material removal rate). The efficacy of the proposed BDA is demonstrated using a case study where the goal is to determine the right states of control variables of dry electrical discharge machining for maximizing material removal rate. It is found that the proposed BDA is transparent and free from BD inequality.

As we described in chapter 2, the BDA framework for smart manufacturing in this study included five interconnected systems developed using the JAVA™ platform, namely 1) BD preparation system, 2) BD exploration system, 3) data visualization system, 4) data analysis system, and 5) knowledge extraction system

It is considered that the framework and systems must be both human- and machine-friendly.

4.1 Big data preparation system

In the proposed framework, the first system is the BD preparation system. This system creates digital manufacturing commons related to CV-EV centric dataset. It also helps form BD for manufacturing processes. Another feature of that it explicitly shows the numerical data relevant to CV-EV combinations. As an example, for a specific machining process, we can identify different categories of information like source, machine type, tool, workpiece, etc. Then in our system, this information will be included in the metafile. Then, the system converts this meta file to an Extensible Markup Language (XML) file that is machine-readable. Then, it will be added to DMC, a cloud-based repository that creates our HCPS-friendly BD in our study. Figure 4-1 shows the main characteristics of the BD preparation system in the proposed BDA framework. In chapter 2, we described dataset preparation for the case of surface roughness of machining operation. This part describes the BD preparation system, which is developed for the comprehensive framework.

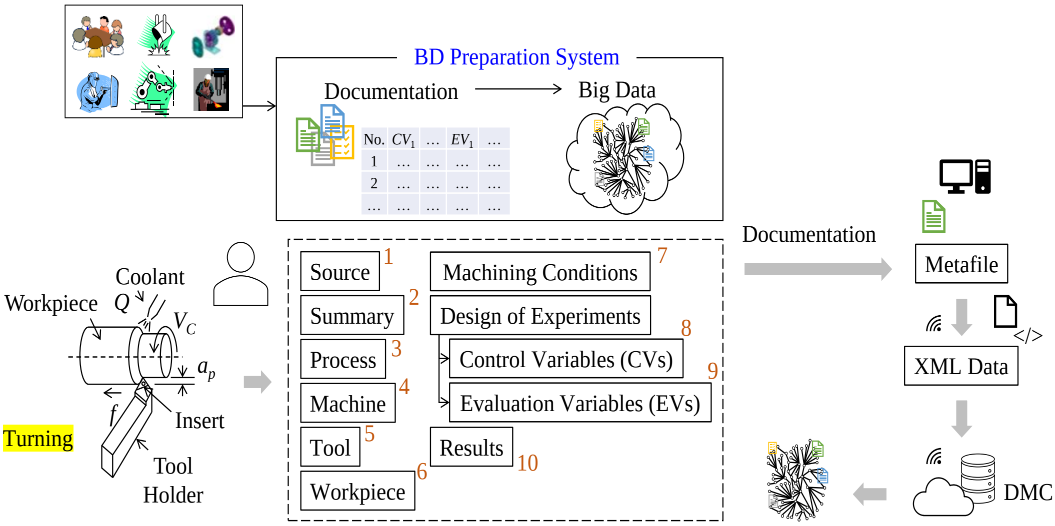


Figure 4-1. Main characteristics of the BD preparation system

4.2 Big data exploration system

The second system is the BD exploration system. It helps search the digital manufacturing commons in the BD and acquire the relevant datasets for the particular situation. BD exploration system consists of search and data acquisition. As we described, plenty of manufacturing processes are already included in DMC. As described, our system is a cloud-based repository. In the provided area in the developed framework, Then, in the provided area of the JAVA™- based developed system, BD Exploration System (BDES), it is possible to search the repository by writing some keywords and selecting and finding the related datasets in the repository. Figure 4-2 shows the BD exploration system developed in JAVA™.

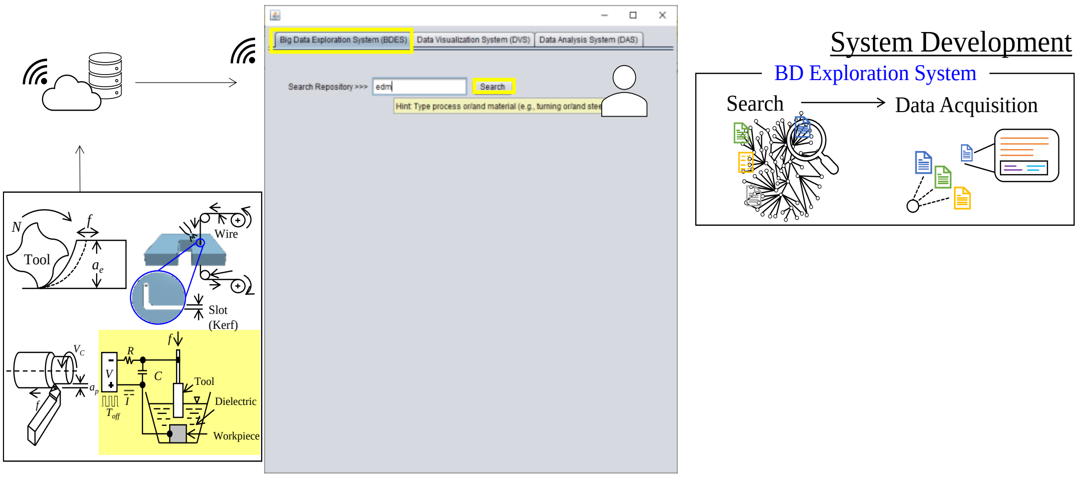


Figure 4-2. JAVA™- based developed BD exploration system

4.3 Data visualization system

The third system is the data visualization system. It represents relevant datasets using various visualization schemes. Provides a facility to visualize/represent the acquired CV-EV-centric datasets using different data visualization techniques (e.g., scatter plots, possibility distribution, and so forth). In this system, the user can choose a specific pair of CV-EV and an appropriate technique and visualize the datasets whenever needed. Figure 4-3 shows the screen print of the data visualization system. Figure 4-3 shows the screen print of the developed data visualization system. First, the user selects a set of CV-EV (Current-

MRR), which are available according to the BD exploration system. Then, possibility distribution is selected as the representation type, and the graph obtained out of the mentioned CV-EV pair is shown on the screen.

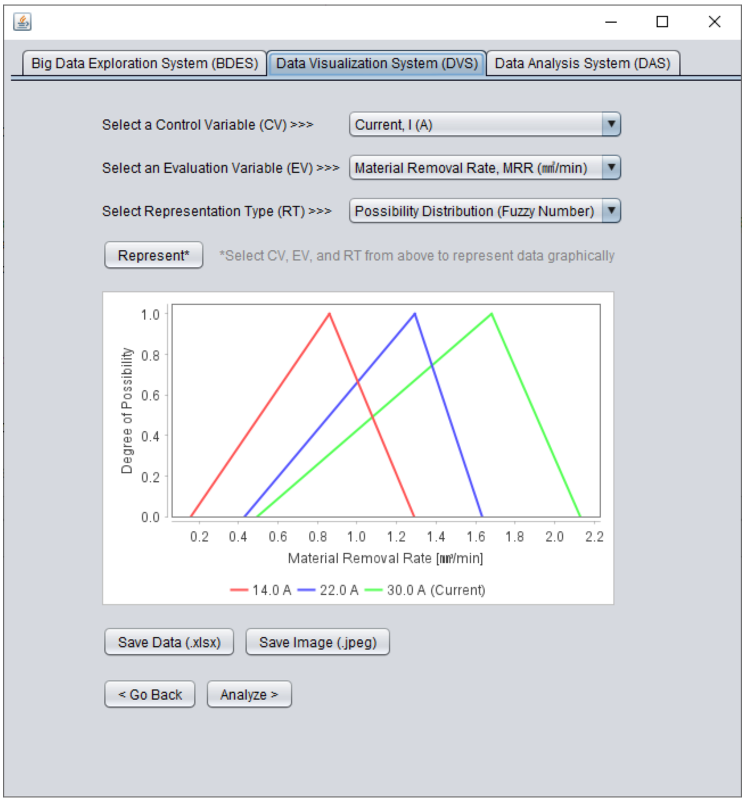


Figure 4-3. Screen print of the developed data visualization system

4.4 Data analysis system

The fourth system is the data analysis system. The main aim of this system is to analyze the CV-EV-centric dataset and find the relationships among all CV-EV combinations for a given situation. This will be possible using different computational methods like correlation and uncertainty analysis. This system can be helpful in the decision-making of manufacturing processes. Figure 4-4 shows the screen print of the data analysis system.

Furthermore, The outcomes of the analyses (correlation and uncertainty analyses) are displayed graphically, quantified by R-values, where R $\in$ [-1,1]. An R-value closer to ‘1’ indicates a strong direct relationship between the CV and

EV. An R-value closer to ‘-1’ indicates a strong indirect relationship between the CV and EV. For instance, shown in Figure 4-4, R-values for correlation and uncertainty analyses are 0.999 and 0.972, respectively. This means that the CV and EV entail a robust direct correlation associated with high uncertainty.

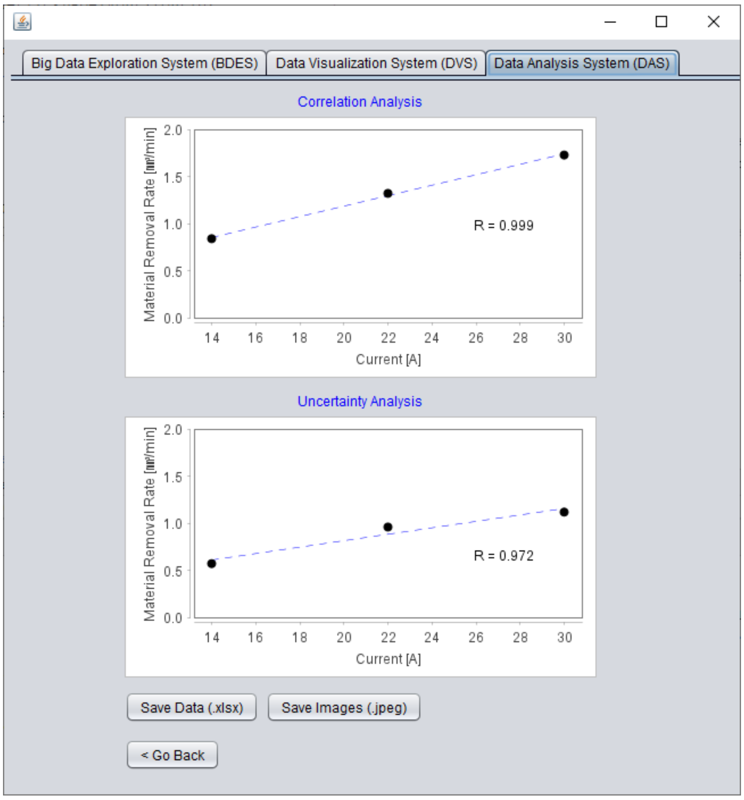


Figure 4-4. Screen print of the developed data analysis system

4.5 Knowledge extraction system

The last system is the knowledge extraction system. This system extracts rules, equations, and other forms of knowledge which can be related to manufacturing processes. Finally, this system directly interacts with IoT-based manufacturing enablers. It has to be mentioned that the users may deploy different techniques for processing the analysis outcomes and extracting underlying knowledge when the outcomes are gathered in the users’ vicinities. This fact makes the knowledge extraction system highly user-dependent.

4.6 Case study

In this part, the output of the proposed framework is analyzed. A case study is performed to validate the developed framework. For this purpose, a specific manufacturing process (dry EDM) is considered. Figure 4-5 shows the schematics of the dry EDM process selected as the case study for validation of the BDA framework. The related CV-EV-centric datasets of this process are also shown in this figure. For CVs, some of them mentioned in this figure are gas pressure (*P*), current (*I*), rotational speed (*N*), pulse off-time (T_{off}), Voltage (*V*), and for EVs (experimental results), some of them are material removal rate, tool wear rate, and surface roughness.

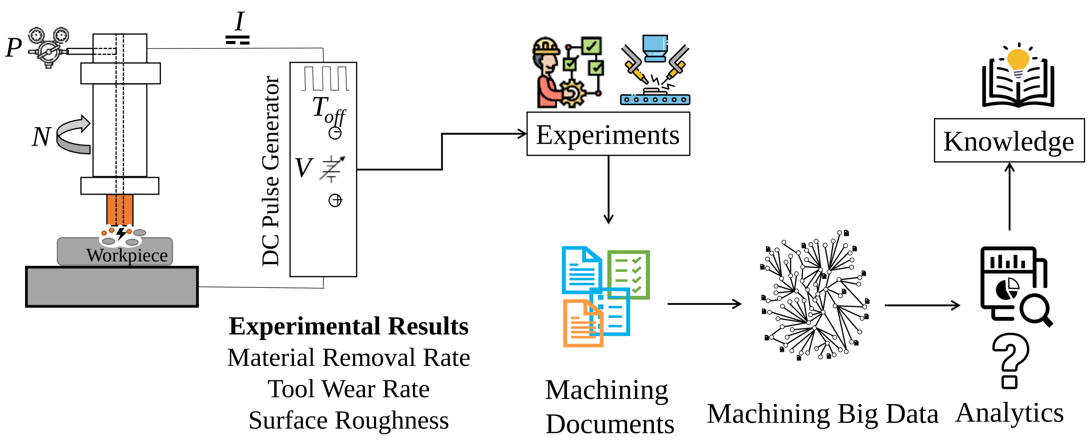


Figure 4-5. Schematics of dry EDM process selected as the case study

4.6.1 Big data preparation system and big data exploration system

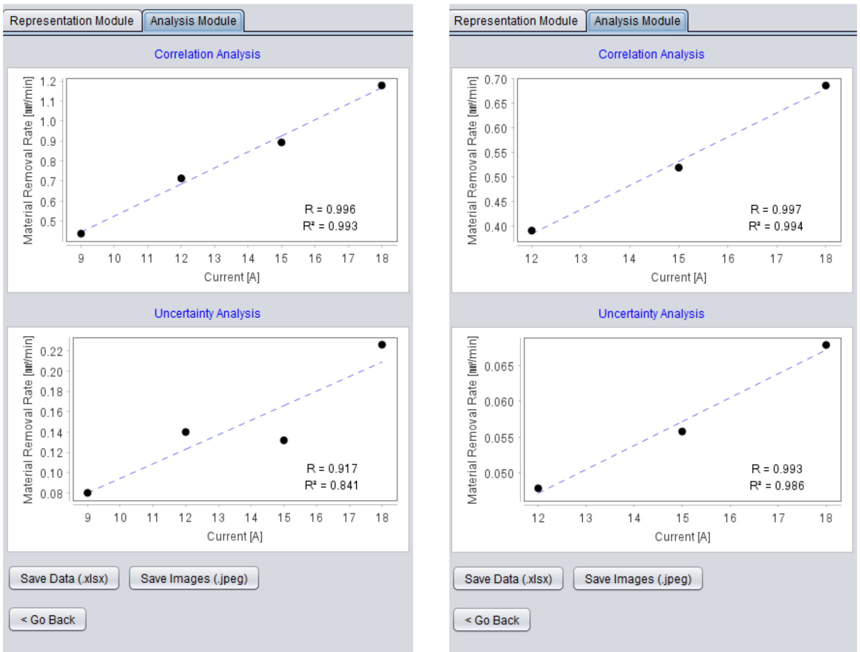
As we mentioned before, one of the primary purposes of this BDA framework is the mitigation of BD inequality and the digital divide. Rather than considering this issue in the designed BDA framework described previously, as mentioned in the introduction, inequality for access to the data is another important matter that must be considered. Nowadays, many sources of BD for scholarly outcomes exist. The BD sources of scholarly outcomes can be divided into primary BD and secondary BD. The primary BD is single publisher-managed BD. The secondary BD aggregates datasets from multiple primary BD sources like ScienceDirect and SpringerLink. In this case study, the secondary BD called Google Scholar is considered the source of documentation of past manufacturing activities. Then, the developed BD preparation system is used to prepare an HCPS-friendly dataset out of the related CV-EV-centric documentation. By using a process-based keyword search and developed BD exploration system, it is found that there are 20 related CV-EV-centric datasets with the term “dry EDM.” One of the main purposes of running the BDA framework for machining operations is extracting knowledge in terms of maximization/minimization of Evs. Accordingly, the process-based search must be more specific, for example, to material, etc. In the BD exploration system, specific materials are also searched: high-speed steel (HSS) and stainless steel (SS). Considering that five sets of CV-EV-centric datasets are found that can be used for the following systems of the proposed BDA framework. These five sources (S1[89], S2[90], S3[91], S4[92], S5[93]) related datasets are considered for analysis and decision-making. Table 4-1 shows each source's CVs, EVs, Number of trials, and workpiece material. In the developed system, each EV-CV combination must be analyzed separately for decision-making (in this case study, optimization of EVs). Due to the nature of the machining process, the user must define pre-determined optimization criteria for the visualization and analysis purpose, just *MRR* is selected as the goal EV, and the predetermined criteria for *MRR* optimization is maximization (MAX).

Table 4-1. CV-EV-centric dataset attributes for the case study of dry EDM

Source No. (S)	No. of Trials	Material	Control Variables (CV)	Evaluation Variables (EV)
S1	15	SS	V, I, T_{off}, P, N	*MRR, TWR, ROC, Z*
S2	39	SS	V, I, T_{off}, P, N, C_b	*MRR, TWR, OS*
S3	31	SS	I, T_{on}, D, N	*MRR, TWR*
S4	81	HSS	I, T_{on}, D, P, N, *Gas Type*	*MRR, SR, ROC*
S5	27	HSS	$I, T_{on}, D, P, N,$	*MRR, SR, ROC*
HSS: High-Speed Steel, SS: Stainless Steel (SS) *V*: Voltage (V), *I*: Current (amp), T_{off} : Pulse off time (µs), T_{on} : Pulse on time (µs), *P*: Gas Pressure (Mpa, bar), N: Spindle Rotational Speed (rpm), C_b: Shielding Clearance (mm) *MRR*: Material Removal Rate (mm^3/min), *TWR*: Tool Wear Rate (mm^3/min), *SR:* Surface Roughness (µm), *Z*: Achieved Depth (µm), *OS*: Oversize (%), *ROC*: Radial Overcut (µm)				

4.6.2 Data visualization system and data analysis system

As it is described, for the data visualization system, first, the specific CV-EV combination, then the representation type can be selected since *MRR* is selected as the selected EV, the visualization system work according to different CVs and *MRR* combinations for different sources. As described before, two types of visualizations are available: scatter plots and possibility distributions. For this case study, since obtaining possibility distribution is not possible for some datasets with fewer data, a scatter plot is selected as the representation type and followed that considered for the data analysis system. For the data analysis system, as it is described, correlation analysis (CA) and uncertainty analysis (UA) graphs together with R values will be available for this system. For example, Figure 4-6 shows the output of the data analysis system for S1 and S2 for *I-MRR*. Other CV-EV combinations can obtain similar data analysis output for other sources.



S1 S2

Figure 4-6. Example of obtained results from the data analysis system

4.6.3 Knowledge extraction system

As described before, according to predetermined criteria (maximization of *MRR* underlying dry EDM), this case study focuses on the *MRR* of the dry EDM process. According to the data analysis system's results, the knowledge extraction system must be performed following that obtained R values. Table 4-2 shows the R values related to each source for the case study of dry EDM. R values for both correlation analysis (CA) and uncertainty analysis (UA) outcomes of the data analysis system are presented. Based on these R values, four main points and their following points can be considered simultaneously for making decisions and extracting knowledge.

Table 4-2. Data Analysis: R Values Table for *MRR*

Type of Analysis	CA	UA	CA	UA	CA	UA	CA	UA	CA	UA
CV \ Source	S1		S2		S3		S4		S5	
V	-0.904	-0.75	-0.852	-0.957						
I	0.996	0.917	0.997	0.993	0.989	0.195	0.999	0.972	0.995	0.998
T_{off}	-0.653	-0.69	-0.661	-0.754						
T_{on}					0.989	0.071	0.978	0.993	0.523	0.593
D					0.222	0.181	0.986	0.99	0.944	0.987
P	0.93	0.926	0.369	0.965			0.902	0.863	0.996	0.798
N	0.993	0.826	1	0.957	0.707	-0.009	-0.419	-0.691	0.265	0.452
C_b			-0.397	0.585						
Gas Type							0.934	0.967		

Point No.1: The first issue is the visualization of the strength of correlation and uncertainty regarding their R values. For this purpose, a gradual coloring representation is presented. Figure 4-7 shows this gradual coloring representation for every CV-EV combination and source, pointing out both uncertainty and correlation analysis. According to this graph, the more filled color shows the higher R-value. This way, at least, it is possible to visually follow up on the strength of uncertainty and correlation trends.

Point No.2: Besides the visual inspection of the strength of uncertainty and correlation, another important factor can be understood from Fig 1. This is how CVs can affect EVs or uncertainty (direct or indirect). In Figure 4-7, Indirect effects are shown in yellow color, and direct effects are shown in green. Indirect (e.g., for CV_i) means if we want to maximize *MRR*, it is needed to minimize CV_i, and direct means if we want to maximize *MRR*, it is needed to maximize CV_i. These are important criteria for decision-making for the optimization of machining operations.

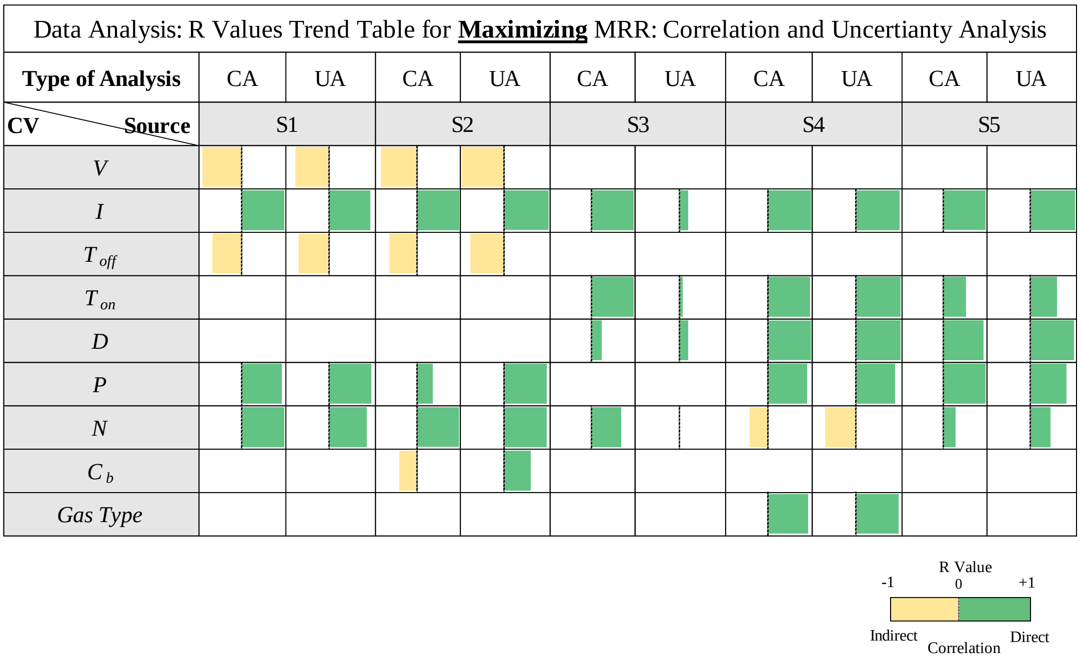


Figure 4-7. Screen print of R values trend table for maximizing *MRR*: correlation and uncertainty Analysis

Point No.3: another important issue is that for each source, which controls variables are significant for the maximization of *MRR*? First, we must consider the R values of correlation analysis. Then, considering their absolute values of them (|R|) and making some range categories, it is possible to categorize the strength of effectiveness of each control variable for our predefined purpose (maximization of *MRR*). The following criteria are defined for making the significance categories:

1- If for a specific CV, $|R|\geq 0.8$, then it is considered the most significant CV.

2- If for a specific CV, $0.4\leq|R|<0.8$, then it is considered the less significant CV.

3- If for a specific CV, $|R|<0.4$, then it is considered a non-significant CV.

Figure 4-8 shows the significance of CVs regarding R values of correlation analysis according to these criteria. In this table, we put symbols for each category to be inspected quickly for decision-making purposes. It has to be mentioned that Initially, just CVs categorized in the “most significant CVs” will be considered for decision-making.

Point No.4: As we mentioned before, one of the most important significances of this framework is the consideration of uncertainty. It implies that uncertainty can be inspected simultaneously in addition to correlation analysis. Considering this point, the main goal is decreasing uncertainty while choosing/selecting an optimal CV-EV relationship. In this case study, we first consider one specific EV for each analysis (*MRR* in this case study). Accordingly, the uncertainty mainly points out the variability of data for each CV and source.

As mentioned in Point No.1., the strength and direction of the effect of uncertainty (Regarding R values) could be visually inspected in Figure 4-8. The strength and direction underlie the amount of uncertainty for each source. Accordingly, less uncertainty is the best choice for decision-making purposes. In this study, our primary goal is to consider each source separately for optimization matters. After focusing on one source, we consider all CVs together to make a rule. In this way, first, we apply all CVs optimized levels for one source. Consequently, the variability of data and the underlying uncertainty will decrease gradually. This shows that for this way of optimization, the main goal of decreasing uncertainty will be taken care of accordingly.

Furthermore, considering these four points, the knowledge-extraction process is as follows:

According to Figure 4-8, for each source, first, the most significant CVs are determined (CVs which are shown in green colors symbol). Then, according to Figure 4-7 and the direction of each CV's effect, a criterion for each of them is written. Max (maximize) for those who affect the EV directly and min (minimize) for those affected indirectly. For example, for S1, for maximization of *MRR*, considering the correlation analysis, and the most significant CVs, it is determined that if we maximize Current (I), Gas pressure (P), and Rotational Speed (N) and

Minimize Voltage (*V*), the best results could be obtained. Table 4-3 shows these defined max and min terms for all sources.

CV Signifcance			
Most significant ($\|R\| \geq 0.8$)	Less significant ($0.4 \leq \|R\| < 0.8$)	Non significant ($\|R\| < 0.4$)	Unavailable
✓	!	✗	

Data Analysis: R Values Significance Table for **<u>Maximizing</u>** MRR - Correlation Analysis					
Type of Analysis	CA				
CV \ Source	S1	S2	S3	S4	S5
V	✓	✓			
I	✓	✓	✓	✓	✓
T_{off}	!	!			
T_{on}			✓	✓	!
D			✗	✓	✓
P	✓	✗		✓	✓
N	✓	✓	!	!	✗
C_b		✗			
Gas Type				✓	

Figure 4-8. Screen print of the CV significance table regarding R values

According to Table 4-3, we can extract rules for each study which are the extracted knowledge of our BDA framework. These rules are:

Rule No.1 (S1): Maximize *I*, *P*, *N*, and Minimize *V* for the Maximization of *MRR*.

Rule No.2 (S2): Maximize I, N, and Minimize V for Maximization of *MRR*.

Rule No.3 (S3): Maximize I, Ton for Maximization of *MRR*.

Rule No.4 (S4): Maximize I, Ton, D, P, and Gas Type for Maximization of *MRR*.

Rule No.5 (S5): Maximize I, D, P for Maximization of *MRR*.

These rules can show us how we can select a desirable CV set for optimizing a machining evaluation variable.

Table 4-3. Decision-making rules out of correlation analysis

Source \ CV	*V*	*I*	T_{off}	T_{on}	*D*	*P*	*N*	C_b	*Gas Type*
S1	Min	Max				Max	Max		
S2	Min	Max					Max		
S3		Max		Max					
S4		Max		Max	Max	Max			Max
S5		Max			Max	Max			

4.6.4 Validation

Based on obtained rules for each source, we tried to find the optimized level/levels in the dataset and find whether, by meeting these criteria for each source, it meets the maximization criteria. First, these rules are applied to the relevant dataset to find the optimized level according to obtained rules. If there is just one related *MRR*, that level is selected. Otherwise, if there are several experiments related to those criteria, the maximum *MRR* among them is selected. Accordingly, we put the obtained *MRR* value by applying maximization rules for each source.

For validation purposes, it is necessary to compare this value with the maximum available *MRR* in the dataset. Since we already have the CV-EV-centric dataset and related *MRR* for each experiment. Table 4-4 shows the relevant *MRR* value considering the dataset's optimization rule and maximum *MRR*, respectively. The results for S1, S2, and S5 show that the optimized levels found are entirely in line with the maximum *MRR* in the dataset. For S3 and S4, the optimized level *MRR* is not already included in the dataset. For these two tests, since this CV combination is not available in the Design of Experiments (DOE) and the dataset for the related source, a predictive statistical approach can help to predict the obtained value based on the available dataset. Results from these validation tests show that our developed framework works efficiently for the optimization of machining operations.

Similarly to this case study and validation performed, other manufacturing processes can be evaluated and optimized using this framework.

Table 4-4. Validation results

Source / Test	Extracted Rule	*MRR* value based on the extracted rule	Maximum *MRR* value in the dataset
S1	Rule No.1	1.497	1.497
S2	Rule No.2	0.811	0.811
S3	Rule No.3	N/A	9.9425
S4	Rule No.4	N/A	3.77
S5	Rule No.5	5.31	5.31

.

Chapter 5: Dealing With Graphical Datasets

Chapter 5 deals with the following issue. BD is often visualized using several two-dimensional plots. These plots are then used to make a decision informally. Consequently, the research question is how to make formal decisions by two-dimensional computing plots, not numerical data. Accordingly, a BDA tool is developed to compute two-dimensional plots generated from BD. The efficacy of the tool is demonstrated by applying it to assess sustainability in terms of Sustainable Development Goal (SDG) 12 (responsible consumption and production). Regarding SDG 12, functional, economic, and environmental issues of engineering materials play a vital role. Accordingly, three two-dimensional plots generated from BD of engineering materials are computed using the proposed analytics. The plots refer to six criteria (strength, modulus of elasticity, cost, density, CO2 footprint, and water usage). The proposed analytics correctly rank the given materials (mild steel, aluminum alloys, and magnesium alloys).

5.1 Sustainable development goals and big data

According to the United Nations, sustainability means meeting current needs without jeopardizing the potential of fulfilling future needs [94]. This is a broad mission. The UN has divided this mission into seventeen goals known as Sustainable Development Goals (SDGs), as follows [95]: 1) No Poverty, 2) Zero Hunger, 3) Good Health and Well-being, 4) Quality Education, 5) Gender Equality, 6) Clean Water and Sanitation, 7) Affordable and Clean Energy, 8) Decent Work and Economic Growth, 9) Industry, Innovation, and Infrastructure, 10) Reduced Inequality, 11) Sustainable Cities and Communities, 12) Responsible Consumption and Production, 13) Climate Action, 14) Life Below Water, 15) Life on Land, 16) Peace and Justice Strong Institutions, and 17) Partnerships to achieve the Goal. These goals can be achieved by meeting the preset targets and indicators [34]. The activities associated with the targets and indicators generate an information silo consisting of a vast array of unstructured, semi-structured, and structured datasets distributed all around the globe. This information silo is often referred to as BD.

However, BD means horizontally networked yet independent data systems containing a vast number of datasets [40,41]. It requires a scalable architecture for efficient storage, manipulation, and analysis. Apart from its volume, its major characteristics are veracity, value, volatility, and validity. Veracity means the accuracy of the datasets. Value deals with generating economic or social wealth from any dataset. Volatility means the tendency for data structures to change over time. Validity means the appropriateness of the datasets for their intended use. The concerned national and international organizations (e.g., International Organizations for Standardization (ISO) and the National Institute of Standards and Technology (NIST)) have been offering vendor-neutral conceptual definitions [98], taxonomies [99], requirements, and usages [44,45], security and privacy [100], reference architectures [102], standardization roadmap [103], and adoption and modernization schemes [104] for BD so that it benefits its stakeholders without causing the phenomenon called BD equalities [105]. Some of the relevant articles that deal with the interplay of BD and SDGs are briefly described in the next section.

5.2 Sustainable development goals and big data

Now, as far as BD-driven value creation is concerned, the relevant datasets must help make decisions. In this case, data analytics are used. In most cases, the analytics help visualize the relevant datasets using some two-dimensional plots. These plots are then used to decide a course of action for achieving predefined objectives. For a better understanding, consider the two-dimensional plot shown in Figure 5-1. This plot (a segment of BD on engineering materials) shows the relative positions of three types of materials (mild steels, aluminum alloys, and magnesium alloys) in terms of two material properties (tensile strength and Yang's modulus). A visual inspection of the plot reveals that mild steels outperform aluminum and magnesium alloys in terms of strength and Yang's modulus. Therefore, if materials with high strength and Yang's modulus are needed, then mild steels should be chosen out of three types of materials. This argument is valid if tensile strength and Yang's modulus are equally important. This kind of decision-making may not always be as easy as described above. For example, if the strength and bending ability of a material are not equally important, then how do we compare aluminum and magnesium alloys?

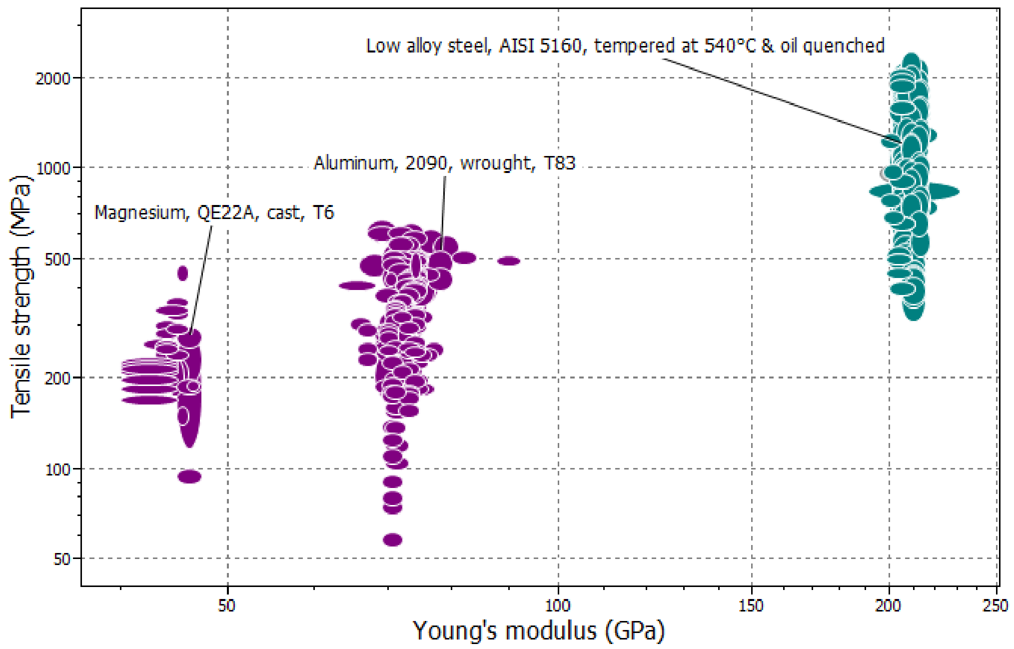


Figure 5-1. Visualization of relevant datasets from BD of engineering materials

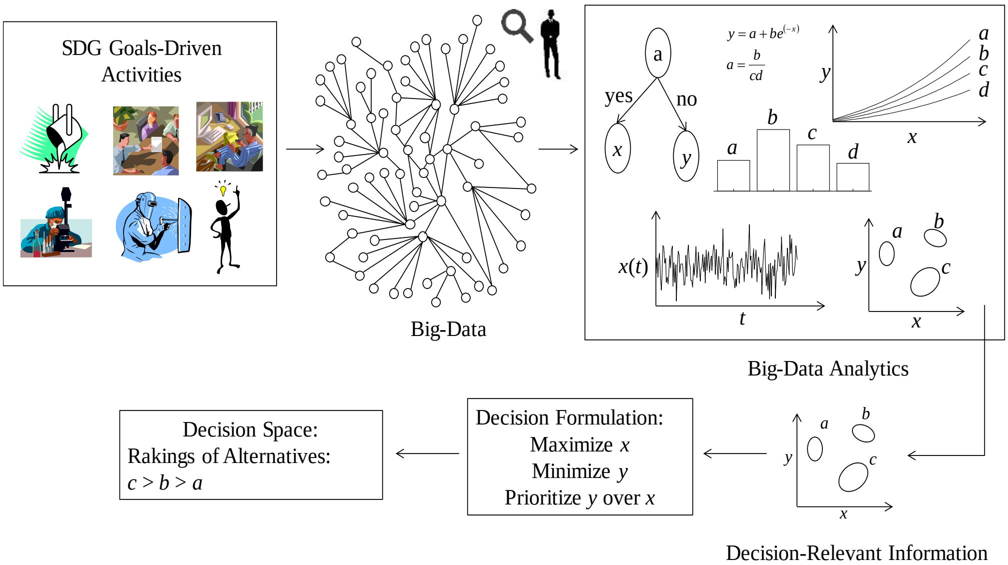
Thus, a formal decision-making approach must be employed to ensure a fair and rational comparison in which the decision-relevant information is graphical (e.g., a set of two-dimensional plots). A general outline of the said decision-making approach is as follows:

The first step is to set the criteria-based objective functions (e.g., maximize tensile strength, minimize Yang's modulus, minimize CO2 emission, etc.). Naturally, this step depends on the underlying situation. The second step is to compute the compliances of the alternatives with respect to criteria-based objective functions. The final step is to consider the importance or weight of each criterion and aggregate compliances so that each alternative receives a decision score for the final decision-making. The remarkable thing is that while performing the second step—while computing compliances—the decision-making approach must be able to compute the graphical information (e.g., two-dimension plots similar to that shown in Figure 5-1) rather than numerical data. Consequently, the conventional decision-making approaches where the primary decision-relevant information is numerical data are not applicable here. This necessitates a new breed of methods and tools for making rational decisions wherein the graphical information is formally computed instead of numerical information. Based on this consideration, this article is written. In particular, decision-making aspects centering on SDG 12 (responsible consumption and production) are focused.

5.3 Decision-making method and tool

This section presents a decision-making method where the decision-relevant information is two-dimensional plots rather than numerical data. The relevant mathematical entities, as well as the computing tool, are also presented in this section. This method and tool are collectively referred to as the decision-making approach.

Figure 5-2 schematically illustrates the proposed decision-making method. It consists of six modules. The first module is denoted as SDG-driven activities, where numerous stakeholders perform activities to achieve SDGs. The second module is denoted as BD, where a vast array of unstructured, semi-structured, and structured datasets collected from the SDG-driven activities are stored for making informed decisions. The third module is denoted as BDA. The analytics search the BD and helps visualize the relevant datasets. The fourth module is denoted as decision formulation, where the decision-maker sets the criteria and their importance. Each criterion is set either in the maximization or minimization format. For example, if social security is a target or indicator of sustainability, then its decision criterion can be "maximize social security." If CO2 emission is a target or indicator of sustainability, then its decision criterion can be "minimize CO2 emission." In addition to setting the criteria, the importance of the criteria is set either in numerical form (setting relative weights of the criteria) or in linguistic form (e.g., criterion x is more important than criterion y). The last module is denoted as decision computation. In this module, decision-relevant information (in this case, two-dimensional plots) is computed to see how well the alternatives comply with the decision criteria. The values of the compliances are then further processed using the importance of the criteria, which results in a ranking list of the alternatives. Finally, a decision is made using the ranking list. It is worth mentioning that Ullah and Noor-E-Alam [106] provided a framework and tool for making decisions using graphical information. The method shown in Figure 2 resembles their work as far as data visualization is concerned.



.

Figure 5-2. Proposed decision-making framework

The concept of compliance can be used to perform decision computation. The mathematical settings for calculating compliance of a numerical range with a given fuzzy number are described in [107]. In this article, a simplified formulation of the compliance analysis is considered based on the formulation shown in [107].

For better understanding, consider an arbitrary two-dimensional plot, as shown in Figure 5-3. Let $X1$ and $X2$ be two parameters associated with the decision-making process and A, B, and C be the alternatives, as schematically illustrated in Figure 3. The relative positions of the alternatives are shown by a two-dimensional plot where the abscissa is $X1$, and the ordinate is $X2$. Each alternative can be represented by some intervals in the $X1$ and $X2$ directions. Each interval complies with the maximization and minimization functions. For example, let $[a, b] \in \Re$ be the scale of $X1$, as shown in Figure 5-3. Here, $x1 \in [a, b]$. The corresponding minimization and maximization functions, denoted as $Min(X1)$ and $Max(X1)$, are as follows.

$$Min(X1) = \min\left(\max\left(\frac{b-x1}{b-a}, 0\right), 1\right) \quad (5\text{-}1)$$

$$Max(X1) = \min\left(\max\left(\frac{x1-a}{b-a}, 0\right), 1\right) \quad (5\text{-}2)$$

For example, consider the case shown in Figure 5-3, where one of the possible intervals underlying A is given as $[p, q]$. In order to calculate the compliance, the following functions for maximization and minimization can be defined:

$$C_{max} = \frac{(p+q)-2a}{2(b-a)} \quad (5\text{-}3)$$

$$C_{min} = \frac{2b-(p+q)}{2(b-a)} \quad (5\text{-}4)$$

For example, consider that $[a, b] = [0, 100]$ and $[p, q] = [30, 70]$. In this case, $C_{max} = C_{min} = 0.5$. This means that the internal [30, 70] equally complies with maximization and minimization functions defined in the universe of discourse [0, 100]. Consider that $[a, b] = [0, 100]$ and $[p, q] = [20, 50]$. In this case, $C_{max} = 0.35$ and $C_{min} = 0.7$. This means that [20, 50] complies with minimization more than it complies with maximization in the universe of discourse [0, 100]. Out of C_{max} and C_{min}, one is considered for the decision-making purpose. The explanation is as follows. For example, if $X1$ represents CO2 emission, then it (CO2 emission) must be minimized. As a result, the criterion becomes “minimize CO2 emission.” As such, the value of C_{min}, not $C_{max,}$ is considered for the decision-making purpose. Similarly, if $X1$ represents productivity, then it (productivity) must be minimized. As such, the value of C_{max}, not C_{min}, is considered for the decision-making purpose. Thus, if a given criterion refers to maximization, then its compliance C is equal to C_{max}. Alternatively, if a given criterion refers to minimization, then its compliance C is equal to C_{min}. This way, each alternative returns a set of compliances, $C1,\ldots,Cn \in [0,1]$, because each alternative entails multiple intervals. A set of compliances can induce a possibility distribution in the form of a triangular fuzzy number. The mathematical procedure to induce fuzzy numbers from a given set of numerical values is presented in [108]. For a lucid description of the induction process, refer to Appendix A in [109].

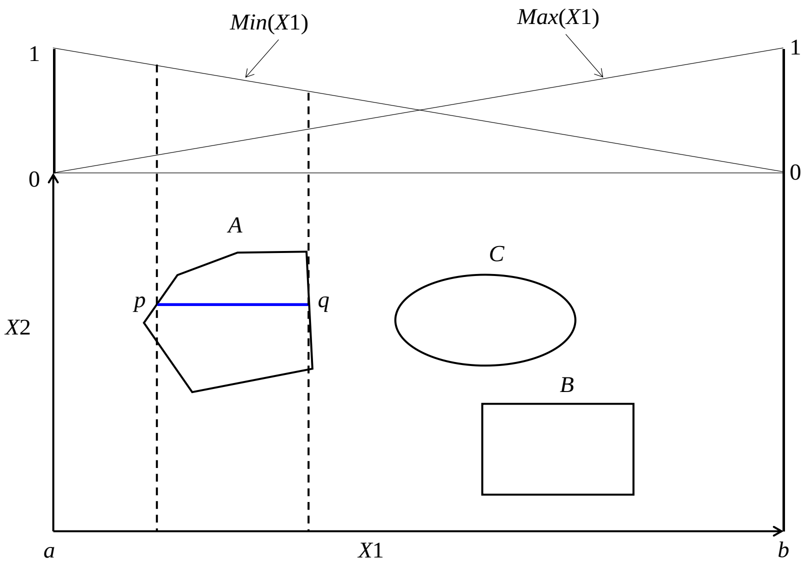


Figure 5-3. Compliance calculation for an interval

Let $[t_1, t_3]$ be the support and t_2 be the core of the induced fuzzy number denoted as T. The compliance of T with the idealistic possibility distribution of compliances is the compliance of an alternative with respect to the criterion. This aggregated compliance is thus the decision score (denoted as D) of the alternative with respect to the criterion. Now, the idealistic possibility distribution can be defined in numerous ways. One of the obvious ways is to define it by a function that ensures the maximization of the possibility of compliance in the universe of discourse of [0, 1] as marked by "Ideal" in Figure 5-4. As such, the decision score is calculated as follows:

$$D = \frac{rt_3 - st_1}{t_3 - t_1} \qquad (5\text{-}5)$$

The parameters r and s in Equations (5) are defined as follows.

$$r = \frac{t_3}{1 + (t_3 - t_2)} \qquad (5\text{-}6)$$

$$s = \frac{t_1}{1 - (t_2 - t_1)} \qquad (5\text{-}7)$$

For example, consider that t_1 = 0.2, t_2 = 0.4, t_3 = 0.7. This results in r = 0.7/1.4 = 0.5, s = 0.2/0.8 = 0.25, D = (0.35-0.05)/0.5 = 0.6. If t_1 = 0, t_2 = t_3 = 1, then r = 1, s = 0, D = 1. This means that when the induced possibility distribution (aggregated compliance) takes the shape of the idealistic possibility distribution (Figure 5-4), the decision score becomes a unit; otherwise, it is less than a unit.

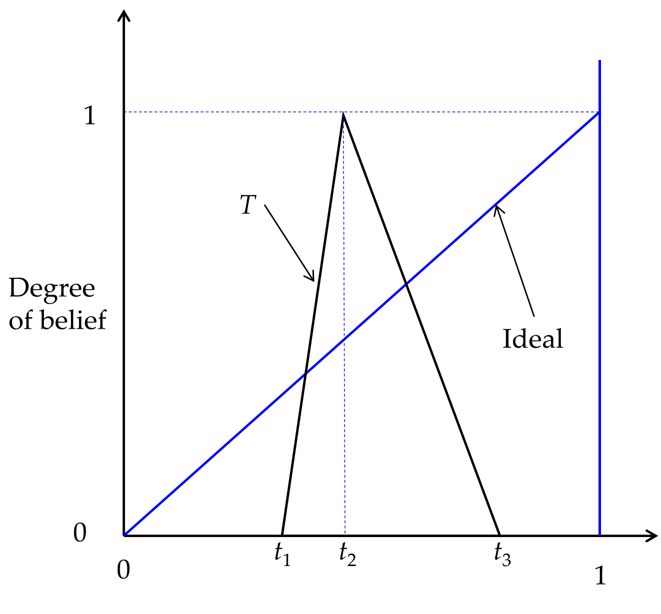


Figure 5-4. Calculating decision score

The abovementioned decision computation needs a special computing tool. The tool must be able to extract the intervals for an alternative from a given two-dimensional plot [110]. The pseudocodes of the tool are as follows:

Step 1:	Start
Step 2:	Upload a two-dimensional plot
Step 3:	Set the scales for the abscissa and ordinate of the uploaded plot (i.e., minimum and maximum values defining the scales of the abscissa and ordinate of the plot)
Step 4:	Select one of the directions (either abscissa or ordinate) for extracting intervals
Step 5:	Select one of the alternatives for extracting intervals
Step 6:	Drag the mouse in the selected direction and extract several intervals covering the region of the selected alternative and direction
Step 7:	Repeat Step 5 for all other alternatives
Step 8:	Go to Step 3 and select the other direction
Step 9:	Repeat Steps 4,…,6
Step 10:	Output all intervals extracted in the above steps
Step 11:	End

Based on the above pseudocodes, a decision tool is developed that runs on Windows™ operating system. Figure 5-5 and Figure 5-6 show two of the user interfaces of the tool developed. Accordingly, a decision-maker first uploads a two-dimensional plot. After that, the decision-maker sets the scale of the abscissa and ordinate. The scale may not have to be the scale of the plot. For example, the arbitrary case shown in Figure 5-5 shows that the decision-maker uploaded a two-dimensional plot (stiffness versus cost) of engineering materials. This plot shows the relative positions of five materials: aluminum, cast iron, stainless steel, composites (GFRP), and composites (CFRP). Thus, these five types of materials are the alternatives for this particular case. Each alternative can be represented by some intervals in the abscissa and ordinate directions. The instance shown in Figure 5-6 is a scenario when the decision-maker extracts the intervals for representing the cost of cast iron.

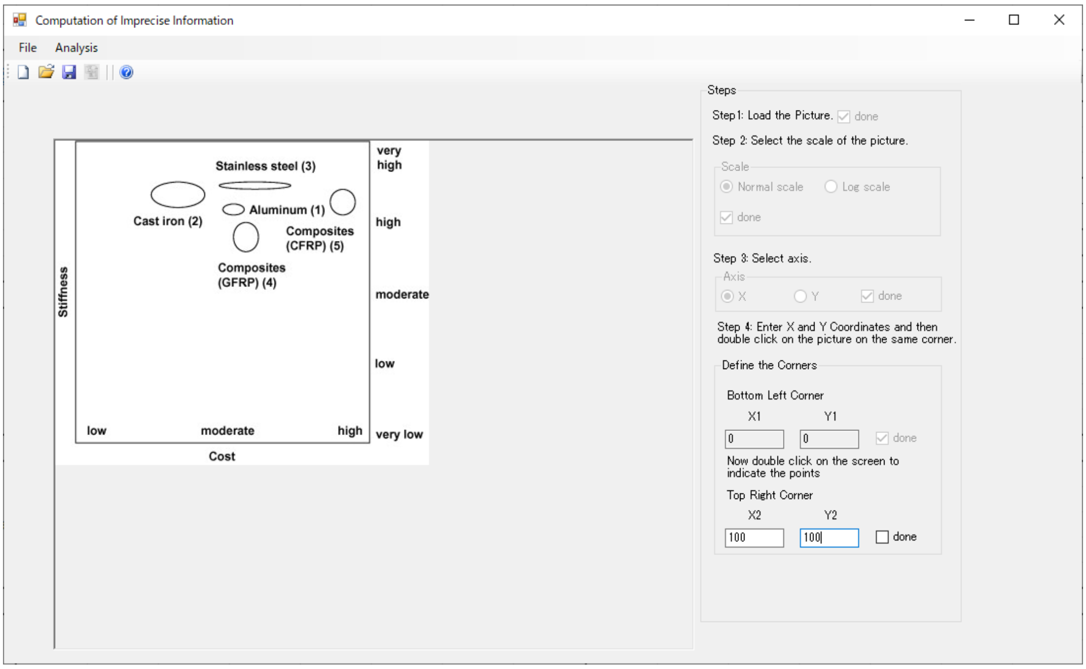


Figure 5-5. The user interface of the decision tool

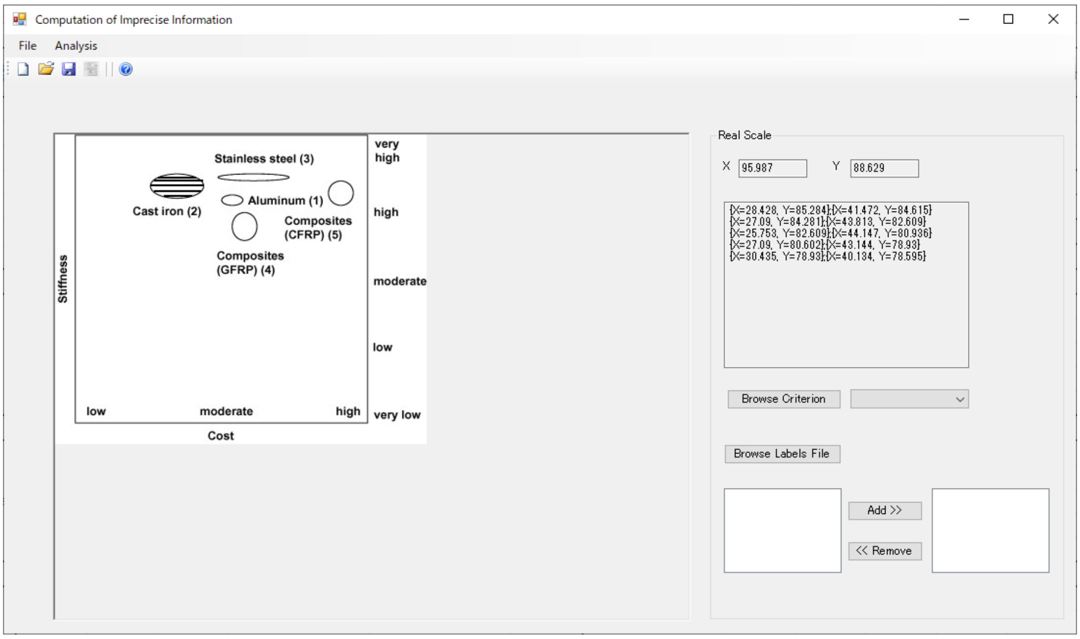


Figure 5-6. An example of extracting intervals from an alternative

5.4 Results and discussions

This section presents results and a discussion considering two case studies. The first case study is conducted to see whether or not the proposed decision-making approach produces reliable results even though the relevant datasets presented by two-dimensional plots exhibit a great deal of uncertainty. The other one is conducted to see whether or not the proposed decision-making approach can be used for multiple-criteria decision-making, even though the relevant datasets presented by several two-dimensional plots exhibit a great deal of uncertainty. Both case studies focus on SDG 12 (responsible production and consumption). For assessing the fulfillment of SDG 12, BD relevant to engineering materials becomes an important issue. The description is as follows.

Fulfilling SDG 12 means scanning the whole product life cycle to identify the harmful factors (carbon emissions and resource depletion) and eliminate those. As such, SDG 12 is highly correlated to the notion of material efficiency. It (material efficiency) deals with yield improvement, downsizing and lightening, cost reduction, and reduction of the CO2 footprint of primary material production [68-71]. The remarkable thing is that material efficiency is more effective than energy efficiency. (Energy efficiency deals with direct energy consumption while manufacturing products [68-71]). As a result, tracking all information regarding the characteristics and properties of a vast array of engineering materials has become important for ensuring SDG 12 and beyond. For this, a new concept denoted as materials passport has been introduced. The explanation is as follows.

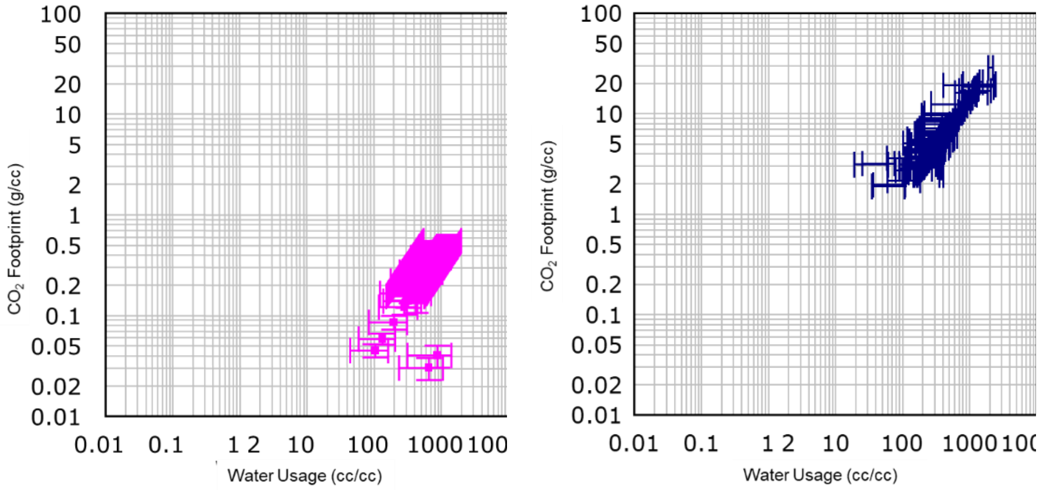
Since a product includes metals (including precious and rare earths), ceramics, polymers, and natural materials, myriad materials processes entail the fabrication of the parts of a product. Furthermore, the reasons and requirements behind using specific materials and their circularity must be known beforehand. Otherwise, the sustainability assessment cannot be performed as expected. In order to deal with material efficiency in a more global context (e.g., circular economy), a concept called material passport has been introduced. The material passport needs material-centric information collected from upstream activities (mining, trade, smelting/refining) and downstream activities (trade, component manufacturing, contract manufacturing and assembly, and end-using). While improving the sustainability of a product, its constituent materials and percentages, and the relevant manufacturing and assembly/service processes-relevant datasets stored in its material passport can be used. Furthermore, the eco-indicators of materials (CO2 emission of primary production of constituent materials, resource depletion) and other governance issues (whether or not forced labor is used in the upstream and downstream activities) populate material passports. Therefore, material-centric BD extracted from the material passports of various products become essential information for ensuring SDG 12 and other SDGs.

Nevertheless, in most cases, a relevant subset of BD is visualized using two-dimensional (scatter) plots. These plots dominate the underlying decision-making processes ensuring better fulfillment of SDG 12. In such cases, the decision-making method and tool presented in the previous section become instrumental.

5.4.1 Case study 1

This case study is conducted to confirm whether or not the proposed decision-making method and tool produce reliable results. In particular, the datasets regarding two eco-indicators (CO2 footprint and water usage) of two types of materials (wooden materials and polymers) are considered.

From the BD of engineering materials, the datasets regarding CO2 footprint and water usage of 447 wooden materials and 244 polymers are shown using two two-dimensional plots, as shown in Figure 5-6. See [113] for more details. The datasets are shown in Figure 5-7 by two scatter plots where the ordinates present Water Usages (cc of water/cc of material) and the abscissas present CO2 emission (grams of CO2/cc of material) of the primary production of the respective materials. To be more specific, consider the following case where two families of materials, namely, polymers and wooden materials, are ranked in terms of CO2 footprint and water usage. As far as water usage (as a measure of resource depletion) is concerned, both materials can be ranked equally. However, on the other hand, as far as CO2 footprint is concerned, wooden materials are far better than polymers.



(***a***) wooden materials (***b***) polymers

Figure 5-7. BD regarding the sustainability of two families of materials. (a) wooden materials. (b) polymers [115].

Figure 5-8 shows the user interface of the decision tool (Figure 5-5 and Figure 5-6) when the decision-maker extracts the ranges representing the uncertainty in water usage. Even though the water usage is plotted using a logarithmic scale starting from 0.01 cc of water/cc of material to 10000 ccs of water/cc of material, the range extraction process rescales the plot to [0, 100]. (Here, “cc” means cubic centimeter.) Therefore, each extracted range becomes a segment of [0, 100]. If preferred, the decision-maker can use other scaling schemes. For example, Figure 5-9 shows the extracted ranges representing the uncertainty in water usage of wooden materials. Similarly, water usage and CO2 footprint ranges are also collected for the rest of the cases.

Table 5-1 and Table 5-2 list the values of the decision-making parameters relevant to decision computation. As seen in Table 5-1 and Table 5-2, the decision score (*D*) values concerning the criterion "minimize water usage" for both groups of materials are comparable (0.395 and 0.410).

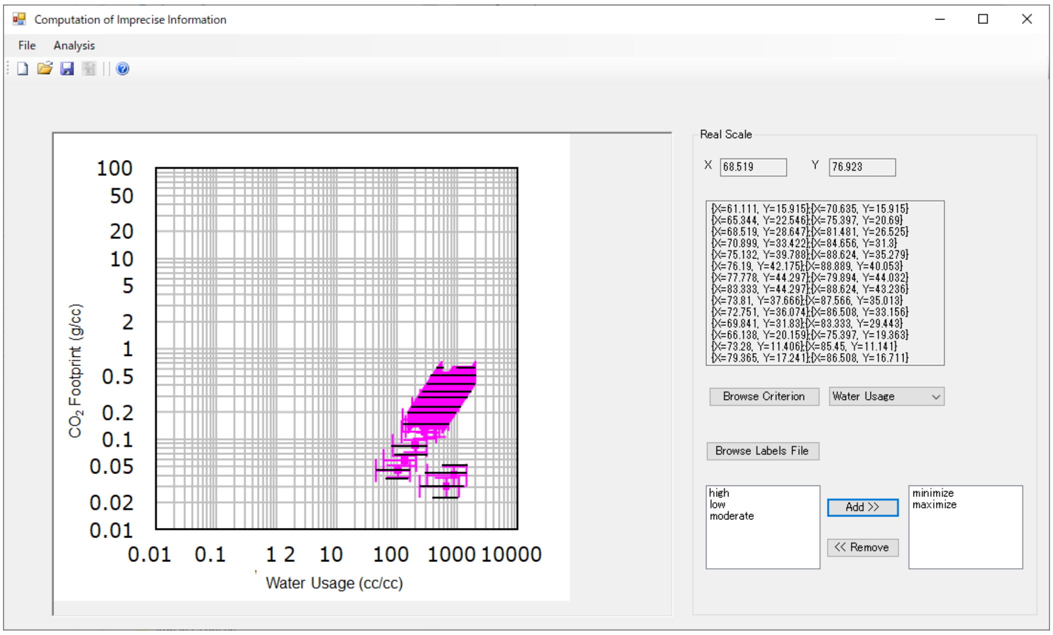


Figure 5-8. The user interface of the decision tool for extracting ranges representing the water usage of wooden materials

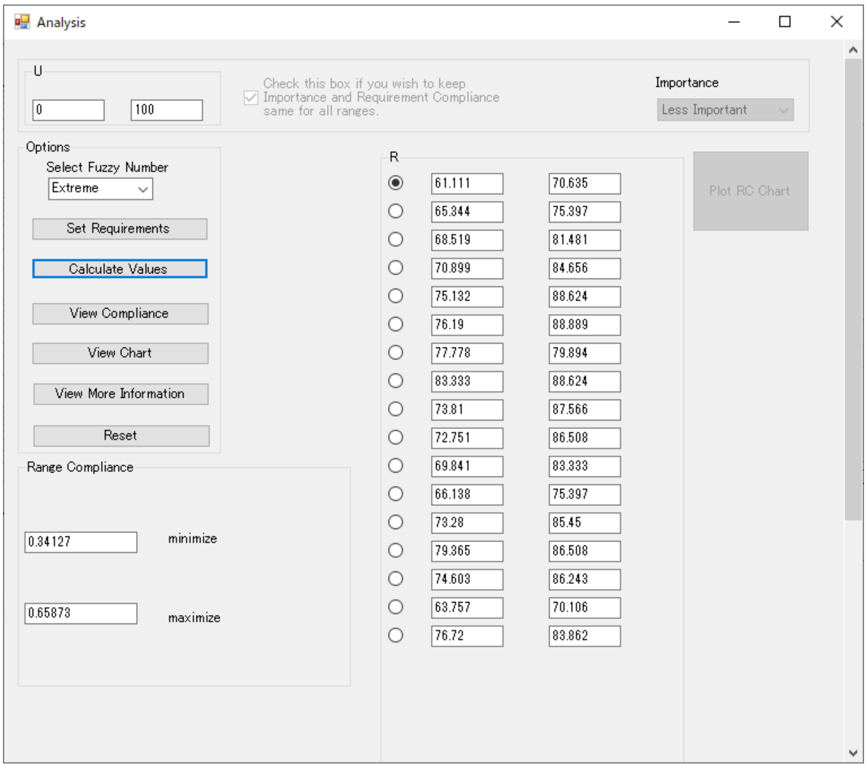


Figure 5-9. Extracted ranges representing water usage of wooden materials.

This means that both natural materials and polymers are equally unsustainable in terms of water usage. This is consistent with the visual inspection of the BD (Figure 5-7). On the other hand, the decision score (*D*) value concerning CO2 footprint is very high (0.904) for natural materials compared to that of polymers (0.449). The above-mentioned results imply that the presented decision-making approach is reliable and it can successfully process the uncertainty underlying BD.

Table 5-1. Sustainability assessment of wooden materials.

	Decision-making parameters					
Sustainability Criteria	t_1	t_2	t_3	r	s	D
Minimize water usage	0.140	0.210	0.320	0.288	0.151	0.395
Minimize CO_2 footprint	0.570	0.680	0.880	0.733	0.640	0.904

Table 5-2. Sustainability assessment of polymers.

	Decision-making parameters					
Sustainability Criteria	t_1	t_2	t_3	r	s	D
Minimize water usage	0.120	0.230	0.360	0.319	0.135	0.410
Minimize CO_2 footprint	0.140	0.260	0.390	0.345	0.159	0.449

5.4.2 Case study 2

Upon confirming the presented decision-making approach's reliability, it is time to apply it in multiple-criteria decision-making, as described below.

In particular, this case study shows how to rank the three most frequently used metallic materials: mild steel, aluminum alloys, and magnesium alloys, using graphical information generated from BD on engineering materials. Six criteria are considered: 1) strength, 2) Yang's modulus, 3) destiny, 4) CO2 footprint, 5) water usage, and 6) cost. As such, three two-dimensional plots—strength versus Yang's modulus, the cost versus density, and water usage versus CO2 footprint—are used to show the relative positions of the mild steel, aluminum alloys, and magnesium alloys on each plot. Figure 5-10 shows a scenario where all possible ranges of Yang's modulus of magnesium alloys are extracted. Figure 5-11 shows a scenario where all possible cost ranges of magnesium alloys are extracted. Figure 5-12 shows a scenario where all possible ranges of the CO2 footprint of magnesium alloys are extracted. The same matter is done for all combinations of criteria and alternatives. To ensure sustainability, the strength and Yang's modulus must be maximized, whereas the density, cost, CO2 footprint, and water usage must be minimized. The decision scores (Ds) of the alternatives respective to the respective criteria are listed in Table 5-3. Accordingly, the alternatives can be ranked using values of the scores listed in Table 5-3. The ranking process is described as follows.

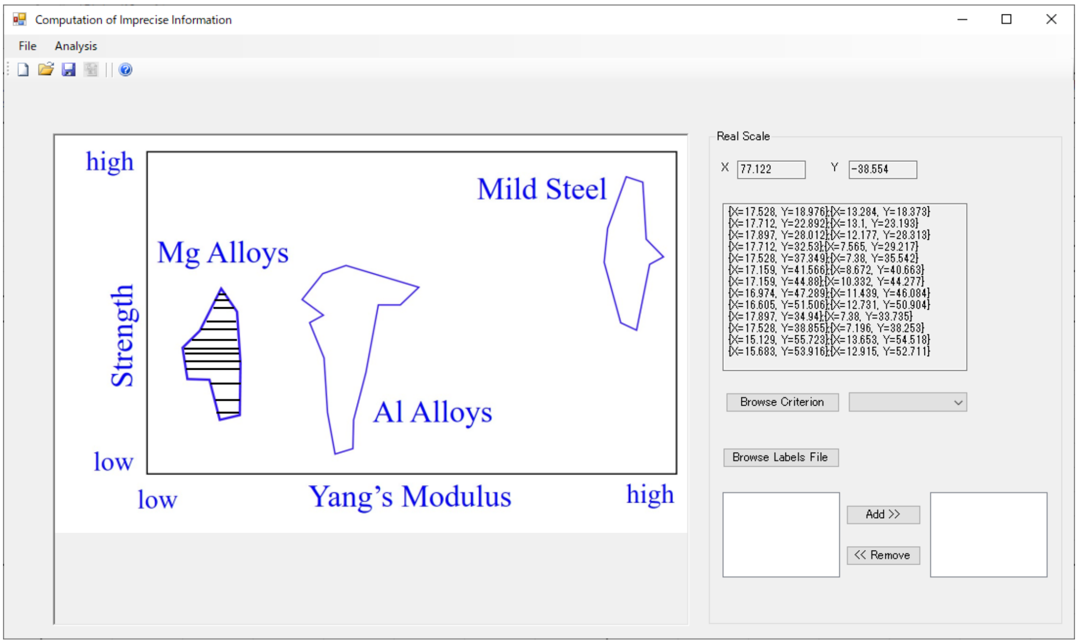


Figure 5-10. Extracting decision-relevant information from the plot of strength versus Yang's modulus.

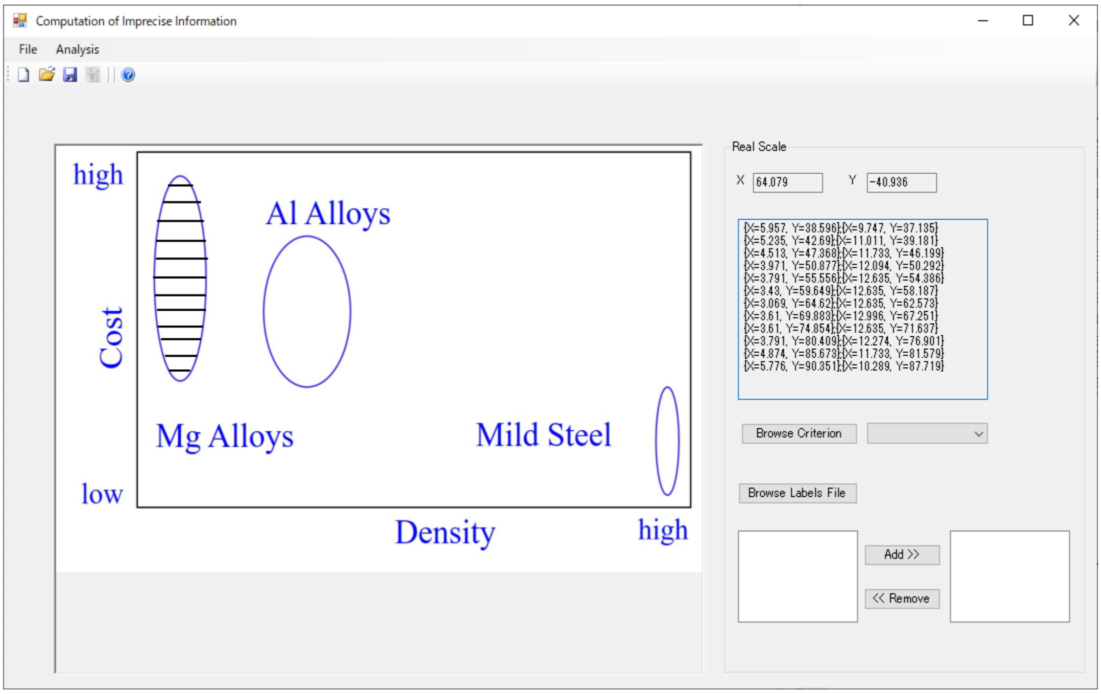


Figure 5-11. Extracting decision-relevant information from the plot of cost versus density.

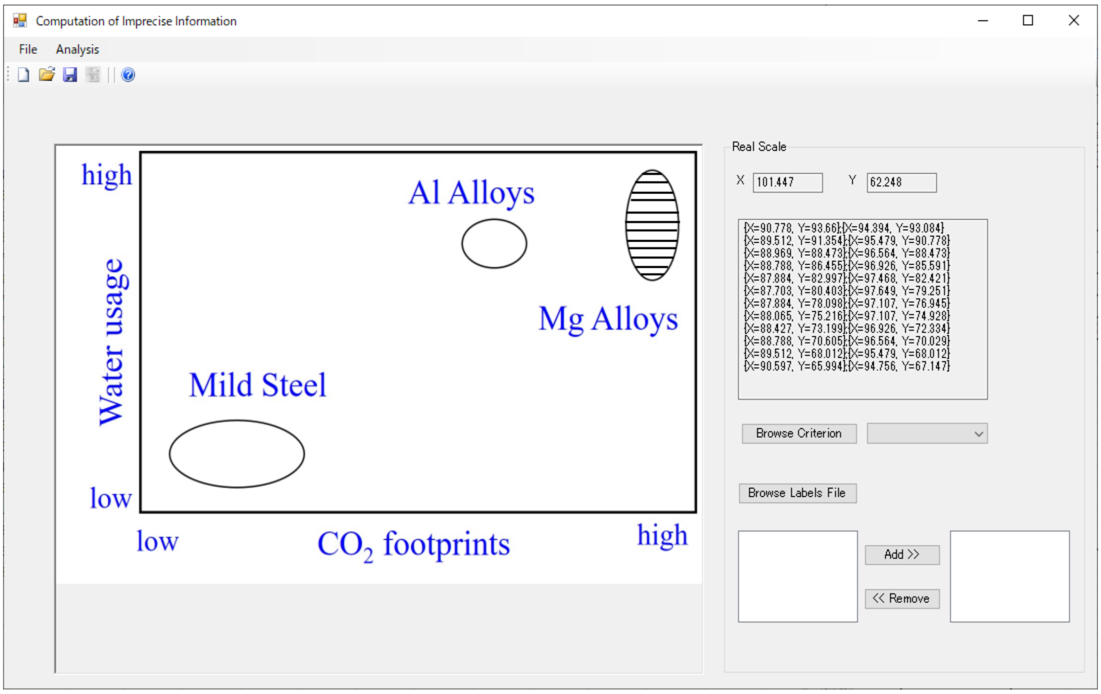


Figure 5-12. Extracting decision-relevant information from the plot of water usage versus CO2 footprint.

Table 5-3. Decision scores (Ds) of the alternatives with respect to the respective criteria.

Alternatives	Criteria					
	strength	Yang's modulus	cost	density	water usage	CO2 footprint
Mild steel	0.684	0.909	0.811	0.043	0.833	0.825
Aluminum Alloys	0.465	0.376	0.448	0.69	0.258	0.359
Magnesium Alloys	0.37	0.139	0.355	0.919	0.205	0.074

The ranking score denoted as R_j of the j-th alternative respective to all criteria is given as follows.

$$R_j = \frac{\sum_{i=1}^{n} D_{ij} w_i}{\sum_{i=1}^{n} w_i} \tag{5-8}$$

In equation (3-8), D_{ij} is the decision score of j-th alternative with respect to the i-th criterion, w_i is the weight or importance of the i-th criterion, and n is the number of criteria. In a multi-criteria decision-making process, the decision-makers can set the values of w_i as they prefer and see the corresponding ranking of the alternatives. For example, if the environmental issues are prioritized over others, the values of the weights of CO2 emission and water usage can be set much higher than that of the others. However, some of the possible weight-setting scenarios are presented below for better understanding.

As shown in Table 5-4, if the CO2 footprint, cost, and water usage are prioritized twice as much as others, then mild steel becomes the best alternative. Aluminum alloys are the second best, whereas magnesium alloys remain the last choice. A similar result is found in the case shown in Table 5-5, where the strength, Yang’s modulus, and cost are prioritized twice as much as others. In this case, mild steel becomes the best alternative. Aluminum alloys are the second best, whereas magnesium alloys remain the last choice.

Table 5-4. Ranking decision score when the CO2 footprint, cost, density, and water usage are prioritized twice as many others.

	Criteria						
	strength	Yang's modulus	cost	density	water usage	CO2 footprint	
Alternatives	weights						
	1	1	2	2	2	2	R_j
Mild steel	0.684	0.909	1.622	0.086	1.666	1.65	0.662
Aluminum Alloys	0.465	0.376	0.896	1.38	0.516	0.718	0.435
Magnesium Alloys	0.37	0.139	0.71	1.838	0.41	0.148	0.362

Table 5-5. Ranking decision score if the strength, Yang's modulus, and cost are prioritized twice as much as others.

	Criteria						
	strength	Yang's modulus	cost	density	water usage	CO2 footprint	
Alternatives	weights						
	2	2	2	1	1	1	R_j
Mild steel	1.368	1.818	1.622	0.043	0.833	0.825	0.723
Aluminum Alloys	0.93	0.752	0.896	0.69	0.258	0.359	0.432
Magnesium Alloys	0.74	0.278	0.71	0.919	0.205	0.074	0.325

However, if the density is about 3.5 times more important than others, all alternatives become indifferent (see Table 5-6). This means that magnesium alloys are the best alternatives when destiny receives very high importance compared to all other criteria.

The above examples show how an informed decision can easily be made using the presented decision-making approach, even though the relevant datasets presented using two-dimensional plots exhibit uncertainty. The remarkable thing is that the subjectivity or preference of the decision-makers can be documented by reporting the weights, as shown in Table 5-6.

Table 5-6. Ranking decision score if the density is given about 3.5 times more importance than others.

	Criteria						
	strength	Yang's modulus	cost	density	water usage	CO2 footprint	
Alternatives	weights						
	1	1	1	3.5	1	1	R_j
Mild steel	0.684	0.909	0.811	0.1505	0.833	0.825	0.496
Aluminum Alloys	0.465	0.376	0.448	2.415	0.258	0.359	0.508
Magnesium Alloys	0.37	0.139	0.355	3.2165	0.205	0.074	0.513

The presented decision-making approach can be studied further in different directions. One of the directions is discussed in this section as follows.

Calculating the decision score (*D*) is a critical issue in the decision-making approach. As defined before, the decision score is calculated using equations (4-5), (4-6), and (4-7). Nevertheless, the central theme of the decision score is to see whether or not an induced fuzzy number denoted as T matches the ideal one (denoted “Ideal” in Figure 5-4). Furthermore, the induced fuzzy number (T) captures the uncertainty in the compliances calculated using the graphical information. For getting more insights into this issue, consider three arbitrary alternatives, A, B, and C, and their induced fuzzy numbers denoted as T(A), T(B), and T(C), as schematically illustrated in Figure 5-13.

Alternative A is very close to the maximum compliance (unit). On the other hand, alternative C is very far from the maximum compliance (unit). The other alternative (B) stays between A and C. Compared to A and C, B entails a large amount of uncertainty, i.e., it is widely spread. Therefore, both distance of the induced fuzzy number from the maximum compliance and its degree of uncertainty must be quantified simultaneously to develop a more insightful decision score. Consequently, the following functions can be used to calculate the components of the decision score.

$$D1 = 1 - \left(\frac{t_1+t_2+t_3}{3}\right) \tag{5-9}$$

$$D2 = \frac{\frac{1}{2}(t_3-t_1)}{\frac{1}{2}} \tag{5-10}$$

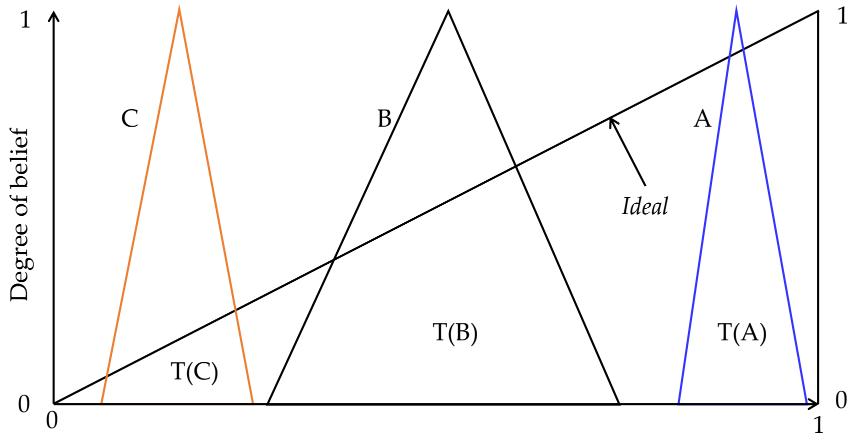


Figure 5-13. Comparing three alternatives using their induced fuzzy numbers

Here, $D1 \in [0,1]$ quantifies the distance of the centroid of an induced triangular fuzzy number from the maximum compliance (unit), whereas $D2 \in [0,1]$ quantifies the degree of uncertainty of an induced fuzzy number, which is equal to its area divided by the largest possible area. The less the value of $D1$, the better the alternative because the less the value of $D1$, the closer the alternative is to maximizing compliance. The less the value of $D2$, the better the alternative because the less the value of $D2$, the less the degree of uncertainty. As a result, the decision score (E) can be calculated as follows.

$$E = 1 - \frac{\sqrt{(D1)^2+(D2)^2}}{\sqrt{2}} \tag{5-11}$$

Thus, the more the value of E, the better the alternative. Based on this new decision-scoring approach, the arbitrary alternative shown in Figure 13 can be shown in the $D2$ versus $D1$ plot, as shown in Figure 5-14.

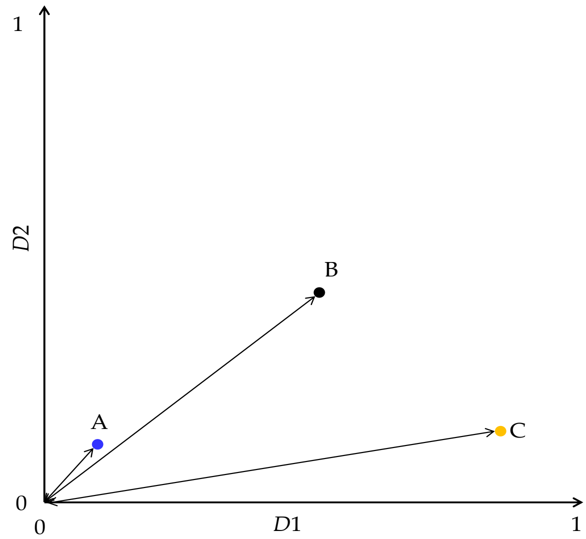


Figure 5-14. Relative positions of the arbitrary alternatives on the $D2$ versus $D1$ plot

The wooden and polymeric materials (case study 1) are reassessed using the new decision score. The results are shown in Table 5-7 and Table 5-8 for wooden and polymeric materials, respectively. As seen in Table 5-7 and Table 5-8, the decision score (E) concerning the criterion "minimize water usage" for both groups of materials is comparable (0.44 and 0.43). This means that both natural materials and polymers are equally unsustainable in terms of water usage. This is consistent with the visual inspection of the plot (Figure 5-7). On the other hand, the decision score (E) concerning CO2 footprint is very high (0.70) for natural materials compared to that of polymers (0.45). The above results are consistent with the results reported in case study 1. Thus, the decision scores given by D and E can serve the purpose of making the right decision. It is worth mentioning that E is perhaps a better decision score compared to D because it considers uncertainty on top of the compliances.

Table 5-7. Sustainability reassessment of wooden materials.

	Decision-making parameters					
Sustainability Criteria	t_1	t_2	t_3	*D1*	*D2*	*E*
Minimize water usage	0.14	0.21	0.32	0.78	0.18	0.44
Minimize CO_2 footprint	0.57	0.68	0.88	0.29	0.31	0.70

Table 5-8. Sustainability reassessment of wooden materials.

	Decision-making parameters					
Sustainability Criteria	t_1	t_2	t_3	*D1*	*D2*	*E*
Minimize water usage	0.12	0.23	0.36	0.76	0.24	0.43
Minimize CO_2 footprint	0.14	0.26	0.39	0.74	0.25	0.45

Chapter 6: Discussions and Future Research Directions

In this study, we investigated the smart manufacturing concept and proposed a comprehensive BDA framework that is free from BD inequality and the digital divide in a way that can be useful for Small and Mid-size Enterprises (SMEs). Furthermore, dataset preparation in the context of Human-cyber-physical systems has been investigated to be preprocessed and used in DTs. Furthermore, in the context of smart manufacturing, an innovative way has been described for using several two-dimensional plots, which are really common types of datasets in manufacturing processes instead of numerical data. Following our research, these are possible future opportunities:

1- The proposed framework for BDA can be extended and more developed used solve problems regarding different types of manufacturing processes in addition to machining operations.

2- The proposed BDA framework can be used for adding other data visualization and data analysis methods in addition to correlation analysis and uncertainty analysis using scatter plots and possibility distribution due to the modularity of our framework.

3- Regarding dataset preparation, since the surface roughness of the turning process has been investigated in this step, for the next phase of the study, it is confirmed that the proposed method can be used to prepare datasets of surface roughness of milling, turning, grinding, electrical discharge machining, and polishing. Surface profile height data of other processes will be considered to comprehensively prepare surface roughness BD.

4- Considering decision-making using graphical data, In the next phase of this study, a decision support system can be developed to automate the decision-making approach presented in the previous sections.

Chapter 7: Concluding Remarks

The concept of Big Data (BD) has been introduced under the umbrella of Industry 4.0 or smart manufacturing. BD of the manufacturing process in this study are categorized into three main parts: 1) Dataset for DT, 2- CV-EV-centric dataset, 3) Graphical dataset. This thesis considers the problem of developing BD and BDA for smart manufacturing, focusing on these categories and the following three related issues.

Issue 1: DTs of manufacturing phenomena are supposed to machine-learn the required knowledge using relevant datasets available in BD. Therefore, a research question is how to preprocess manufacturing phenomena-relevant datasets (dataset for DT) for using them directly in DTs.

Issue 2: Considering the CV-EV-centric dataset, BD and analytics require expensive resources and sophisticated computation arrangements. Thus, BD hardly benefits small and medium-sized manufacturing organizations, resulting in "BD inequality." Consequently, a research question is how to eliminate BD inequality.

Issue 3: BD is often visualized using several two-dimensional plots (graphical dataset). These plots are then used to make a decision informally. Consequently, a research question is how to make formal decisions by computing two-dimensional plots, not numerical data.

A specific chapter in this thesis has been presented to cover each issue. Chapter 2 generally describes the proposed BDA framework and its functional requirements. Chapter 3, BDA framework for dataset for DT, describes the dataset preparation and covers issue 1. Chapter 4, BDA framework for CV-EV-centric dataset, covers issue 2. Chapter 5, BDA framework for graphical data, describes decision-making using graphical data and covers issue 3. The following concluding remarks can be obtained according to each issue and its related chapter.

7.1 Big data analytics framework concluding remarks

BDA offers many data visualization facilities. It is also equipped with machine learning and computational intelligence-driven arrangements. This makes the decision-making process more formal. However, adding these computational arrangements makes the analytics computationally heavy and highly resource-depended. As a result, only large organizations can sustain BDA, and medium and small organizations fall behind. Thus, BDA results in an inequality referred to as BD inequality or digital divide. Measures are needed to mitigate this matter. Accordingly, we developed a BDA framework than can be free from BD inequality and the digital divide. For this purpose, considering a systematic viewpoint, we developed a BDA framework showing all the subsystems.

Furthermore, we have elucidated the functional requirements of the subsystems. This system has been developed using the JAVA™ Platform. We considered both human- and machine-friendly nature of the subsystems.

The proposed BDA framework consists of five interconnected systems: the BD preparation system, BD exploration system, data visualization system, data analysis system, and knowledge extraction system.

7.2 Preparing datasets for digital twin concluding remarks

The datasets included in human-cyber-physical system-friendly BD must have the following three characteristics: (1) The datasets must be readily accessible to all stakeholders through the Internet; (2) The datasets must be both human- and machine-readable; (3) The datasets can effortlessly be integrated with the machine-learning-based knowledge extraction segment of a DT. Unfortunately, BD exhibiting the abovementioned characteristics is not readily available. Thus, it (BD) needs to be prepared from the documentation of past research and operational activities. This is a challenge because of the lack of a steadfast procedure. This study fills this gap by presenting a state-of-the-art method for preparing the datasets of surface roughness to be included in industrial BD from the context of smart manufacturing and cognitive computing.

The surface roughness datasets included in human-cyber-physical system-friendly BD consist of four segments: semantic annotation, roughness model, simulation algorithm, and simulation system. These segments provide input information for the input module, modeling module, simulation module, and validation module, respectively, of a DT dedicated to administering surface roughness in a human-cyber-physical system.

The semantic annotation segment of the dataset boils down to a concept map. A human- and machine-readable concept map is developed for the dataset of surface roughness. The information of other segments (roughness model, simulation algorithm, and simulation system) can be integrated with the semantic annotation, which is done in this article.

The delay map of surface roughness profile heights plays a pivotal role in the dataset preparation.

Instead of the surface roughness modeling method used in this study, other modeling methods such as Markov chain, DNA-based computing, non-stationary Gaussian process, and semantic modeling can be used. The roughness model, simulation algorithm, and simulation system segments accord with the modeling method. Even though the dataset segments are reconstructed according to the modeling method, the data structure remains valid. Thus, the dataset structure (semantic annotation, roughness model, simulation algorithm, and simulation system) serves as the metadata of the surface roughness.

7.3 Machining decision making concluding remarks

According to the developed BDA framework, for the CV-EV-centric dataset, the BD preparation system creates digital manufacturing commons related to the CV-EV-centric dataset. It also helps form BD for manufacturing processes. BD exploration system helps search the digital manufacturing commons in the BD and acquire the relevant datasets for particular situations. The data visualization system represents relevant datasets using various visualization schemes like scatter plots and possibility distribution. The main goal of the data analysis system is to analyze the CV-EV-centric dataset and find the relationships among all CV-EV combinations for a given situation. Finally, the knowledge extraction system extracts rules, equations, and other forms of knowledge which can be related to manufacturing processes.

A case study is performed in this study To analyze and validate the functionality of the proposed BDA framework. Dry electrical discharge machining (dry EDM) is selected as the case study. And for knowledge extraction, the optimization (maximization) of one EV (*MRR*) is considered as the predetermined criteria. Considering google scholar as the source of documentation (scholarly articles), by defining process-based keyword search and using BD preparation and BD exploration system, we determined and made five HCPS-friendly and CV-EV-centric sources of the dataset (S1, S2, S3, S4, and S5). According to that, data visualization (using scatter plots) and a data analysis system were implemented. We performed the knowledge extraction system according to the R-values obtained from the data analysis system. According to the strength and direction of R values (positive or negative), we determine the significance of each CV (by defining criteria for |R| values), then their direction of effect. For example, for Current (*I*) and correlation analysis, almost R values for all sources are positive and significant ($R>0.9$). So, *I* is a significant CV for maximizing *MRR*. For maximization of *MRR*, we have to maximize *I*. According to this definition and considering other CVs, we obtained some rules for each source. For example, for S1: Maximize I, P, N, and Minimize V for the Maximization of *MRR*.

After obtaining these results for the validation purpose, we applied these rules to the dataset and found the related *MRR*. Then, this value was compared with the maximum *MRR* in the total dataset for each source. The interesting point is that for S1, S2, and S5, the validation tests were completely similar to the maximum *MRR* level, and for the other two sources, the applied rules data were unavailable in the dataset. The validation shows the efficacy of the proposed BDA framework for a specific purpose.

7.4 Dealing with graphical datasets concluding remarks

The information regarding indicators of SDGs and product life-cycle-based arrangements (e.g., material passport) create a vast information silo manifesting BD. Therefore, making the right decisions using such BD will play a pivotal role in achieving sustainability. Since BD is “big,” the relevant datasets in the visual form (a set of two-dimensional plots) become the decision-relevant information. This necessitates novel decision-making methods and tools capable of handling two-dimensional plots rather than numerical data. Furthermore, the methods and tools must accommodate the preferences of the decision-maker. Accordingly, this article presents a decision-making method and a tool to formally compute the two-dimensional plots of numerical data. The proposed method and tool can directly extract the decision-relevant information from two-dimensional plots (generated from BD) and compute the decision scores based on the maximization or minimization principle (e.g., minimizing CO2 emission).

The efficacy of the presented method and tool is shown using two case studies. The first case study shows that the proposed decision-making method and tool produce reliable results even though the relevant datasets presented using two-dimensional plots exhibit uncertainty. In this case study, the datasets regarding two eco-indicators (CO2 footprint and water usage) of two types of materials (wooden materials and polymers) are plotted using two scatter plots. The datasets regarding CO2 footprint and water usage of 447 wooden materials and 244 polymers are presented. As the plots show, both wooden materials and polymers can be ranked equally regarding water usage. However, on the other hand, as far as CO2 footprint is concerned, wooden materials are far better than polymers. The parameter of the proposed decision-making approach, denoted as the decision score, makes the same decision (wooden materials and polymers are indifferent in terms of water usage, but wooden materials are better than polymers in terms of CO2 emission).

The other case study applies the method and tool for multiple-criteria decision-making. In particular, this case ranks the three most frequently used metallic materials—mild steel, aluminum alloys, and magnesium alloys—using graphical information generated from BD on engineering materials. Six criteria are considered: 1) strength, 2) Yang's modulus, 3) destiny, 4) CO2 footprint, 5) water usage, and 6) cost. In this case study, a parameter denoted as the raking parameter is introduced that aggregates the decision scores of all criteria for an alternative by considering each criterion's importance or weight. It is shown that weight is instrumental in exercising the preferences of the decision-maker.

Numerous policymakers, practitioners, and researchers have been acting in a coordinated manner yet remaining independent to achieve SDGs. The SDG-centric activities, particularly 93 Tier 1, 72 Tier 2, and 62 Tier 3 indicators-centric activities all around the globe, manifest BD. Making sense of such a huge data silo is essential for ensuring the degree of fulfillment of SDGs or putting forward the right set of solutions if the fulfillment is not satisfactory. In such cases, decisions are most likely made from the visualized BD rather than numerical datasets. Therefore, the presented decision-making method and tool can contribute to the advancement of BD-related research for achieving SDGs.

.

List of References

[1] “Schuh, G.; Anderl, R.; Dumitrescu, R.; Krüger, A.; Hompel, M.t. (Eds.). Industrie 4.0 Maturity Index. Managing the Digital Transformation of Companies – UPDATE 2020. acatech STUDY, April 2020. Available online: https://en.acatech.de/publication/industrie-4-0-maturity-indexupdate-2020/.”

[2] A. Sharif Ullah, “Modeling and simulation of complex manufacturing phenomena using sensor signals from the perspective of Industry 4.0,” *Advanced Engineering Informatics*, vol. 39, pp. 1–13, Jan. 2019, doi: 10.1016/j.aei.2018.11.003.

[3] S. Fattahi, T. Okamoto, and S. Ura, “Preparing Datasets of Surface Roughness for Constructing Big Data from the Context of Smart Manufacturing and Cognitive Computing,” *Big Data and Cognitive Computing*, vol. 5, no. 4, p. 58, Oct. 2021, doi: 10.3390/bdcc5040058.

[4] “Big data analytics, URL https://www.ibm.com/analytics/hadoop/bigdata-analytics. Last accessed 2022.11.17.”

[5] “What is big data? URL https://www.oracle.com/big-data/what-is-bigdata/ Last accessed 2021.04.17.”

[6] F. X. Diebold, “On the Origin(s) and Development of the Term ‘Big Data,’” *Available at SSRN*, Sep. 2012, doi: 10.2139/ssrn.2152421.

[7] Nicolaus Henke, Michael Chui, James Manyika, Tamim Saleh, Bill Wiseman, and Guru Sethupathy, “The age of analytics: Competing in a data-driven world,” McKinsey Global Institute, Dec. 2016.

[8] S. Yin and O. Kaynak, “Big Data for Modern Industry: Challenges and Trends [Point of View],” *Proceedings of the IEEE*, vol. 103, no. 2, pp. 143–146, Feb. 2015, doi: 10.1109/JPROC.2015.2388958.

[9] L. Andre, “53 Important Statistics About How Much Data Is Created Every Day,” FinancesOnline. [Online]. Available: https://financesonline.com/how-much-data-is-created-every-day/

[10] R. R. Sreenivasan, “Characteristics of Big Data – A Delphi study,” Master’s Thesis, Memorial University of Newfoundland, Newfoundland, Canada, 2017. Accessed: Sep. 12, 2022. [Online]. Available: https://research.library.mun.ca/13080/1/thesis.pdf

[11] T. Verevka, A. Mirolyubov, and J. Makio, “Opportunities and Barriers to Using Big Data Technologies in the Metallurgical Industry,” in *Innovations in Digital Economy*, Cham, 2021, pp. 86–102.

[12] S. Fattahi and A. S. Ullah, “Optimization of Dry Electrical Discharge Machining of Stainless Steel using Big Data Analytics,” in *Procedia CIRP*, Gulf of Naples, Italy, 2021.

[13] S. Qureshi, “Overcoming Technological Determinism in Understanding the Digital Divide: Where Do We Go From Here?,” *Information Technology for Development*, vol. 20, no. 3, pp. 215–217, Jul. 2014, doi: 10.1080/02681102.2014.930981.

[14] J. Cinnamon, “Data inequalities and why they matter for development,” *Information Technology for Development*, vol. 26, no. 2, pp. 214–233, Apr. 2020, doi: 10.1080/02681102.2019.1650244.

[15] M. Hilbert, “Big Data for Development: A Review of Promises and Challenges,” *Development Policy Review*, vol. 34, no. 1, pp. 135–174, Jan. 2016, doi: 10.1111/dpr.12142.

[16] M. Favaretto, E. De Clercq, and B. S. Elger, “Big Data and discrimination: perils, promises and solutions. A systematic review,” *Journal of Big Data*, vol. 6, no. 1, p. 12, Feb. 2019, doi: 10.1186/s40537-019-0177-4.

[17] E. Sisinni, A. Saifullah, S. Han, U. Jennehag, and M. Gidlund, “Industrial Internet of Things: Challenges, Opportunities, and Directions,” *IEEE Transactions on Industrial Informatics*, vol. 14, no. 11, pp. 4724–4734, Nov. 2018, doi: 10.1109/TII.2018.2852491.

[18] B. Beckmann, A. Giani, J. Carbone, P. Koudal, J. Salvo, and J. Barkley, “Developing the Digital Manufacturing Commons: A National Initiative for US Manufacturing Innovation,” *44th North American Manufacturing Research Conference, NAMRC 44, June 27-July 1, 2016, Blacksburg, Virginia, United States*, vol. 5, pp. 182–194, Jan. 2016, doi: 10.1016/j.promfg.2016.08.017.

[19] Md. A. Moktadir, S. M. Ali, S. K. Paul, and N. Shukla, “Barriers to big data analytics in manufacturing supply chains: A case study from Bangladesh,” *Computers & Industrial Engineering*, vol. 128, pp. 1063–1075, Feb. 2019, doi: 10.1016/j.cie.2018.04.013.

[20] “Motohashi, K., Survey of Big Data Use and Innovation in Japanese Manufacturing Firms, RIETI Policy Discussion Paper, 17-P-027, August

2017. Available from https://www.rieti.go.jp/jp/publications/pdp/17p027.pdf."

[21] "Kim, Y.G. and Motohashi, K. Use and sharing of big data, firm networks and their performance, RIETI Policy Discussion Pa-per, 19-E-016, March 2019. Available from https://www.rieti.go.jp/jp/publications/dp/19e016.pdf."

[22] M. Syafrudin, G. Alfian, N. Fitriyani, and J. Rhee, "Performance Analysis of IoT-Based Sensor, Big Data Processing, and Machine Learning Model for Real-Time Monitoring System in Automotive Manufacturing," *Sensors*, vol. 18, no. 9, p. 2946, Sep. 2018, doi: 10.3390/s18092946.

[23] M. Wiech, A. Boffelli, C. Elbe, P. Carminati, T. Friedli, and M. Kalchschmidt, "Implementation of big data analytics and Manufacturing Execution Systems: an empirical analysis in German-speaking countries," *Production Planning & Control*, vol. 33, no. 2–3, pp. 261–276, Feb. 2022, doi: 10.1080/09537287.2020.1810766.

[24] C. A. Escobar, M. E. McGovern, and R. Morales-Menendez, "Quality 4.0: a review of big data challenges in manufacturing," *Journal of Intelligent Manufacturing*, vol. 32, no. 8, pp. 2319–2334, Dec. 2021, doi: 10.1007/s10845-021-01765-4.

[25] W. Yu, T. Dillon, F. Mostafa, W. Rahayu, and Y. Liu, "A Global Manufacturing Big Data Ecosystem for Fault Detection in Predictive Maintenance," *IEEE Transactions on Industrial Informatics*, vol. 16, no. 1, pp. 183–192, Jan. 2020, doi: 10.1109/TII.2019.2915846.

[26] M. Faheem, R. A. Butt, R. Ali, B. Raza, Md. A. Ngadi, and V. C. Gungor, "CBI4.0: A cross-layer approach for big data gathering for active monitoring and maintenance in the manufacturing industry 4.0," *Journal of Industrial Information Integration*, vol. 24, p. 100236, Dec. 2021, doi: 10.1016/j.jii.2021.100236.

[27] P. O'Donovan, K. Leahy, K. Bruton, and D. T. J. O'Sullivan, "An industrial big data pipeline for data-driven analytics maintenance applications in large-scale smart manufacturing facilities," *Journal of Big Data*, vol. 2, no. 1, p. 25, Nov. 2015, doi: 10.1186/s40537-015-0034-z.

[28] D. Shah, J. Wang, and Q. P. He, "Feature engineering in big data analytics for IoT-enabled smart manufacturing – Comparison between deep learning and statistical learning," *Computers & Chemical Engineering*, vol. 141, p. 106970, Oct. 2020, doi: 10.1016/j.compchemeng.2020.106970.

[29] W. Fang, Y. Guo, W. Liao, K. Ramani, and S. Huang, "Big data driven jobs remaining time prediction in discrete manufacturing system: a deep learning-based approach," *International Journal of Production Research*, vol. 58, no. 9, pp. 2751–2766, May 2020, doi: 10.1080/00207543.2019.1602744.

[30] C. Zhang, Z. Wang, K. Ding, F. T. S. Chan, and W. Ji, "An energy-aware cyber physical system for energy Big data analysis and recessive production anomalies detection in discrete manufacturing workshops," *International Journal of Production Research*, vol. 58, no. 23, pp. 7059–7077, Dec. 2020, doi: 10.1080/00207543.2020.1748904.

[31] Y.-C. Ko and H. Fujita, "An evidential analytics for buried information in big data samples: Case study of semiconductor manufacturing," *Information Sciences*, vol. 486, pp. 190–203, Jun. 2019, doi: 10.1016/j.ins.2019.01.079.

[32] S. Mihály *et al.*, "Earth observation and geospatial big data management and engagement of stakeholders in Hungary to support the SDGs," *Big Earth Data*, vol. 5, no. 3, pp. 306–351, Jul. 2021, doi: 10.1080/20964471.2021.1940733.

[33] J. Li, W.-H. Chen, Q. Xu, N. Shah, and T. Mackey, "Leveraging Big Data to Identify Corruption as an SDG Goal 16 Humanitarian Technology," in *2019 IEEE Global Humanitarian Technology Conference (GHTC)*, Oct. 2019, pp. 1–4. doi: 10.1109/GHTC46095.2019.9033129.

[34] "UNSDG. 2019. Available online: https://unstats.un.org/sdgs/indicators/database/ (accessed on 5 December 2021)."

[35] R. Mark and G. Anya, "Ethics of Using Smart City AI and Big Data: The Case of Four Large European Cities," *The ORBIT Journal*, vol. 2, no. 2, pp. 1–36, Jan. 2019, doi: 10.29297/orbit.v2i2.110.

[36] S. MacFeely, "The Big (data) Bang: Opportunities and Challenges for Compiling SDG Indicators," *Global Policy*, vol. 10, no. S1, pp. 121–133, 2019, doi: 10.1111/1758-5899.12595.

[37] "Background Document for the Note by the Secretary-General Transmitting the Report of the Global Working Group on Big Data for Official Statistics (E/CN.3/2020/24). Available online: https://unstats.un.org/unsd/statcom/51st-

session/documents/UN_BigData_report_v6.0-E.html (accessed on 27 September 2021).”

[38] K. S. Mwitondi, I. Munyakazi, and B. N. Gatsheni, “A robust machine learning approach to SDG data segmentation,” *Journal of Big Data*, vol. 7, no. 1, p. 97, Nov. 2020, doi: 10.1186/s40537-020-00373-y.

[39] A. Kharrazi, “Challenges and opportunities of urban big-data for sustainable development,” *Asia-Pacific Tech Monitor*, vol. 34, no. 4, pp. 17–21, 2017.

[40] C. S. Kruse, R. Goswamy, Y. J. Raval, and S. Marawi, “Challenges and Opportunities of Big Data in Health Care: A Systematic Review,” *JMIR Medical Informatics*, vol. 4, no. 4, p. e5359, Nov. 2016, doi: 10.2196/medinform.5359.

[41] Y. Ma *et al.*, “Remote sensing big data computing: Challenges and opportunities,” *Future Generation Computer Systems*, vol. 51, pp. 47–60, 2015.

[42] B. Ferreira, M. Iten, and R. G. Silva, “Monitoring sustainable development by means of earth observation data and machine learning: a review,” *Environmental Sciences Europe*, vol. 32, no. 1, p. 120, Sep. 2020, doi: 10.1186/s12302-020-00397-4.

[43] R. Kashyap and F. C. J. Verkroost, “Analysing global professional gender gaps using LinkedIn advertising data,” *EPJ Data Science*, vol. 10, no. 1, p. 39, Jul. 2021, doi: 10.1140/epjds/s13688-021-00294-7.

[44] H. Hassani, X. Huang, S. MacFeely, and M. R. Entezarian, “Big Data and the United Nations Sustainable Development Goals (UN SDGs) at a Glance,” *Big Data and Cognitive Computing*, vol. 5, no. 3, 2021, doi: 10.3390/bdcc5030028.

[45] C. Allen, M. Smith, M. Rabiee, and H. Dahmm, “A review of scientific advancements in datasets derived from big data for monitoring the Sustainable Development Goals,” *Sustainability Science*, vol. 16, no. 5, pp. 1701–1716, Sep. 2021, doi: 10.1007/s11625-021-00982-3.

[46] J. Wang, C. Xu, J. Zhang, and R. Zhong, “Big data analytics for intelligent manufacturing systems: A review,” *Journal of Manufacturing Systems*, vol. 62, pp. 738–752, Jan. 2022, doi: 10.1016/j.jmsy.2021.03.005.

[47] S. Kahveci, B. Alkan, M. H. Ahmad, B. Ahmad, and R. Harrison, “An end-to-end big data analytics platform for IoT-enabled smart factories: A case study of battery module assembly system for electric vehicles,” *Journal of*

Manufacturing Systems, vol. 63, pp. 214–223, Apr. 2022, doi: 10.1016/j.jmsy.2022.03.010.

[48] J. Woo, S.-J. Shin, W. Seo, and P. Meilanitasari, "Developing a big data analytics platform for manufacturing systems: architecture, method, and implementation," *Int J Adv Manuf Technol*, vol. 99, no. 9–12, pp. 2193–2217, Dec. 2018, doi: 10.1007/s00170-018-2416-9.

[49] C. Jun, J. Y. Lee, and B. H. Kim, "Cloud-based big data analytics platform using algorithm templates for the manufacturing industry," *International Journal of Computer Integrated Manufacturing*, vol. 32, no. 8, pp. 723–738, Aug. 2019, doi: 10.1080/0951192X.2019.1610578.

[50] Y. Lu and X. Xu, "Cloud-based manufacturing equipment and big data analytics to enable on-demand manufacturing services," *Robotics and Computer-Integrated Manufacturing*, vol. 57, pp. 92–102, Jun. 2019, doi: 10.1016/j.rcim.2018.11.006.

[51] W. Ji and L. Wang, "Big data analytics based fault prediction for shop floor scheduling," *Journal of Manufacturing Systems*, vol. 43, pp. 187–194, Apr. 2017, doi: 10.1016/j.jmsy.2017.03.008.

[52] W. Ji, S. Yin, and L. Wang, "A big data analytics based machining optimisation approach," *J Intell Manuf*, vol. 30, no. 3, pp. 1483–1495, Mar. 2019, doi: 10.1007/s10845-018-1440-9.

[53] C.-C. Chen *et al.*, "A Novel Efficient Big Data Processing Scheme for Feature Extraction in Electrical Discharge Machining," *IEEE Robot. Autom. Lett.*, vol. 4, no. 2, pp. 910–917, Apr. 2019, doi: 10.1109/LRA.2019.2891498.

[54] A. Sharif Ullah, "Modeling and simulation of complex manufacturing phenomena using sensor signals from the perspective of Industry 4.0," *Advanced Engineering Informatics*, vol. 39, pp. 1–13, Jan. 2019, doi: 10.1016/j.aei.2018.11.003.

[55] A. S. Ullah, "What is knowledge in Industry 4.0?," *Engineering Reports*, vol. 2, no. 8, p. e12217, 2020, doi: 10.1002/eng2.12217.

[56] T. Wuest, D. Weimer, C. Irgens, and K.-D. Thoben, "Machine learning in manufacturing: advantages, challenges, and applications," *Production & Manufacturing Research*, vol. 4, no. 1, pp. 23–45, Jan. 2016, doi: 10.1080/21693277.2016.1192517.

[57] Y. Zhang, C. Qian, J. Lv, and Y. Liu, "Agent and Cyber-Physical System Based Self-Organizing and Self-Adaptive Intelligent Shopfloor," *IEEE Transactions on Industrial Informatics*, vol. 13, no. 2, pp. 737–747, Apr. 2017, doi: 10.1109/TII.2016.2618892.

[58] Son Bui and Theodore Vorburger, "Surface Metrology Algorithm Testing System," no. 31, Jan. 2007.

[59] Md. A. Moktadir, S. M. Ali, S. K. Paul, and N. Shukla, "Barriers to big data analytics in manufacturing supply chains: A case study from Bangladesh," *Computers & Industrial Engineering*, vol. 128, pp. 1063–1075, Feb. 2019, doi: 10.1016/j.cie.2018.04.013.

[60] "ISO TC 213, Available online: http://www.iso.org/iso/home/store/catalogue_tc/catalogue_tc_browse.htm?commid=54924&published=on (accessed on 30 May 2021)."

[61] S. H. Bui and T. V. Vorburger, "Surface metrology algorithm testing system," *Precision Engineering*, vol. 31, no. 3, pp. 218–225, Jul. 2007, doi: 10.1016/j.precisioneng.2007.01.002.

[62] A. M. M. Sharif Ullah, A. Fuji, A. Kubo, J. Tamaki, and M. Kimura, "On the Surface Metrology of Bimetallic Components," *Machining Science and Technology*, vol. 19, no. 2, pp. 339–359, Apr. 2015, doi: 10.1080/10910344.2015.1018536.

[63] A. Kubo, R. Teti, A. S. Ullah, K. Iwadate, and T. Segreto, "Determining Surface Topography of a Dressed Grinding Wheel Using Bio-Inspired DNA-Based Computing," *Materials*, vol. 14, no. 8, 2021, doi: 10.3390/ma14081899.

[64] T. Li, L. A. Blunt, X. Jiang, and W. Zeng, "An Information Model for Surface Metrology," *Procedia CIRP*, vol. 10, pp. 251–258, Jan. 2013, doi: 10.1016/j.procir.2013.08.038.

[65] M. Farboodi, R. Mihet, T. Philippon, and L. Veldkamp, "Big Data and Firm Dynamics," *AEA Papers and Proceedings*, vol. 109, pp. 38–42, 2019, doi: 10.1257/pandp.20191001.

[66] "What Is Big Data? Available online: https://www.oracle.com/big-data/what-is-big-data/ (accessed on 20 October 2021)."

[67] "Big Data Analytics. Available online: https://www.ibm.com/analytics/hadoop/big-data-analytics (accessed on 20 October 2021)."

[68] W. Ji, S. Yin, and L. Wang, “A big data analytics based machining optimisation approach,” *Journal of Intelligent Manufacturing*, vol. 30, no. 3, pp. 1483–1495, Mar. 2019, doi: 10.1007/s10845-018-1440-9.

[69] S. Ren, X. Zhao, B. Huang, Z. Wang, and X. Song, “A framework for shopfloor material delivery based on real-time manufacturing big data,” *Journal of Ambient Intelligence and Humanized Computing*, vol. 10, no. 3, pp. 1093–1108, Mar. 2019, doi: 10.1007/s12652-018-1017-7.

[70] O. Oleghe, “A predictive noise correction methodology for manufacturing process datasets,” *Journal of Big Data*, vol. 7, no. 1, p. 89, Oct. 2020, doi: 10.1186/s40537-020-00367-w.

[71] P. M. LaCasse, W. Otieno, and F. P. Maturana, “A hierarchical, fuzzy inference approach to data filtration and feature prioritization in the connected manufacturing enterprise,” *Journal of Big Data*, vol. 5, no. 1, p. 45, Nov. 2018, doi: 10.1186/s40537-018-0155-2.

[72] C. Bizer, T. Heath, and T. Berners-Lee, “Linked Data - The Story So Far,” *International Journal on Semantic Web and Information Systems (IJSWIS)*, vol. 5, no. 3, pp. 1–22, 2009, doi: 10.4018/jswis.2009081901.

[73] S. R. Bader, I. Grangel-Gonzalez, P. Nanjappa, M.-E. Vidal, and M. Maleshkova, “A Knowledge Graph for Industry 4.0,” in *The Semantic Web*, Cham, 2020, pp. 465–480. doi: 10.1007/978-3-030-49461-2_27.

[74] X. Li, P. Zhuang, and C. Yin, “A metadata based manufacturing resource ontology modeling in cloud manufacturing systems,” *J Ambient Intell Human Comput*, vol. 10, no. 3, pp. 1039–1047, Mar. 2019, doi: 10.1007/s12652-018-0964-3.

[75] M. Antunes, D. Gomes, and R. L. Aguiar, “Towards IoT data classification through semantic features,” *Future Generation Computer Systems*, vol. 86, pp. 792–798, Sep. 2018, doi: 10.1016/j.future.2017.11.045.

[76] L. Ramos, “Semantic Web for manufacturing, trends and open issues: Toward a state of the art,” *Computers & Industrial Engineering*, vol. 90, pp. 444–460, Dec. 2015, doi: 10.1016/j.cie.2015.10.013.

[77] S. R. Bader and M. Maleshkova, “SOLIOT—Decentralized Data Control and Interactions for IoT,” *Future Internet*, vol. 12, no. 6, Art. no. 6, Jun. 2020, doi: 10.3390/fi12060105.

[78] S. R. Bader, M. Maleshkova, and S. Lohmann, “Structuring Reference Architectures for the Industrial Internet of Things,” *Future Internet*, vol. 11, no. 7, Art. no. 7, Jul. 2019, doi: 10.3390/fi11070151.

[79] S. Liang, K. Stockinger, T. M. de Farias, M. Anisimova, and M. Gil, “Querying knowledge graphs in natural language,” *Journal of Big Data*, vol. 8, no. 1, p. 3, Jan. 2021, doi: 10.1186/s40537-020-00383-w.

[80] A. M. M. S. Ullah, N. Arai, and M. Watanabe, “Concept Map and Internet-aided Manufacturing,” *Procedia CIRP*, vol. 12, pp. 378–383, Jan. 2013, doi: 10.1016/j.procir.2013.09.065.

[81] A. K. Ghosh, A. S. Ullah, A. Kubo, T. Akamatsu, and D. M. D’Addona, “Machining Phenomenon Twin Construction for Industry 4.0: A Case of Surface Roughness,” *Journal of Manufacturing and Materials Processing*, vol. 4, no. 1, Art. no. 1, Mar. 2020, doi: 10.3390/jmmp4010011.

[82] A. K. Ghosh, A. S. Ullah, R. Teti, and A. Kubo, “Developing sensor signal-based digital twins for intelligent machine tools,” *Journal of Industrial Information Integration*, vol. 24, p. 100242, Dec. 2021, doi: 10.1016/j.jii.2021.100242.

[83] A. M. M. Sharif Ullah and M. Shamsuzzaman, “Fuzzy Monte Carlo Simulation using point-cloud-based probability–possibility transformation,” *SIMULATION*, vol. 89, no. 7, pp. 860–875, Jul. 2013, doi: 10.1177/0037549713482174.

[84] T. Berners-Lee, W. Hall, J. A. Hendler, K. O’Hara, N. Shadbolt, and D. J. Weitzner, “A Framework for Web Science,” *WEB*, vol. 1, no. 1, pp. 1–130, Sep. 2006, doi: 10.1561/1800000001.

[85] A. M. M. S. Ullah, D. D’Addona, and N. Arai, “DNA based computing for understanding complex shapes,” *Biosystems*, vol. 117, pp. 40–53, Mar. 2014, doi: 10.1016/j.biosystems.2014.01.003.

[86] B. Denkena, M.-A. Dittrich, S. Stamm, M. Wichmann, and S. Wilmsmeier, “Gentelligent processes in biologically inspired manufacturing,” *CIRP Journal of Manufacturing Science and Technology*, vol. 32, pp. 1–15, Jan. 2021, doi: 10.1016/j.cirpj.2020.09.015.

[87] D. M. D’Addona, A. M. M. S. Ullah, and D. Matarazzo, “Tool-wear prediction and pattern-recognition using artificial neural network and DNA-based computing,” *J Intell Manuf*, vol. 28, no. 6, pp. 1285–1301, Aug. 2017, doi: 10.1007/s10845-015-1155-0.

[88] B. Denkena, M.-A. Dittrich, S. Stamm, M. Wichmann, and S. Wilmsmeier, "Gentelligent processes in biologically inspired manufacturing," *CIRP Journal of Manufacturing Science and Technology*, vol. 32, pp. 1–15, Jan. 2021, doi: 10.1016/j.cirpj.2020.09.015.

[89] G. Puthumana and S. S. Joshi, "Investigations into performance of dry EDM using slotted electrodes," *Int. J. Precis. Eng. Manuf.*, vol. 12, no. 6, pp. 957–963, Dec. 2011, doi: 10.1007/s12541-011-0128-2.

[90] P. Govindan and S. S. Joshi, "Experimental characterization of material removal in dry electrical discharge drilling," *International Journal of Machine Tools and Manufacture*, vol. 50, no. 5, pp. 431–443, May 2010, doi: 10.1016/j.ijmachtools.2010.02.004.

[91] R. Murickan, "Experimental investigation of Dry Electrical Discharge Machining on SS 316L," *International Journal of Latest Trends in Engineering and Technology*, vol. Vol. 2, p. 100, May 2013.

[92] S. Fattahi and H. Baseri, "Analysis of dry electrical discharge machining in different dielectric mediums," *Proceedings of the Institution of Mechanical Engineers, Part E: Journal of Process Mechanical Engineering*, vol. 231, no. 3, pp. 497–512, Jun. 2017, doi: 10.1177/0954408915611540.

[93] S. Fattahi, I. Shyha, and H. Baseri, "Optimisation of Dry Electrical Discharge Machining of High Speed Steel using Grey-Relational Analysis," *International Journal of Robotics and Mechatronics*, vol. 2, no. 4. 2016. [Online]. Available: http://ojs.unsysdigital.com/index.php/ijrm/article/view/886

[94] "Report of the World Commission on Environment and Development: Our Common Future." Annex to General Assembly Document A/42/427.

[95] "#Envision2030: 17 Goals to Transform the World for Persons with Disabilities 2019. Available online: https://www.un.org/ development/desa/disabilitiesenvision2030.html (accessed on 5 December 2021)." https://www.un.org/development/desa/disabilities/envision2030.html (accessed Nov. 21, 2022).

[96] "Big Data, Preliminary Report, ISO/IEC JTC1: Information Technology. pp. 21–23. Available online: http://www.iso.org/iso/big_data_report-jtc1.pdf. (accessed on 29 March 2022)."

[97] S. Fattahi, T. Okamoto, and S. Ura, "Preparing Datasets of Surface Roughness for Constructing Big Data from the Context of Smart

Manufacturing and Cognitive Computing," *Big Data and Cognitive Computing*, vol. 5, no. 4, Art. no. 4, Dec. 2021, doi: 10.3390/bdcc5040058.

[98] NIST Big Data Public Working Group, Definitions and Taxonomies Subgroup., "NIST Big Data Interoperability Framework:: volume 1, definitions version 3," National Institute of Standards and Technology, Gaithersburg, MD, NIST SP 1500-1r2, Oct. 2019. doi: 10.6028/NIST.SP.1500-1r2.

[99] NIST Big Data Public Working Group, "NIST Big Data Interoperability Framework :: volume 2, big data taxonomies version 3," National Institute of Standards and Technology, Gaithersburg, MD, NIST SP 1500-2r2, Nov. 2019. doi: 10.6028/NIST.SP.1500-2r2.

[100] Wo Chang, Arnab Roy, and Mark Underwood, "NIST Big Data Interoperability Framework: Volume 4, Security and Privacy," Oct. 2019, doi: 10.6028/NIST.SP.1500-4r2.

[101] Wo Chang, Sanjay Mishra, and NBD-PWG NIST, "NIST Big Data Interoperability Framework: Volume 5, Architectures White Paper Survey," Oct. 2015, doi: 10.6028/NIST.SP.1500-5.

[102] Wo Chang, David Boyd, and Orit Levin, "NIST Big Data Interoperability Framework: Volume 6, Reference Architecture," Oct. 2019, doi: 10.6028/NIST.SP.1500-6r2.

[103] NIST Big Data Public Working Group, "NIST Big Data Interoperability Framework: volume 7, standards roadmap, version 2," National Institute of Standards and Technology, Gaithersburg, MD, NIST SP 1500-7r1, Jun. 2018. doi: 10.6028/NIST.SP.1500-7r1.

[104] NIST Big Data Public Working Group, Definitions and Taxonomies Subgroup., "NIST Big Data Interoperability Framework:: volume 9, adoption and modernization version 3," National Institute of Standards and Technology, Gaithersburg, MD, NIST SP 1500-10r1, Oct. 2019. doi: 10.6028/NIST.SP.1500-10r1.

[105] "Big Data and discrimination: perils, promises and solutions. A systematic review | Journal of Big Data | Full Text." https://journalofbigdata.springeropen.com/articles/10.1186/s40537-019-0177-4 (accessed Nov. 22, 2022).

[106] A. M. M. S. Ullah and M. Noor-E-Alam, "Big data driven graphical information based fuzzy multi criteria decision making," *Applied Soft Computing*, vol. 63, pp. 23–38, Feb. 2018, doi: 10.1016/j.asoc.2017.11.026.

[107] S. Shahinur, A. M. M. S. Ullah, M. Noor-E-Alam, H. Haniu, and A. Kubo, "A decision model for making decisions under epistemic uncertainty and its application to select materials," *Artificial Intelligence for Engineering Design, Analysis and Manufacturing*, vol. 31, no. 3, pp. 298–312, 2017, doi: 10.1017/S0890060417000191.

[108] A. M. M. Sharif Ullah and M. Shamsuzzaman, "Fuzzy Monte Carlo Simulation using point-cloud-based probability–possibility transformation," *SIMULATION*, vol. 89, no. 7, pp. 860–875, Jul. 2013, doi: 10.1177/0037549713482174.

[109] M. a. K. Chowdhury, A. M. M. Sharif Ullah, and S. Anwar, "Drilling High Precision Holes in Ti6Al4V Using Rotary Ultrasonic Machining and Uncertainties Underlying Cutting Force, Tool Wear, and Production Inaccuracies," *Materials*, vol. 10, no. 9, Art. no. 9, Sep. 2017, doi: 10.3390/ma10091069.

[110] M. Ryan, J. Antoniou, L. Brooks, T. Jiya, K. Macnish, and B. Stahl, "Technofixing the Future: Ethical Side Effects of Using AI and Big Data to Meet the SDGs," in *2019 IEEE SmartWorld, Ubiquitous Intelligence & Computing, Advanced & Trusted Computing, Scalable Computing & Communications, Cloud & Big Data Computing, Internet of People and Smart City Innovation (SmartWorld/SCALCOM/UIC/ATC/CBDCom/IOP/SCI)*, Aug. 2019, pp. 335–341. doi: 10.1109/SmartWorld-UIC-ATC-SCALCOM-IOP-SCI.2019.00101.

[111] J. M. Allwood, J. M. Cullen, and R. L. Milford, "Options for Achieving a 50% Cut in Industrial Carbon Emissions by 2050," *Environ. Sci. Technol.*, vol. 44, no. 6, pp. 1888–1894, Mar. 2010, doi: 10.1021/es902909k.

[112] J. M. Allwood, M. F. Ashby, T. G. Gutowski, and E. Worrell, "Material efficiency: A white paper," *Resources, Conservation and Recycling*, vol. 55, no. 3, pp. 362–381, Jan. 2011, doi: 10.1016/j.resconrec.2010.11.002.

[113] A. M. M. S. Ullah, H. Hashimoto, A. Kubo, and J. Tamaki, "Sustainability analysis of rapid prototyping: material/resource and process perspectives," *International Journal of Sustainable Manufacturing*, vol. 3, no. 1, pp. 20–36, Jan. 2013, doi: 10.1504/IJSM.2013.058640.

[114] A. M. M. S. Ullah, A. Fuji, A. Kubo, and J. Tamaki, "Analyzing the Sustainability of Bimetallic Components," *International Journal of Automation Technology*, vol. 8, no. 5, pp. 745–753, Sep. 2014, doi: 10.20965/ijat.2014.p0745.

[115] A. M. M. S. Ullah, H. Hashimoto, A. Kubo, and J. Tamaki, “Sustainability analysis of rapid prototyping: material/resource and process perspectives,” *International Journal of Sustainable Manufacturing*, vol. 3, no. 1, pp. 20–36, Jan. 2013, doi: 10.1504/IJSM.2013.058640.

List of Achievements

Refereed Journal Papers:

[1] Saman Fattahi, Sharifu Ura, and Md. Noor-E-Alam, "Decision-Making Using Big Data Relevant to Sustainable Development Goals (SDGs)," *Big Data and Cognitive Computing*, vol. 6, no. 2, ar. no. 64, June 2022, doi: 10.3390/bdcc6020064. CiteScore 6.1 [Scopus and WoS Indexed]

[2] Saman Fattahi, Takuya Okamoto, and Sharifu Ura, "Preparing Datasets of Surface Roughness for Constructing Big Data from the Context of Smart Manufacturing and Cognitive Computing," *Big Data and Cognitive Computing*, vol. 5, no. 4, ar. no. 58, October 2021, doi: 10.3390/bdcc5040058. CiteScore 6.1 [Scopus and WoS Indexed]

Conference Proceedings Papers:

[1] Saman Fattahi, Sharifu Ura, and Angkush Kumar Ghosh, "Developing Big Data Analytics to Optimize Cutting Conditions of Machining Operations," *Proceedings of the 19th International Conference on Precision Engineering (ICPE 2022)*, Nara, Japan, November 28–December 2, 2022. [Paper Number C000291]

[2] Saman Fattahi and AMM Sharif Ullah, "Optimization of Dry Electrical Discharge Machining of Stainless Steel using Big Data Analytics," *Proceedings of the 15th CIRP Conference on Intelligent Computation in Manufacturing Engineering (CIRP ICME'21)*, Gulf of Naples, Italy, 14–16 July 2021. [Virtual Conference]. In *Procedia CIRP*, vol. 112, pp. 316–321, 2022, Elsevier, doi: 10.1016/j.procir.2022.09.004. CiteScore 3.9. [Scopus Indexed]

[3] Takuya Okamoto, Sharif Ullah, Aakihiko Kubo, Saman Fattahi, and Angkush Kumar Ghosh, "Preparing Big Data of Surface Roughness for Smart Manufacturing," *Proceedings of International Conference on Leading Edge Manufacturing in 21st century: LEM21*, vol. 2021.10, no. 0, pp. 009–110, 2021, JSME, doi: 10.1299/jsmelem.2021.10.009-110. [Presented in the 10th International Conference on Leading Edge Manufacturing in 21st Century (LEM21), November 14–18, 2021, Kitakyushu, Fukuoka, Japan] [Scopus Indexed]

Acknowledgments

Throughout the accomplishment of my doctoral study, research, and writing this thesis, I have received a great deal of support and assistance from many individuals and organizations. I acknowledge them as follows.

I sincerely thank my academic advisor, Professor Sharifu Ura, for his unwavering support and advice. He has given me all the freedom to pursue my research while silently and non-obtrusively ensuring that I stay on course and do not deviate from the core of my research. With his able guidance, this thesis was possible. I also express my sincere gratitude to my deputy academic advisors, Professor Yohei Hoshino, and Professor Kazuhiro Hayashida, for their invaluable support.

I express my deep appreciation to my thesis examination committee members, Professor Sharifu Ura (Chair), Professor Yohei Hoshino, Professor Kazuhiro Hayashida, Professor Noriaki Miura, and Professor Hiroshi Masui, for generously offering their time, support, guidance, goodwill, and invaluable advice throughout the preparation and review of this thesis.

I thank Dr. Angkush Kumar Ghosh for his help and advice during my research. My labmates also cooperated well during this period.

Last but not least, I want to express my deep appreciation to my beloved wife and daughter, who emotionally supported me.

This page is kept blank intentionally.