DATA-BASED MANUFACTURING PROCESS OPTIMIZATION USING IMAGE DATA

Case: Veneer and LVL

Lappeenranta–Lahti University of Technology LUT

Master of Science in Technology, Master's thesis

2022

Ahsan Muneer

Examiner(s): Professor Pasi Luukka

Post-doc Researcher Jyrki Savolainen

ABSTRACT

Lappeenranta–Lahti University of Technology LUT

LUT School of Engineering Science

Computational Engineering (Business Analytics)

Ahsan Muneer

Data-Based Manufacturing Process Optimization Using Image Data - Case: Veneer and LVL

Master's thesis

2022

89 pages, 37 figures, 7 tables and 0 appendix

Examiner(s): Professor Pasi Luukka, Post-doc Researcher Jyrki Savolainen

Keywords: Industry 4.0, Smart manufacturing, Big Data, IoT, Data Mining, Image Processing, Textural feature extraction, Veneer drying, LVL

Manufacturing industries are continually exploring ways to optimize their manufacturing processes to improve production by maximizing raw materials usage and minimizing operational costs. Industry 4.0 brings a digital revolution to the manufacturing industries in which Machine learning, IoT, Big Data, Data Mining, and Image processing play a vital role in optimizing manufacturing processes. This thesis aims to examine how the manufacturing process in the veneer/LVL industry can be optimized by using data from veneer sheets and applying machine learning methods. To achieve the thesis objective, a comprehensive literature review has been conducted of the existing methodologies used in the veneer/LVL industry.

In the empirical part of this thesis textural image data of veneer sheets is extracted and utilized to identify similar veneer sheets images after the drying process. Veneer sheets contain different variations in the textures; this thesis studied approaches to select the sheets containing high textures as well as low textures. A model has been developed in this thesis to select the candidate dry sheets with high and low textures after the drying process. Three approaches have been studied with Gray-Level Co-Occurrence Matrix (GLCM) and Canny edge detection methods on the image data set, and 20 features have been extracted from the image's texture. Future work and research required for further analyse the veneer sheet images with different edge detection methods are discussed.

ACKNOWLEDGEMENTS

When COVID-19 was at its peak and the whole world was facing lockdowns and challenges, at that time, securing admission at LUT University on a scholarship was not less than a miracle for me. I want to express my warmest gratitude to the LUT University and its faculty, especially those teachers who taught us different subjects during the entire master's program.

After graduation studying at an international university was a dream come true which had finally been fulfilled at LUT University after a gap of six years from when I left the studies. I am confident that the experience I gained while studying at LUT University will accelerate my career.

I want to express my profound gratitude to my supervisor Jyrki Savolainen for considering me for this thesis. During this thesis, he had not just guided me but also provided me with his research experience and gave his utmost fullest availability wherever I needed him for the discussions or the implementation and research. I am also thankful to Raute Corporation for giving me the thesis opportunity and required data. I am pretty confident that the findings in this thesis will also help me to further pursue the research direction.

Lastly, I would like to express my sincere thanks to my parents and wife. There are not enough words to describe how your support has helped me throughout these years. Thank you for always motivating, guiding, and encouraging me in my most difficult times.

Lappeenranta, 20.06.2022

Ahsan Muneer

SYMBOLS AND ABBREVIATIONS

Abbreviations

AI	Artificial Intelligence
AM	Additive Manufacturing
ANN	Artificial Neural Network
APC	Advanced process control systems
ASM	Angular second moment
AUC	Area under the curve
BDA	Big Data and Analytics
CAD	Computer aided design
CAM	Computer aided manufacturing
CAPP	Computer aided processing planning
CASOA	Cloud assisted self-organized architecture
CBM	Cloud-based manufacturing
CC	Cloud Computing
CIM	Computer integrated manufacturing
CNC	Computer numerical control
CNN	Convolutional Neural Network
CPS	Cyber physical systems
CSF	Connected smart factory
DCS	Distributed control systems
DE	Difference entropy
DNNs	Deep neural networks
DSS	Decision support systems

DV	Difference variance
EDA	Exploratory data analysis
FKNN	Fuzzy k-Nearest Neighbor
FMS	Flexible manufacturing systems
GAN	Generative adversarial networks
GLCM	Gray-Level Co-Occurrence Matrix
GLVQ	Generalized learning vector
I4.0	Industry 4.0
ICT	Information and communication technologies
IDM	Inverse difference moment
IOC	Information measures of correlation
IoT	Internet of Things
KDD	Knowledge discovery in database
KNN	K-Nearest Neighbour
KPCA	Kernel principal component analysis
LDA	Linear Discriminant Analysis
LVL	Laminated veneer lumber
MA	Metaheuristic algorithm
MaaS	Manufacturing-as-a-Service
MCC	Maximum correlation coefficient
ML	Machine learning
MLP	Multilayer perceptron
MSPC	Multivariate statistical process control
PCA	Principal Component Analysis

PLS	Partial Least Squares
PLSR	Partial least squares regression
PM	Primitive maintenance
PMS	Process monitoring systems
RDBMS	Relational database management schemas
RNNs	Recurrent neural networks
SA	Sum average
SAD	Sum of absolute difference
SDCM	Software-defined cloud manufacturing
SE	Sum entropy
SED	Squared Euclidean Distance
SMOS	Smart manufacturing objects
SOA	Service-oriented architecture
SPC	Statistical process control
SPCA	Sparse principal component analysis
SPD	Social product development
SPE	Squared Prediction Error
SQL	Structure query language
SS	Sum of squares
SV	Sum variance
SVD	Singular Vector Decomposition
SVM	Support vector machine
t-SNE	T-distributed stochastic neighbour embedding

Table of contents

Abstract

(Acknowledgements)

(Symbols and abbreviations)

1 Introduction 12
1.1 Motivation and background 12
1.2 Objectives and research questions 12
1.3 Data and methodology 13
1.4 Structure of thesis 13
2 Theoretical background 15
2.1 Industry 4.0 paradigm 15
2.1.1 Internet of things 17
2.1.2 Intelligent manufacturing 19
2.1.3 Big data analytics 23
2.2 Theoretical core concepts 30
2.2.1 Process control and optimization 30
2.2.2 Multivariate statistical process control 31
2.3 Data mining 32
2.3.1 Data exploration 35
2.3.2 Data pre-processing and cleaning 36
2.3.3 Dimensionality reduction and de-noising 40
2.3.4 Reliability and validity 42
2.4 Data Management 43
2.4.1 Data Conversion 44
2.5 Machine learning methods 45
3 Literature review 48
3.1 Review methodology 49
3.1.1 Selection of research databases 49
3.1.2 Keyword selection 50

3.2 Results of the literature review 53
4 Methodology and Data 59
4.1 Production process – Peeling and drying 59
4.2 Data and Methodology 63
4.2.1 GLCM texture features used in this research 66
4.2.2 Edge detection and Fingerprints 71
5 Results 74
5.1 Result analysis 76
6 Conclusion and discussion 79
6.1 Future research 80
References 81

Figures

Figure 1. Capturing images of veneer sheets during peeling and drying process

Figure 2. Basic design principles of I4.0

Figure 3. Characteristics of cloud-based manufacturing

Figure 4. Big data challenges

Figure 5. Dig data applications in manufacturing industry

Figure 6. Steps for extracting deep insights from big data

Figure 7. Big data analytics methods

Figure 8. Big data analytics techniques

Figure 9. Process challenges

Figure 10. Applications of quality improvement

Figure 11. Knowledge discovery methods in the process industry

Figure 12. Process industry KDD

Figure 13. Categorization of data pre-processing methods

Figure 14. Data cleaning steps

Figure 15. Outlier detection methods

Figure 16. Transformation of Big data for data analysis

Figure 17. Data conversion from source to target

Figure 18. ML algorithm types and usage in the process industry

Figure 19. Deep learning architectures

Figure 20. Research processes, sub-processes, and its outcomes

Figure 21. Papers published by authors on veneer peeling

Figure 22. Papers published by authors on veneer drying

Figure 23. Methods frequently used in optimizing the drying process according to current literature

Figure 24. Overview of plywood and LVL manufacturing process

Figure 25. Pre-processing of the log before peeling

Figure 26. Overview of blocks to veneer sheet process

Figure 27. Overview of veneer sheets manufacturing

Figure 28. Overview of drying process

Figure 29. GLCM w.r.t four directions of edge detection

Figure 30. Proposed system for extracting and comparing image similarity presented in a block diagram

Figure 31. GLCM texture on different angle with distance of one pixel

Figure 32. Original veneer sheet image (left) and Fingerprint image of the same veneer sheet (right)

Figure 33. Candidates by using GLCM texture features of original sheet image

Figure 34. Low texture candidates by using GLCM texture features of original sheet image

Figure 35. Candidates by using GLCM texture features of fingerprint image

Figure 36. Consecutive veneer sheets similarity using fingerprints

Figure 37. Candidate images with distance d = (1,5,7,9) and angles (0°, 45°, 90° and 135°)

Tables

Table 1. EDA methods according to data type

Table 2. EDA methods according to objective

Table 3. Research publication platforms

Table 4. Numbers of papers related to the keyword search in each database

Table 5. Combination of search queries

Table 6. Papers according to the concept centric approach

Table 7. GLCM texture features

1 Introduction

1.1 Motivation and background

This thesis focuses on the peeling-drying process optimization in the veneer/LVL industry using data mining methodologies and machine learning (ML) algorithms for image data. In recent years terms related to industry 4.0 such as “big data analytics”, “IoT”, and “machine learning” have become more common because the size of available data for analysis is exponentially increasing while the cost of data capturing is decreasing. In the last few years, the volume of structured and unstructured data has increased exponentially (Vialetto and Noro, 2020). In industry, exponentially increasing use of process operations, process control systems and information systems generate massive volumes of data, making the existing manufacturing databases gigantic. Furthermore, with improvements and advancements in IoT and image processing, the data collected from the smart manufacturing process will likely expand in multi-dimensions.

In recent years, big data has received a lot of attention due to its significant impact on the optimization of manufacturing processes. Because of the widespread usage of distributed control systems (DCS) and improvements in information and communication technologies (ICT), industrial processes are increasingly running in an unpredictable and complex environment with challenging procedures and complicated constraints. (Belhadi et al., 2019).

1.2 Objectives and research questions

This thesis focuses on, first, how to optimize the manufacturing processes using Industry 4.0 big data analytics, IoT and machine learning. Secondly, a focused attempt is undertaken to develop a data-based method to identify similar veneer sheets after the drying process according to their textural features and grains. The motivation to identify veneer sheets is that implementing a full-fledged data-optimization of the veneer production process requires matching the dried (processed) sheet images with the original ones, which was only partially

possible at the time of writing this thesis (spring 2022). The Figure 1 shows the capturing of veneer images at different stages during the veneer/LVL manufacturing process.

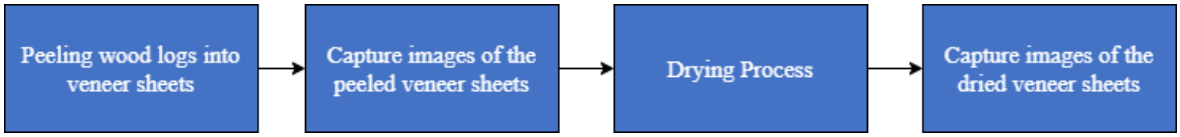


Figure 1. Capturing images of veneer sheets during peeling and drying process

The research questions are formulated as follows based on the above description:

1. *According to the current literature, how Machine Learning algorithms have been applied in the data-based optimization of drying process specifically in the veneer / LVL industry?*
2. *How to match peeling and drying sheet images to enhance the peeling-drying process in the given LVL-manufacturing case example?*

1.3 Data and methodology

This thesis is based on quantitative research, and the data of this study is provided by Raute Corporation, which is composed of two parts. The first part contains captured images of peeling and drying veneer sheets after the peeling and drying process, and the second part contains black and white, “fingerprint” images of the same sheets produced with edge detection algorithms. Key methodologies applied in this thesis are methods of data mining and machine learning models.

1.4 Structure of thesis

This thesis consists of six chapters, First Chapter is about the introduction of the thesis consisting of motivation and background, objectives and research questions, methodologies, and thesis structure. Chapter 2 concentrates on theoretical background and consists of five sub-chapters. The first sub-chapter describes the industry 4.0 paradigm and the three main

pillars of Industry 4.0 related to this project. The second sub-chapter explains the theoretical core concept for the multivariate process control and process controls optimization in the manufacturing industry. The third sub-chapter explains the core principles of data mining and their importance in big data analytics specifically in manufacturing 4.0. This sub-chapter focuses on studies related to the data exploration, data pre-treatment methodologies for the industrial data, dimensionality, and noise reduction methods and lastly, this sub-chapter explains the importance of reliability and validity of the data as project uses the data obtained from different IoT devices on time intervals. The fourth sub-chapter explains the stages of data management after the data mining to be used in ML methods. This sub-chapter focuses on Data acquisition, Data storage, Data conversion, Feature engineering and data representation. Lastly, the fifth sub-chapter describes the relevant Machine learning methods from the literature.

Chapter 3 consists of the Literature review related to industry 4.0. The first sub-chapter describes the followed Review methodology for the literature review related to the topic of this thesis. Chapter 4 consists of the description of the data and the methodologies applied in the manufacturing process. The first sub-chapter briefly explains the current manufacturing process from inputting of woodblock, its peeling, quality evaluation, and final output of the dried veneer sheets. The second sub-chapter thoroughly explain the data structure and data types. How data is being collected and transformed from different data points such as peeled wet sheets images, sensors (temperature, humidity, light, pressure) and dry sheet image. Chapter 5 explains and interprets the results and the characteristics of the evaluation criteria, how the benchmarks are set, and quality evaluation decisions are made.

The last chapter, Chapter 6 discusses key findings during and at the end of project implementation, answering the research questions, limitations, and lastly, the future research potentials to optimize the manufacturing process using big data analytics, IoT and image processing.

2 Theoretical background

2.1 Industry 4.0 paradigm

In 2013, the fourth industrial revolution was initiated by Germans as a decisive initiative to take an innovative position in industries that are revolutionizing the manufacturing sector. I4.0 has emerged as a prospective technological paradigm for extending and integrating industrial processes at inter- and intra-organizational levels, radically changing the manufacturing sector and its economic environment. The previous three industrial revolutions focused on increasing production led by mechanization, electrical energy, and information technology (IT) (Veza et al., 2015). During the first industrial revolution, the age of steam, manufacturing capacities were established with the use of water and steam energies. In the second industrial revolution, the age of electricity, electrical energy was used to accomplish mass manufacturing in the industries.

The third industrial revolution, the age of information, digital and information technologies, broadened production automation. Flexible manufacturing systems (FMS) were introduced in this industrial revolution, which was made possible by adopting computer numerical control systems (CNC) and industrial robots. Computer integrated manufacturing (CIM) is made possible with computer aided processing planning (CAPP), computer aided manufacturing (CAM), and computer aided design (CAD) applications. In the fourth industrial revolution, Cyber physical systems (CPS) introduced a paradigm change in industries and businesses to improve productivity, particularly in the manufacturing industry (Lasi et al., 2014; Xu et al., 2018).

The use of intelligent machines and technologies to digitally modify manufacturing processes is relatively new. Rapid advancement in information and communication technologies (ICT) has led to the development of I4.0. The underlying technology behind I4.0 is CPS, which makes manufacturing systems modular and flexible, allowing mass production of highly customized products. The development and technological advancement in I4.0 will provide a series of feasible solutions for the expanding need for informatization in the manufacturing industry. This viability is demonstrated by the fact that an increasing number of businesses are investigating the benefits of digitizing their horizontal and vertical

supply chains by adopting I4.0 to become leading digital businesses in tomorrow's complex industrial manufacturing ecosystems (PWC, 2016). Due to swift and revolutionary changes in the industry, information is transferred in a short period, globalizing the markets, and resulting in intense competition between companies that were never observed before (Rajnai and Kocsis, 2018).

The main objective of I4.0 is to make traditional factories and manufacturing processes intelligent to achieve higher levels of operational efficiency, productivity, and automation. By incorporating sensors, autonomous systems, and actuators into the manufacturing process, I4.0 makes factories smarter, more dynamic, and adaptive. As a result, machines and equipment can self-optimize and automate to a high degree. In addition, the production process is capable of meeting more complicated requirements and product quality standards (Roblek et al., 2016).

Implementation of I4.0 must be interdisciplinary with close contact between different vital areas (Alcácer and Cruz-Machado, 2019). Multiple technologies are available for implementing I4.0: Additive Manufacturing (AM), Artificial Intelligence (AI), Automation and Industrial Robots, Big Data and Analytics (BDA), Blockchain, Internet of Things (IoT), Cloud Computing (CC), Cyber-Physical Systems (CPS), Industrial information integration, Simulation and Modelling, and Visualization. With enabling technologies, tools, and methods, I4.0 helps in reducing manufacturing costs and improving time efficiency and product quality. As a result, I4.0 will accelerate the manufacturing industry to achieve exceptional operational efficiency and improved productivity (Lu, 2017).

I4.0 is an approach for developing an interaction system enabling manufacturing production lines and products to communicate through a connected smart factory (CSF); there are six underlying design principles of I4.0, as in Figure 2 (Nascimento et al., 2019). CSF is a methodology designed to construct a highly interlinked network-based integrated manufacturing model that provides real-time monitoring and autonomous control of the whole manufacturing process. Optimizes supplies, energy, and raw material usage and adds value through the coordinated cooperation of products and services, resulting in low production costs and high-value products. CSF is the outcome of applying a new paradigm to the industry called ICT-based smart (manufacturing, building, grid, etc.), where machines enable automatic manufacturing through simulation (Park, 2016).

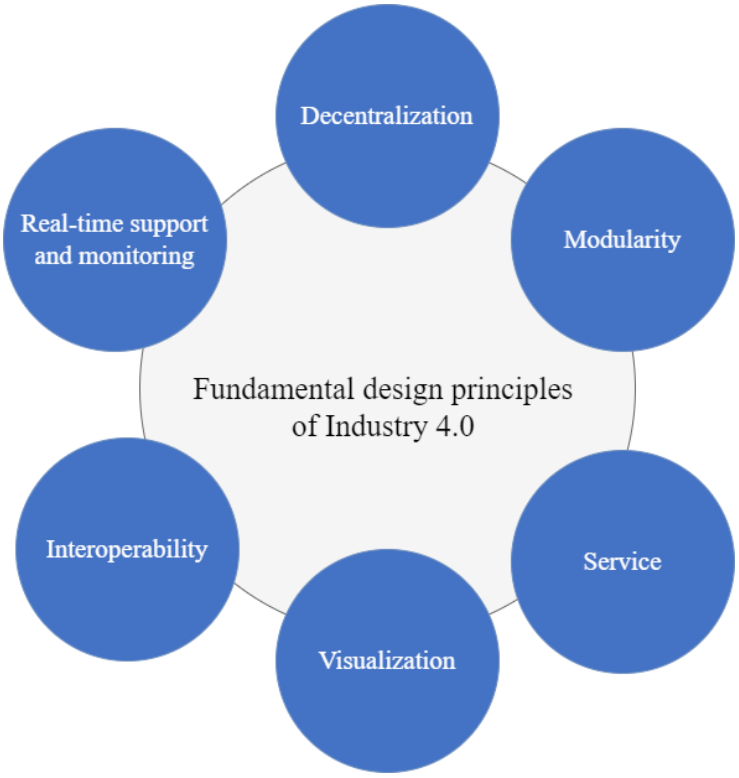


Figure 2. Basic design principles of I4.0 (Adapted from Nascimento et al., 2019)

CPS communicates over the IoT to connect infrastructures, machines, human users, and processes across organizational units, enabling integration between the physical and virtual environments by using sensor devices, actuators, and computational resources to broadcast data in real-time for decentralized decision-making processes (Trappey et al., 2016).

2.1.1 Internet of things

(Sezer et al., 2018) describes the Internet of Things as *"IoT allows people and things to be connected anytime, anyplace, with anything and anyone, ideally using any path/network and any service."* According to (Bortolini et al., 2017), IoT is the ubiquitous presence with the common goal of different objects interacting and collaborating to digitize physical systems. The internet has dramatically changed the way how we live, allowing people to communicate in a virtual environment in several contexts covering professional and social relationships. IoT is adding a new dimension to this process by enabling connectivity to and between smart

devices, thus leading to the goal of communicating *"anytime, anywhere, any media, anything."*

The potential offered by the IoT enables the development of a vast number of applications, only a tiny fraction of which are currently available to our society. Most of these are the domains and environments in which these applications of IoT are improving the quality of our lives, such as smart manufacturing, smart factories, e-health, smart homes, smart cities, and energy management. According to a UN report, a new age of ubiquity is approaching, in which internet traffic generated by humans will be overshadowed by the networking of everyday objects (Atzori et al., 2010).

The internet of things is a paradigm that is exponentially gaining popularity in today's wireless telecommunications era; the main idea behind IoT is the ubiquitous presence of various types of objects around us, such as sensor devices, radio-frequency identification (RFID) tags, mobile devices, and actuators, that can communicate and cooperate to achieve common goals using unique addressing protocols (Giusto et al., 2010). The development of IoT and I4.0 is interlinked; in fact, IoT is the trend and direction of the I4.0. IoT comprises two terms: Internet and Things; The first refers to a network-oriented vision, while the second focuses on the integration of things to share data by following standard protocols (Atzori et al., 2010; Gubbi et al., 2013).

(Gubbi et al., 2013) presents a taxonomy from a high-level perspective to define the required components of the Internet of Things. There are three major components, Hardware, Middleware, and Presentation, that enables the unified ubicomp for IoT. Hardware is based on communication devices, actuators, and sensors, middleware is based on computing tools and storage, and presentation is based on visualization. Visualization provides deep insights and concise interpretation tools that can be accessed on various platforms and can be tailored for specific applications. Visualization is critical for IoT applications because it lets users interact with the environment to monitor the process and make decisions in real-time.

IoT can be added to the historical list of forces that drove last three industrial changes (mechanization, electricity, and information technology). The changes that I4.0 brings with itself will make it the global language for smart manufacturing. I4.0 guarantees to improve and increase the manufacturing productivity by fifty percent and to reduce the utilization of required resources to half (GTAI, 2018). Digitalization of information can be utilized to

change production patterns by using virtual models of the current physical environment and sensor data to meet manufacturing objectives (Peruzzini et al., 2017). Networking of the products over the IoT enables the whole manufacturing process to connect, transforming the factories with different value streams into a smart environment (Kerin and Pham, 2019).

2.1.2 Intelligent manufacturing

Traditional enterprises are facing new business challenges due to globalization, mass customization and competitive business environments (Simmert et al., 2018). I4.0 introduces new opportunities that can potentially change the current manufacturing processes in the companies. With the acceptance and growth of digital technologies over the past years, technologies can affect different processes differently; some techniques have cross-sectional effects on all processes, while some focus only on one specific process. The area which is highly impacted by I4.0 is manufacturing, enhancing the production process to optimize the operational performance, product or service development and supply chain planning (Zheng et al., 2020).

Intelligent manufacturing is a multidimensional concept that optimizes production processes through advanced data analysis, manufacturing technologies, and system engineering. It is a new manufacturing model to improve the design, production, management, testing, and integration of a product throughout its life cycle with the implementations of intelligence science and technology. In I4.0, an integrated management system (IMS) employs service-oriented architecture (SOA) to offer collective, adaptable, reconfigurable, and customizable services to end-users over the network, resulting in a highly unified human-machine manufacturing architecture.

The agent paradigm is observed as one of the most effective ways of achieving intelligent manufacturing (Lu, 2017). The word *"agent"* refers to a process or entity designed to accomplish a task constantly and independently in a non-deterministic context with other processes and elements. Agents perform tasks in a condition from which they are separated and have their own knowledge and understanding of their surrounding environment; they employ preference in interacting within their environment, developing plans, making autonomous choices, and performing actions to change the environment (Adeyeri et al., 2015). The multi-agent system describes autonomous and intelligent entities as agents or

entities that can work together to achieve global tasks. The fundamental requirement for collaboration arises from the reality that agents lack sufficient information to make decisions that can be applied globally (Tang et al., 2017). The author presented a cloud-assisted self-organized architecture (CASOA) based on smart agents and the cloud computing to communicate and interact over the network for agent-based manufacturing in the perspective of I4.0 and classified agents into four types according to their functionalities: machining agents, product agents, suggestion agents, and conveying agents.

AI performs an integral part in IMS by offering particular features such as learning, analysing, and responding. Human participation in the information management system can be minimized by applying AI. For example, raw material and production components can be organized automatically, and production operations can be controlled and monitored in real-time. With the continuous acceptance of I4.0, autonomous sensing, intelligent decision-making, smart interconnectivity, learning, and analysis are becoming a reality (Zhong et al., 2017).

Intelligence and digitalization are applied from the procurement of raw material to the production system, product consumption, and the end of the product's life (Lu, 2017). I4.0 provides horizontally and vertically value-added integration in the industrial process. The horizontal approach integrates value-creation modules across the entire product lifecycle, from material flow to logistics. Whereas the vertical approach combines product, machine, and user requirements with several aggregation levels of value-creation and production systems (Castelo-Branco et al., 2022; Lu, 2017; Veza et al., 2015). (Shafiq et al., 2015) stated three approaches of integrations in the industrial process, horizontal, vertical, and end-to-end, to achieve the design principles of I4.0. The author also presented three levels of integration at which virtual manufacturing can be integrated: virtual processes, virtual devices, and virtual factories. The integration of these three virtual levels will assist in building the architecture of I4.0 in order to achieve higher levels of smart machines, industrial automation, and semantic analytics.

End-to-end digital integration refers to a universal digital engineering concept that reduces the gaps between product designing, production, and the customer. There are two approaches, application pull and the technology push, through which I4.0 drives manufacturing. Technology push requires higher degrees of automatization, digitalization,

networking, and miniaturization, whereas application-pull generates dynamic changes as a result of a new generation of industrial infrastructure (Lasi et al., 2014).

IoT plays an integral part in the architecture of intelligent manufacturing; IoT-enabled manufacturing is a cutting-edge approach that transforms traditional manufacturing devices into smart manufacturing objects (SMOs) that can connect, communicate, and interact with one another to execute manufacturing logic automatically and adaptively (Zhong et al., 2017). The manufacturing process in I4.0 will require more microcontrollers, actuators, sensors, autonomous systems, advanced methods of data analytics, CPS, and ECM due to the rapid development of technologies (Lasi et al., 2014; Lu, 2017).

Cloud-based manufacturing (CBM) is a technology that has the potential to considerably assist in the implementation of I4.0. Figure 3 shows the characteristics of cloud-based manufacturing. Similar to cloud computing, CBM uses resources that are highly distributed over a network which is increasing acceptability of the Manufacturing-as-a-Service (MaaS) in the industry (Xu et al., 2018). Cloud technologies are helpful towards the implementation of CBM by minimizing the setup and maintenance cost and increasing scalability through virtual resources (Nascimento et al., 2019). (Thames and Schaefer, 2016) presents software-defined cloud manufacturing (SDCM) based on the network of software and hardware elements communicating over a TCP/IP. The purpose is to use elements that make up a I4.0 system, such as an Industrial internet of Things (IIoT), CBM, or social product development (SPD), individually or in combination.

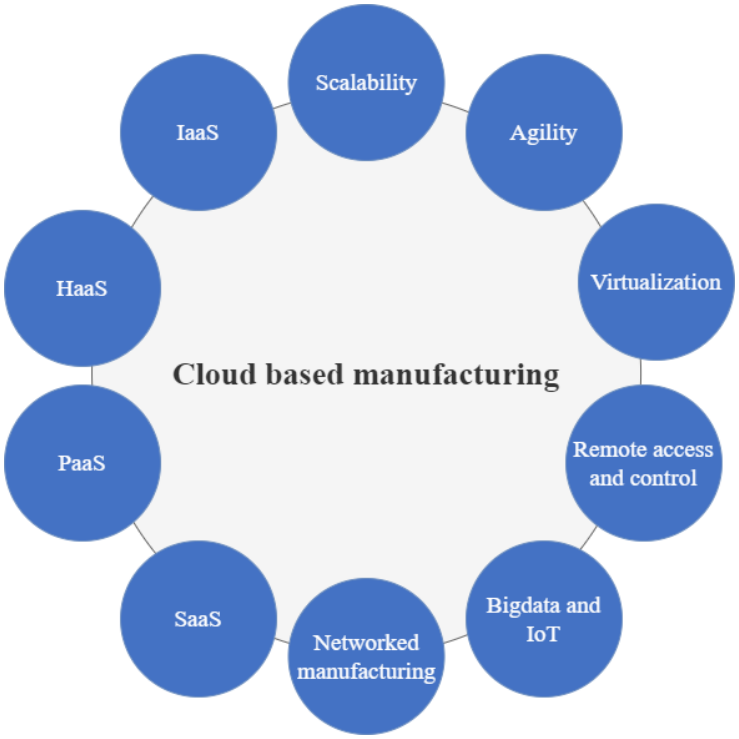


Figure 3. Characteristics of cloud-based manufacturing (Adapted from Thames and Schaefer, 2016)

It is crucial to understand that a company's digital transformation is not just about investing in the advancement and implementation of technology. It is also based on the business strategy and both internal and external business processes. Digitalization in an organization is not a one-time event but an advancement through several steps including technological and organizational changes (Rajnai and Kocsis, 2018). (Peruzzini et al., 2017) points out that factories are made up of more than just machines; there are also human operators in the factories that collaborate with machines and each other in various ways, such as performing jobs, regulating, and monitoring the process, and interpreting machine results. People may therefore be seen as "things" in a smart factory to be monitored and connected to one another and the machines. Despite advancements in production line automation, humans perform essential roles in factories and are primarily responsible for smooth production and high product quality.

2.1.3 Big data analytics

The amount of data generated by industrial systems has been rapidly increasing as a result of developments in IoT, 5G, and CC technologies. Improvements and advancements in product design, manufacturing processes, and primitive maintenance (PM) have been achieved because of the effective use of industrial data. Big Data Analytics has been identified as a vital component in the development of intelligent industrial systems (Wang et al., 2022). Humans, machines, and nature all contribute to the generation of data. With the advancement of technologies and services, an enormous volume of data is generated from several sources, which might be structured, semi-structured, or unstructured (Ishwarappa and Anuradha, 2015).

I4.0 brings a new era of the industrial revolution, which accelerated the integration of information technology with the industrial systems, and enterprise data has become increasingly rich (Wang et al., 2022). The continuous manufacturing process, multiple sensors, and real-time data generation and transfer make the data have three characteristics (3 V's): Volume, Variety, and Velocity (Belhadi et al., 2019; Gandomi and Haider, 2015; Hashem et al., 2015; Jagadish, 2015; Kerin and Pham, 2019; Philip Chen and Zhang, 2014; Wang et al., 2022). These 3 V's were the initial suggestions for characterizing big data as these are the three main aspects that surfaced as a base framework for data management and manipulation challenges.

Other dimensions, as in Figure 4 big data challenges, including volatility, variation, veracity, validation, verification, vision, and value have been sought to assign for a better characterization of big data in order to continually analyse a substantial volume of unstructured data acquired from several different sources (Alcácer and Cruz-Machado, 2019; Belhadi et al., 2019; Gandomi and Haider, 2015; Hashem et al., 2015; Ishwarappa and Anuradha, 2015; Sivarajah et al., 2017). (Wang et al., 2022) stated that industrial data has also been characterized as multi- sources, multi-dimensions, multi-noise, imbalanced, and time series.

Figure 4. Big data challenges (Adapted from Sivarajah et al., 2017)

The volume presents the most immediate challenge to conventional IT structures. The main benefit of big data analytics is the advantage gained from the potential to process enormous volumes of data (Ishwarappa and Anuradha, 2015). The structural heterogeneity in data is called variety (Gandomi and Haider, 2015) generated in multiple formats from multiple sources based on multidimensional data features (Alcácer and Cruz-Machado, 2019). Advancement in technology allows organizations to process different types of structured, semi-structured and unstructured data. Structured data consists of five percent of all existing data in the form of tables available in worksheets or relational databases (MySQL, Microsoft SQL Server, Oracle etc.). Raw text, audios, images, and videos are examples of the unstructured data that need to be transformed into a structured form for the analysis by the machines. Manufacturing industries have been collecting such data from internal resources (sensors, production lines) and external sources (marketing campaigns, social media). With the development of new data management and data analytics tools enables the companies to utilize data in their business processes, which is one of the pioneering aspects of smart manufacturing (Gandomi and Haider, 2015).

Velocity refers to the rate at which data is being generated as well as the rate at which it should be examined and acted upon (Alcácer and Cruz-Machado, 2019; Belhadi et al., 2019; Hashem et al., 2015). Veracity represents the inaccuracy or irregularity in the data, such as outliers, noise, and missing values (Alcácer and Cruz-Machado, 2019). IBM introduced this term as the fourth V for the big data characteristic. For example, Sentiments on social media are unreliable since they involve an individual's judgement. They do, however, include useful information resulting in the requirement to deal with inaccuracy and uncertainty in the data as an additional aspect of big data that can be addressed by utilizing tools, methodologies, and analytics designed for the mining and management of uncertainty in the data (Gandomi and Haider, 2015).

Variability is the term used to describe the variation in data flow rates (Alcácer and Cruz-Machado, 2019; Belhadi et al., 2019; Gandomi and Haider, 2015; Sivarajah et al., 2017). The SAS introduces variability and complexity as two new dimensions of big data (Belhadi et al., 2019; Gandomi and Haider, 2015). Velocity of big data is often inconsistent, with frequent peaks and troughs. The complexity of big data is defined by the fact that it is generated from different sources, which poses a significant challenge in connecting, matching, cleaning, and transforming data generated from multiple sources (Gandomi and Haider, 2015).

Value is the most essential characteristic of big data. According to Oracle's (2021) definition, BD is generally characterized by *"low-value density."* That is, when data is received in its raw form, it does not have much value as compared to the volume of the data. However, analysing vast volumes of such data can yield a high value (Belhadi et al., 2019; Gandomi and Haider, 2015).

Data-driven and model-driven approaches are the two primary paradigms for BDA. Model-driven paradigm refers to the process of developing a solid understanding of how a physical system works. It is an effective technique based on a detailed understanding of a system and can take advantage of scientifically proven relationships. Data-driven is a model-free approach and is based on the correlation between the parameters of a system indicating its status and estimates predicted by different ML methods with high accuracy for optimization (Wang et al., 2022). Due to the widespread usage of distributed control systems in the manufacturing industry over the last few decades, massive volumes of data have been generated. It can be challenging to develop first-principal models because of their immensely

complex processes, whereas data-driven process modelling, monitoring, control, and prediction have recently gained significant attention (Ge et al., 2017).

Figure 5 illustrates applications in manufacturing industry that are based on data driven paradigm. Industrial big data allows industries to precisely perceive changes in the system's internal and external environment by enabling scientific analysis and decision-making to improve productivity, reduce costs, and enhance operating efficiency. Industrial data can be seen as a way to promote smart manufacturing, which in results brings new business models to promote socioeconomic development (Wang et al., 2022).

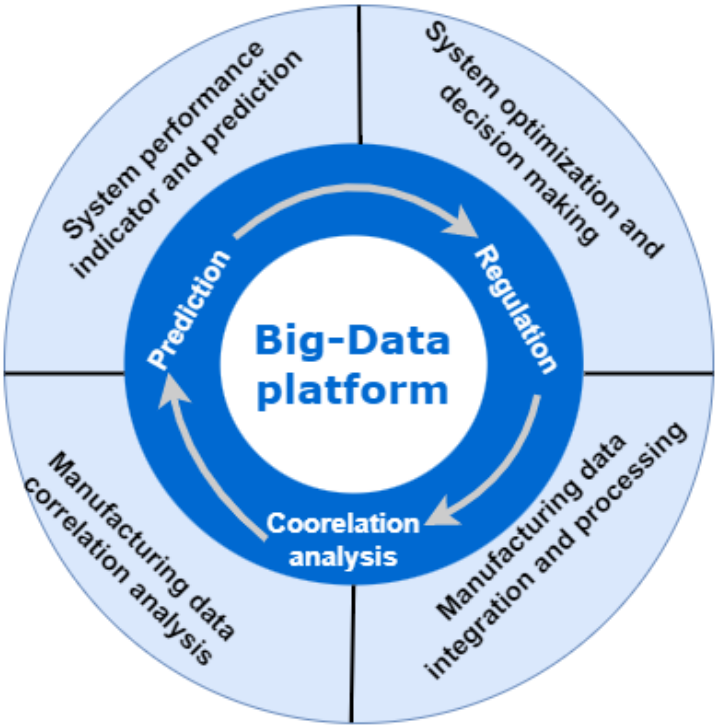


Figure 5. Dig data applications in manufacturing industry (Adapted from Wang et al., 2022)

Organizations require effective methods to transform enormous amounts of fast-moving and diverse data into actionable insights to make these data-driven choices. Figure 6 shows the stages of extracting deep insights from big data. Data management and analytics are two critical sub-processes of extracting insights from big data. The methods and technologies used to capture, store, process, and retrieve data for the analysis are referred to as data

management. Whereas analytics refers to the techniques for analysing and extracting information from large amounts of data.

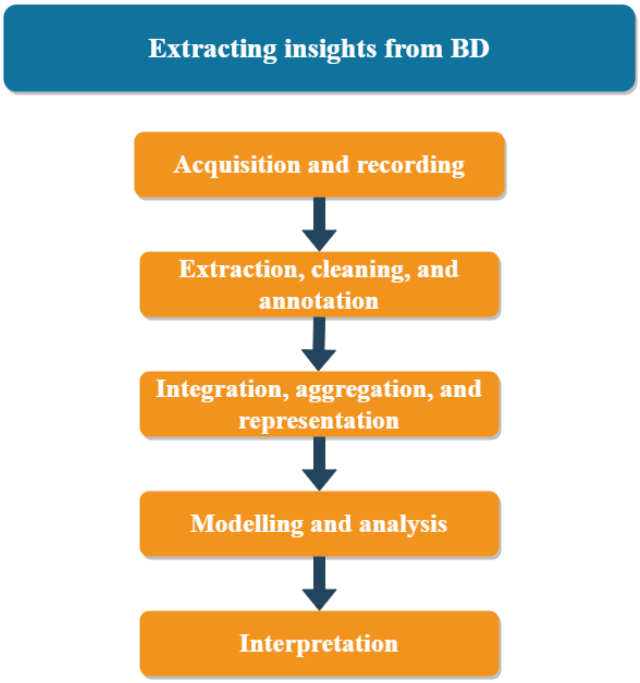


Figure 6. Steps for extracting deep insights from big data (Adapted from Gandomi and Haider, 2015)

Big data is defined by data collection and storage, but data analysis and extracting meaningful insights from unstructured data is its core feature, and without it, big data is useless. Big data can provide systematic guidance for manufacturing processes over the entire product lifecycle, ensuring cost-effective and fault-free operation and assisting management with decision-making and solving problems related to the manufacturing process. The ability to generate value through big data provides advantages to businesses (Alcácer and Cruz-Machado, 2019).

Advanced data analysis is required to explore data and it depends on the type of insights required as in Figure 7. Offline and real-time data can be analysed and mined using cloud computing (CC) and by applying advanced analytics methodologies: machine learning forecasting, semantic analysis, sentiment analysis, network analysis, and simulation. Data insights derived from a large amount of data let manufacturers understand the various

product lifecycle stages. Furthermore, advanced big data analytics can be utilized to detect and resolve bottlenecks caused by data generated by IoT devices in intelligent manufacturing (Alcácer and Cruz-Machado, 2019).

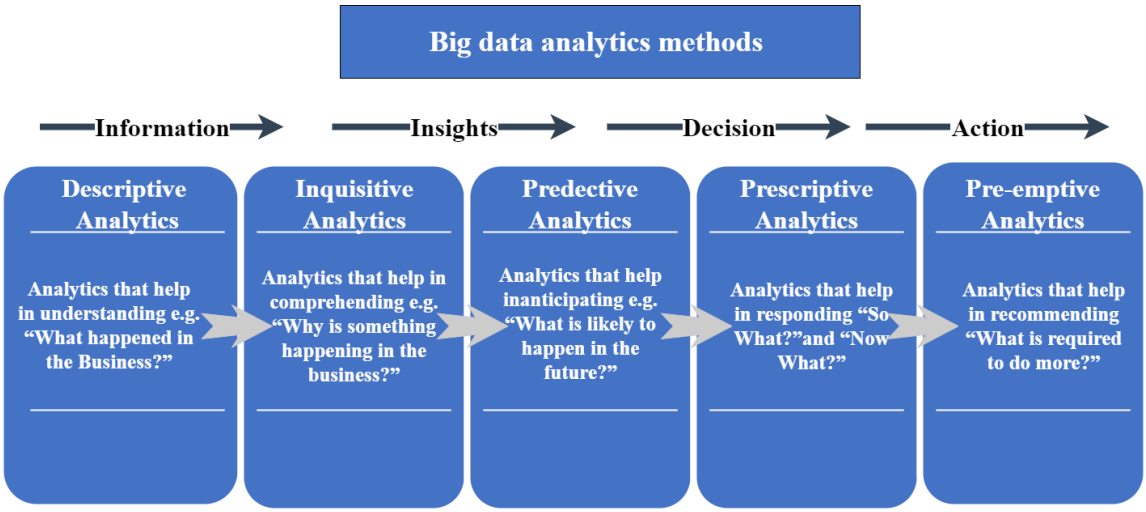


Figure 7. Big data analytics methods (Adapted from Sivarajah et al., 2017)

In order to promote evidence-based decision-making, Organizations require efficient strategies for processing vast amounts of heterogeneous data into meaningful data insights. Big data can improve decision-making and improve organizational production; by applying analytical methods to extract deep insights from data (Sivarajah et al., 2017), as shown in the Figure 8.

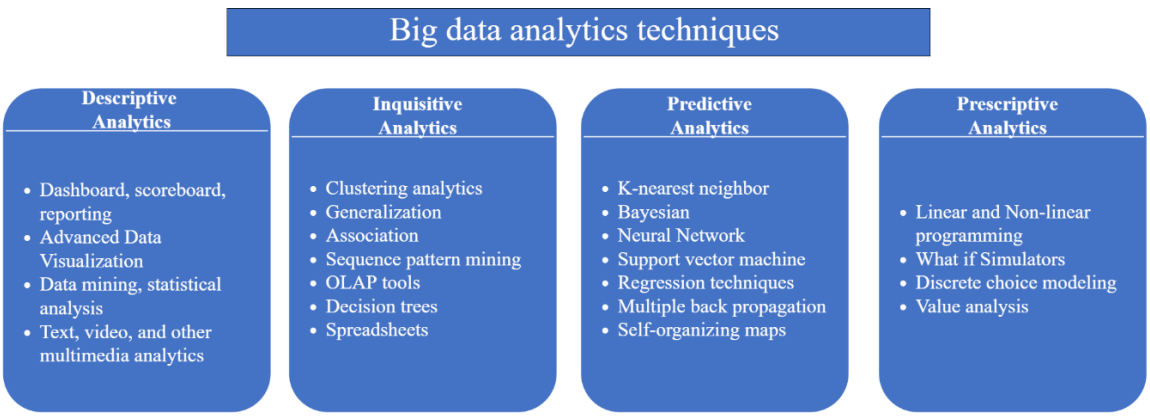


Figure 8. Big data analytics techniques (Adapted from Belhadi et al., 2019)

The benefits of big data are significant and undeniable, but there are plenty of challenges that must be handle before the full potential of big data can be realized. These challenges in big data can be categorized into three groups (Sivarajah et al., 2017).

1. Data challenges in big data are related to the data characteristics, commonly known as the V's of big data, as in Figure 4.

2. Process challenges are the collection of issues that appear during the data processing and analysis that can arise during data collection, interpretation or pre-processing and presentation of the results. These challenges are related to the how approaches: How to collect, integrate, transform, select a suitable model, and interpret the findings. Process challenges are divided into five steps, as shown in the Figure 9.

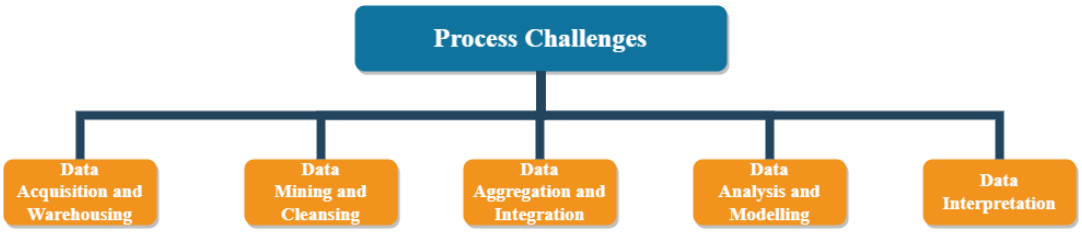


Figure 9. Process challenges (Adapted from Wang et al., 2022)

3. Management challenges arises during retrieving, handling, and governing of data, such as privacy of the data, data sharing, and data ownership. Personal data, financial transactions, or other sensitive data belonging to individuals and organizations are stored in databases. Organizations must guarantee a secure infrastructure that allows each employee and staff to only access relevant data for which the employee is authorized.

BDA has been significantly improved with the developments in AI. With the implementation of AI, industrial data can be effectively explored for intelligent manufacturing. Continuous learning from big data enables a manufacturing process to learn, optimize, and regulate itself (Wang et al., 2022). Furthermore, as manufacturing industries become more complicated and knowledge-intensive, enormous amount of data is generated from I4.0 applications. The drawbacks of heterogeneous data will impede industrial advancement. As a result, big data

management has become a significant problem; depending on the security and safety protocols, CC architecture can be utilized to overcome data management issues. Data mining, machine learning, and CC will determine the future of big data analytics (Lu, 2017).

2.2 Theoretical core concepts

One of the main goals of industrial operations that handle biomaterials of varying quality, such as wood, is to use effective control systems to reduce the unpredictability of the end products and ensure that they meet the required quality standards. For this purpose, Industrial facilities use statistical process control (SPC) and multivariate statistical process control (MSPC) to monitor process flows continually and adjust process parameters as needed (Lestander et al., 2012).

2.2.1 Process control and optimization

By analysing patterns of the manufacturing data and correlations between features, valuable information can be retrieved. Statistical models for several applications, including process monitoring, fault detection, and quality factor indicators, can be implemented using such information (Ge et al., 2017). In manufacturing and chemical industries, SPC techniques have become essential for monitoring the performance of a process to detect any unusual events. Improvements in the manufacturing processes and product qualities can be achieved by identifying and removing issues (Kourti et al., 1996). After training and validating a model, the trained model can be applied for online or offline process control applications to monitor a manufacturing process's operational status by fault diagnosis, fault classification, dimensionality reduction, trend analysis, and quality prediction.

Squared prediction error (SPE) and Hotelling's T2 are commonly implemented in multivariate process control methods. The process's desirable and undesirable operating conditions can be distinguished by setting control limits for these monitoring indicators. The objective of fault diagnosis is to provide a detailed explanation for the indicated fault in an industrial process. Based on the fault detection method applied, the underlying cause of the defect can be identified in a specific component, sensor, or actuator (Ge et al., 2017). The

majority of quality improvement applications are categorized into five types, as shown in Figure 10.

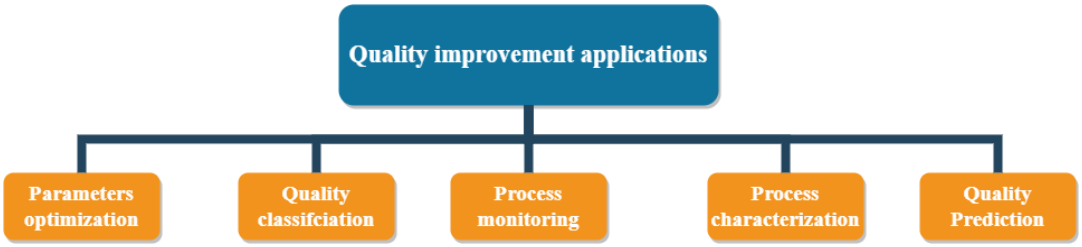


Figure 10. Applications of quality improvement (Adapted from Cheng et al., 2018)

Parameters optimization (see Fig. 10) provides the optimal number of process related parameters to deliver the required quality based on learned features of a high-quality product or a process. These parameters can be utilized as control chart boundaries or assist in developing an accurate model. Quality classification divides product quality into multiple levels and employs effective measures to improve the quality of a product based on the level predicted ahead of time. Process monitoring monitors the production process and detects any unusual patterns. It supports identifying and resolving the source of signals outside the control limits or the point where quality changes as soon as possible. Processes characterization identifies features or parameters that impact quality, then classifies these features or parameters according to their level of significance. When the output is a real value variable, quality prediction builds a model that correlates the input variable features with the output and uses it to forecast the quality of a given set of input parameters (Cheng et al., 2018).

2.2.2 Multivariate statistical process control

In the manufacturing industry, there are several cases where concurrent monitoring or handling of two or more process quality components is required. Multivariate statistical process control refers to scenarios in which numerous interlinked process variables of interest are monitored (Bersimis et al., 2009). When multiple monitoring parameters are

included to control micro fluctuations in the manufacturing process, SPC becomes more complicated, potentially resulting in information overload for process operators.

In order to reduce a large set of variables into a small number of components, MSPC uses multidimensional data analysis techniques, and shows the results in charts along with process limits. This allows the monitoring system to integrate a large number of parameters while maintaining a broad perspective of enormous datasets (Bersimis et al., 2009; Lestander et al., 2012). PCA and PLS are multivariate statistical projection methods that can handle enormous volumes of data and compress the information into low-dimensional latent variable components, making monitoring and interpreting results more manageable (Kourti et al., 1996). Any multivariate process control approach must fulfil the four conditions: (i) Is the process in control? (ii) An overall likelihood for the occurrence of event "Process detects an out-of-control erroneously condition" (iii) correlation among the variables (iv) What is the problem if the process is out-of-control (Bersimis et al., 2009).

(Lestander et al., 2012) uses PCA and PLSR to simulate an effective MSPC based on data from the wood pellet manufacturing process to classify deviations in the supervised parameters over a period in order to forecast wood pellet dryness and to identify the possibilities of using MSPC for monitoring and predictions in the wood pellet industry.

(Tiryaki and Aydin, 2022) did a study on multivariate Hotelling T2 statistical process control in terms of various quality characteristics that are used to monitor the manufacturing process of medium density fiber boards. The quality attributes that contributed the most to each signal were identified by decomposing T2 values.

2.3 Data mining

Data mining is a process of extracting information from data by applying machine learning techniques. Data generated from the manufacturing process can be divided into three categories: operational data, non-operational data, and metadata. Data patterns, correlations, and linkages among these data formats can provide deep insights (Luo, 2008). Data mining may be defined as the process of determining correlations in large databases by applying multiple levels of analysis. It is a high-potential and robust tool that assists organizations or

businesses in increasing sales and profit and optimizing the manufacturing process from the available data (Agarwal, 2014).

According to (Luo, 2008), it is a knowledge discovery method in which data is analysed from multiple dimensions and summarized into meaningful information, that can be used to increase revenue, enhance production, optimize resource utilization and cost. Data mining is a rapidly growing field of study that is the intersection of multiple disciplines: statistics, databases, AI, data visualization, and CC. Data mining provides valuable information that standard queries or reports cannot deliver efficiently.

Data mining is an essential component of the Knowledge discovery in database (KDD) process (Luo, 2008). Figure 11 shows the KDD methods for the process industry, and Figure 12 shows the complete overview of knowledge discovery in a process industry.

Figure 11. Knowledge discovery methods in the process industry (Adapted from Zhang et al., 2018)

KDD process as in Figure 12, including data integration, data pre-processing, data warehouse, data selection, and data transformation are used to handle and prepare data.

Whereas data mining, data analytics, and pattern evaluation are used to extract deep insights from the processed data (Agarwal, 2014; Luo, 2008). Data mining techniques can be generally categorized into descriptive and predictive. Descriptive data mining methods (association, generalization, clustering, and sequence pattern mining) are applied when a dataset's properties and structure have to be explored. Predictive data mining methods (classification, prediction, exception knowledge mining, and time series analysis) are applied to forecast or predict the trends in the data (Cheng et al., 2018).

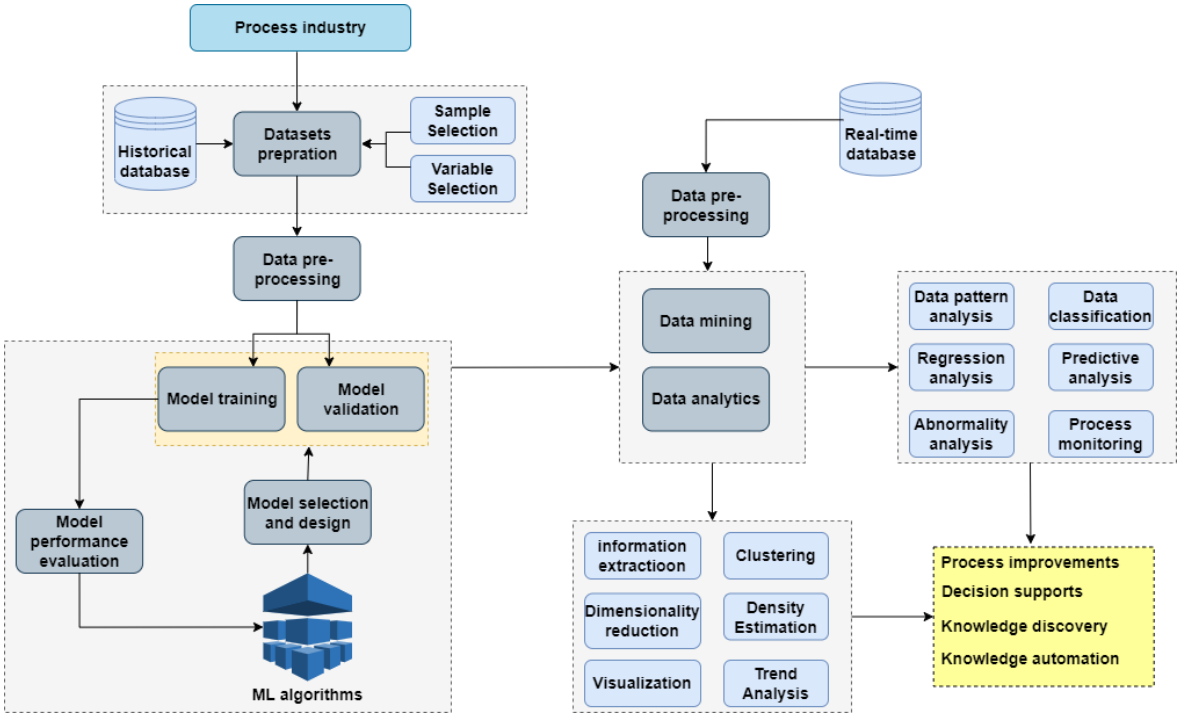


Figure 12. Process industry KDD (Adapted from Ge et al., 2017)

By incorporating data mining insights into process monitoring systems (PMS) and decision support systems (DSS), process optimization can be achieved through closed-loop feedback gathering system and early rectification of manufacturing process faults. Data mining and analysis techniques hold an essential relationship between operational and information technologies in order to form advanced process control systems (APC) that allow cognitive and self-learning abilities to handle real-time data in decision-making systems (Belhadi et al., 2019).

Data mining tasks are diverse and distinct because multiple hidden patterns exist in extensive datasets. Different data mining techniques are required to extract distinct patterns from a dataset. Data association, data summarization, data clustering, data trend analysis, and data classification are the techniques that can be classified according to the patterns or deep insights found in a dataset. These techniques are inherited from multiple research areas: ML, statistics, probability, and association rules (Luo, 2008).

2.3.1 Data exploration

It will not be easy to understand, analyse and apply information if the information extracted from data mining techniques cannot be presented in a simple and straightforward way. Existing data mining approaches do not provide visual analysis, exploration, and optimization of data sets. Data visualization is an important phase of data analysis since it helps in analysing outliers, selecting features, and tweaking the ML processes.

Data can be visualized in several ways, such as heatmaps, histograms, or two-dimensional scatter plots that summarize several complicated correlations into simple, instructive plots. For example, using histograms and violin plots during Exploratory data analysis (EDA) helps identify outliers and prune the data (Komorowski et al., 2016; Ward et al., 2018). EDA minimizes presumption and assists in applying optimal models for further exploration. In simple terms, it is utilized to visualize and extract important but less evident information from data. EDA applies a number of techniques, including descriptive statistics and visualization methods, for in-depth exploration of a dataset.

Data patterns should be thoroughly analysed to fully understand the dominant behaviour and unique characteristics in the data. It helps in determining advanced statistical methods that should be applied to data to extract meaningful information. It also helps in determining the scale on which data is originally represented is applicable or not (Camizuli and Carranza, 2018). (Komorowski et al., 2016) presented a few exploratory data analysis techniques according to the data types (Table 1) and objectives of data analysis (Table 2).

Table 1. EDA methods according to data type (Adapted from Komorowski et al., 2016)

Data Type	EDA method
Categorical	Descriptive statistics (mean, median, mode, SD, Variance)
Univariate continuous	Scatter plots, Line plots, Histograms
Bivariate continuous	XY scatter plots
Two dimensional arrays	Heatmaps
Multi-variate	3D scatter plots
Multi groups	Box plots side-by-side

Table 2. EDA methods according to objective (Adapted from Komorowski et al., 2016)

Objective	EDA method
Distribution of a feature	Histograms, Kernel density estimation
Outlier identification	Scatter plots, Histograms, Box plot
Relationship between two features	Scatter plot with covariance curve and correlation
Relationship between two input features and one outcome feature	Heatmaps
Multi-dimensional data visualizations	PCA or t-SNE

When graphically examining the data distribution, it is apparent that the distributions of the features might seem quite different. Instead of graphically analysing range of data distributions, it may be preferable to characterize the data distribution using statistical parameters to describe the data in terms of central tendency (minimum value, maximum value, arithmetic mean, geometric mean, median, mode), data spread (minimum value, maximum value, interquartile range, variance, standard deviation), and data distribution (kurtosis, skewness) (Komorowski et al., 2016; Reimann et al., 2008).

2.3.2 Data pre-processing and cleaning

Data pre-processing is a collection of techniques used to eliminate inconsistencies and noise from data in order to improve data mining performance. Data integration, cleansing, selection, and transformation are some of the essential data pre-processing methods. In data analysis, data pre-processing techniques are applied to structure the data because unstructured data having inconsistencies (outliers and missing) cannot be utilized directly for data mining (Agarwal, 2014). Figure 13 presents the categorization of data pre-processing methods into different subcategories.

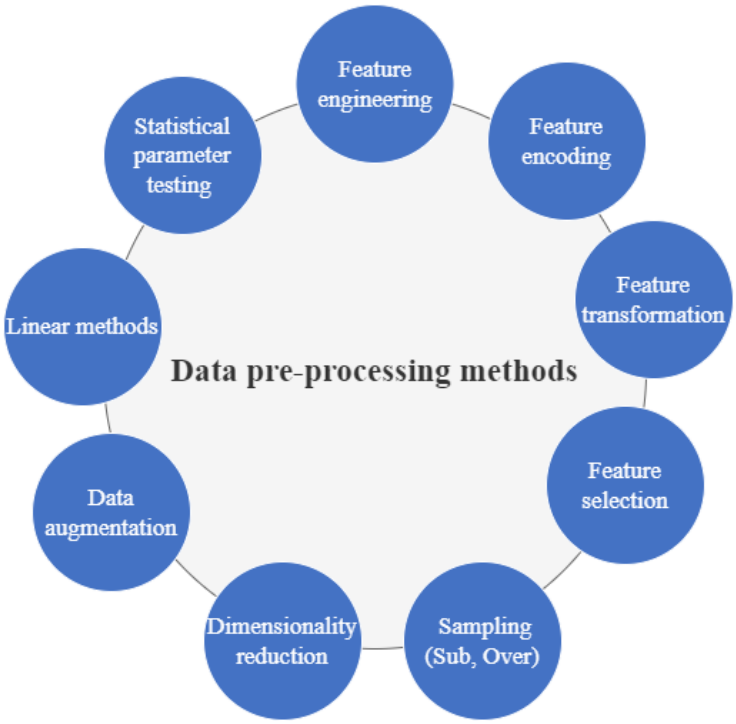


Figure 13. Categorization of data pre-processing methods (Adapted from Nguyen et al., 2019).

Feature extraction is a technique for developing new features dependent on the original input features to decrease the high dimensionality of the feature vector without losing valuable data (Zebari et al., 2020). The process of integrating data from numerous resources and loading it in a data warehouse is known as data integration. Databases, data cubes, and text files can all be used as data sources. Data heterogeneity and redundancy is the most significant challenge of data integration.

Data transformation is the process of converting raw data into a uniform format required for data mining (Agarwal, 2014; Sun et al., 2018). The data transformation technique filters and summarizes data according to the objectives of data mining. Directional, planned data aggregation can make data analysis more efficient. Data transformation consists of data aggregation, data smoothing (noise removal), and data normalization.

Data normalization is required to scale down the data into smaller range like [0,1], which is helpful in avoiding the data properties being dependent on the measurement units that can influence the model results. Min-max, zero mean, and fractional scale are the most used

normalization methods to normalize the data. Normalization approaches perform better on NN or classification algorithms based on distance measure (Sun et al., 2018).

Feature engineering is described as the *"Process of using domain knowledge of the data to create features that make machine learning algorithms work more efficiently"* (Zhang et al., 2018). (Agarwal, 2014) states feature engineering is a process of new attribute constructions based on existing attributes. For example, the area attribute can be created if height and width are available in the data.

Data reduction techniques are applied to massive datasets having multiple features to get a derived version of the data while preserving the veracity of the original data. Data reduction techniques include aggregation of the data cube, selecting the subset of attributes, dimensionality and numerosity reduction, and concept hierarchy. Raw data inputted from different or identical sources may contain many inconsistencies in measuring units or recorded values, so data cleaning techniques must perform on such inconsistent data for accurate results (Agarwal, 2014; Xu et al., 2015; Zhang et al., 2018). Generally, data cleaning is based on four steps, as shown in Figure 14.

Figure 14. Data cleaning steps (Adapted from Xu et al., 2015)

The simplest method for handling missing values is to ignore the data row; however, this is not a preferable option unless the missing ratio of the data is relatively large. One way to calculate the missing value is by calculating the mean value of the available data in a feature and replacing the missing value with it. Regression is a highly effective technique that can be used to calculate the missing value; it is a statistical approach that can be used to fill in missing data by statistically estimating the missing value (Agarwal, 2014; Xu et al., 2015). (Xu et al., 2015) presented a detailed overview and selection of missing value imputation methods: mean replacement, regression replacement, hot deck replacement, MLP, fuzzy similarity-based, pairwise, and likewise deletion, EM, matrix factorization, decision trees, random forests, and maximum likelihood.

Outliers are observations that do not exhibit constant statistical relation with the majority of the data set. Before applying data mining methods, outliers should be removed from the data as these can influence the model parameter estimation results and data analysis process. In the manufacturing industry processes, outliers can arise due to multiple reasons, including sensor failure and improper processing of missing data. Outliers can be of univariate or multivariate. Univariate outliers appear in the context of a single feature, whereas multivariate appears when a combination of variables violates a specific threshold (Xu et al., 2015). (Xu et al., 2015) also presented a significant overview and assessment of outlier detection techniques, shown in Figure 15.

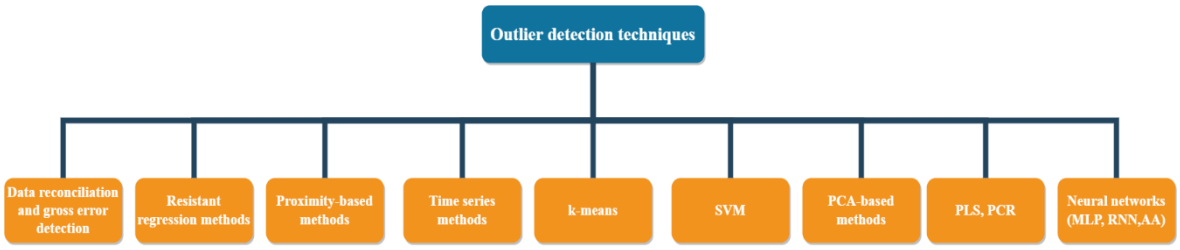


Figure 15. Outlier detection methods (Adapted from Xu et al., 2015)

Outliers are not always bad; sometimes, outliers provide helpful information that can lead to the unearthing of new information (Cheng et al., 2018; Xu et al., 2015). Sometimes outliers are difficult to identify because they involve an expansion of unseen regions. Detecting outliers become more difficult because noise distorts common data value, making it

impossible to distinguish between average data value and outliers. (Agarwal, 2014) presents a five steps process for outlier analysis: (i) Input data for data cleaning (ii) Applying algorithms to distinguish the outliers from the common values (iii) Presentation of an outlier (iv) Describing the profile of the outlier (v) Exploring outliers.

2.3.3 Dimensionality reduction and de-noising

Data analysis has become much more complicated due to recent trends in gathering large and diverse datasets. One of the characteristics of such large datasets is that they contain significant levels of redundancy. The use of massive multidimensional data brings increased noise, redundancy, and the likelihood of unrelated data entities (Houari et al., 2016). The process of transforming a multidimensional data representation into a low dimensional representation is known as dimensionality reduction.

The use of multiple dimensionality reduction techniques has become widespread in many areas of application due to the massive growth of multidimensional data. Furthermore, a number of new techniques are constantly emerging. Dimensionality reduction techniques take a multidimensional dataset and transform it into a low-dimensional dataset while preserving as much of the data's original meaning. The low-dimensional approximation of the original data helps in resolving the curse of the dimensionality problem. Data in low dimensions is simple to examine, process, visualize and interpret (Zebari et al., 2020).

In scientific studies and industrial production, multidimensional data is ubiquitous. It provides ample information but also poses significant challenges to data mining and pattern recognition methods due to its sparseness and redundancy. Dimensionality reduction is essential in pattern recognition methods to reduce the noise and redundancy in the data and to improve the efficiency and classification accuracy of the learning algorithm (Velliangiri et al., 2019). Several benefits can be achieved by applying dimensionality reduction techniques to multidimensional data. (i) Storage space required by multidimensional data can be optimized and reduced when the data is transformed into lower dimensions, (ii) Processing time can be reduced, (iii) Data redundancy, irrelevancy, or noise can be eliminated, (iv) Data quality can be enhanced, (v) Certain algorithms do not perform well when the data is in multi-dimensions, so reducing data dimensions helps algorithms perform efficiently and improve accuracy results, (vi) Exploratory data analysis of multidimensional

data is complicated, so reducing data dimensions may help in clearly developing and analysing patterns, (vii) It simplifies the classification process and enhances efficiency.

Dimensionality reduction methods can be classified into feature selection and feature extraction approaches. Feature selection is considered the most crucial technique; some major multidimensional challenges can be handled by applying it, such as eliminating unnecessary data, efficiently reducing duplication, decreasing redundancy in the data and increasing result interpretation. Whereas feature extraction deals with the challenge of identifying the most unique, explainable, and limited group of features to improve the efficiency of data processing and storage. Information can be missed during feature selection because some features should be removed throughout the feature subset selection process, but in feature extraction, the dimension may be reduced without losing much of the original feature dataset (Zebari et al., 2020). Feature extraction and feature selection are crucial parts of feature engineering for dimensionality reduction. In contrast to feature selection, feature extraction attempts to create a new feature subspace by projecting the original feature space with specified criteria.

The well-known feature extraction techniques are PCA and PLS, other approaches are also available, such as linear discriminant analysis, kernel principal component analysis (KPCA), and Sparse principal component analysis (SPCA) that can be implemented in the same way as PCA (Xu et al., 2015; Zhang et al., 2018). *"PCA uses orthogonal transformations to transform a set of features into linear uncorrelated features called principal components."* Although the method has a strong theoretical foundation and is used to reduce linear dimensionality, it is not suitable for the dimensionality reduction of a non-linear data.

KPCA and SPCA can effectively perform dimensionality reduction in non-linear data. Sparse data refers to features with a large number of missing values. Although multidimensional data comprises an enormous number of features, only some of them are linked with specific learning models, so it can train the model by reconstructing a lower-dimensional dataset of features. Sparse models can filter out a significant amount of redundancy and noise from data, leaving just the features that are relevant to the objective. The fundamental principle of the spares model for feature selection is the multi-purpose optimization of a problem (Velliangiri et al., 2019).

(Houari et al., 2016) state that dimensionality reduction techniques are mainly divided into two types: Linear dimensionality reduction (PCA, SPCA, SVD, LDA, FA) and non-linear dimensionality reduction (LLE, ISO MAP, KPCA, LE Maps, MDS). Most of the data in manufacturing industries originate from sensors that generate signals via electrical, electromechanical, or electro-optical methods. The environment is likely to affect sensor data, i.e., high-frequency noise can affect signals. Usually, noise in the data must be examined thoroughly since it can indicate changes in the operational environment. For this, an interdisciplinary approach known as "acoustic chemometrics" can be used to manipulate the noise to uncover information about the process (Xu et al., 2015).

Noise refers to an abnormal attribute value that differs from the rest of the data in a feature. For example, after drying moisture content in veneer sheet is 80% (the normal range is between 8% and 15%), the body temperature of a patient having 26.9 centigrade, and a pH of 2.26 (the normal range is between 4 and 8). Noisy data can be processed by regression, binning, outlier analysis, and data extraction from other data sources (Sun et al., 2018). Multiple approaches are available to filter noise from the data and can be classified into model-based and data-driven approaches. Kalman filter is one of the most widely used model-based filters for filtering noise from the data. Whereas digital filters, Savitzky-Golay filters, and wavelet filters are used as data-driven filters.

2.3.4 Reliability and validity

Due to uncertainty in the manufacturing process, dynamic changes in real-time data often cause volatility in the models and even declare the earlier obtained model invalid. Meanwhile, accumulating unseen data may cause previously established information to be invalidated. Data mining algorithms may uncover hundreds of patterns, some of which are incorrect or irrelevant in a specific perspective. However, there is currently no definite system or procedure for assessing such issues. Furthermore, data mining uses ML methods to extract meaningful insights, the model may be unable to interactively identify the obtained knowledge, not resulting in completely adaptable or helpful knowledge. Some data mining approaches, such as Gaussian regression, may analyse the quality of the mining and the degree of uncertainty in the outcomes (Cheng et al., 2018).

(Raschka, 2018) summarizes the main reason to assess the predictive performance of a model: (i) To evaluate the model's generalized performance, which is how the model will behave when given unseen or new data. (ii) To improve prediction performance by fine-tuning the learning algorithm and selecting the optimal model with high performance accuracy. (iii) To find a ML model that is most suited for the current problem, by applying and testing different algorithms and selecting the model with high prediction accuracy.

A reliable method for testing a model's performance involves training an ML model with existing data and evaluating its classification performance using newly acquired data or unseen data. Train/Test split is a reliable method to validate the model's performance in which a train set is used for the training, and a test set is applied for the validation of the model. Utilizing unseen data to test a machine learning model provides an impartial assessment of how well the model will perform when the model is used to make real-world predictions.

There are cases when the available datasets are limited; in that case, cross-validation is a useful approach to validate the model. Instead of training a model only once, multiple models are developed iteratively on different data segments. K-Fold, leave-one-out, and hold-out are the most popular cross-validation techniques (Vabalasid et al., 2019). Results of the machine learning classification model can be evaluated on the values of sensitivity, specificity, and area under the curve (AUC) (Adams, 2017).

2.4 Data Management

The volume of big data is growing exponentially, and the current capacity to work with big data is relatively low levels of petabytes, exabytes and zettabytes. Along with the benefits of data analytics, big data also brings many challenges, including difficulties with data acquisition, storage, analysis, and visualization (Philip Chen and Zhang, 2014). In order to process vast volumes of data from multiple sources, BDA and data mining require an intelligent architecture that should be based on data storage techniques, data governance, data management and data risk management (Belhadi et al., 2019).

2.4.1 Data Conversion

Data transformation into a suitable form to perform the analysis is a barrier to adopt big data analytics. Big data can be turned into an analytical process in two ways; the first is for unstructured data, and the second is for structured data, as shown in the Figure 16. If the data is in a structured format, then data is pre-processed before being saved in relational databases to meet the required constraints of schema related to the structure of the data, and after that, the data can be accessed for analysis. On the other hand, Unstructured data must be kept in distributed databases before being accessed for analysis. Unstructured data can be accessed from distributed databases after meeting the schema-on-read constraint (Hashem et al., 2015).

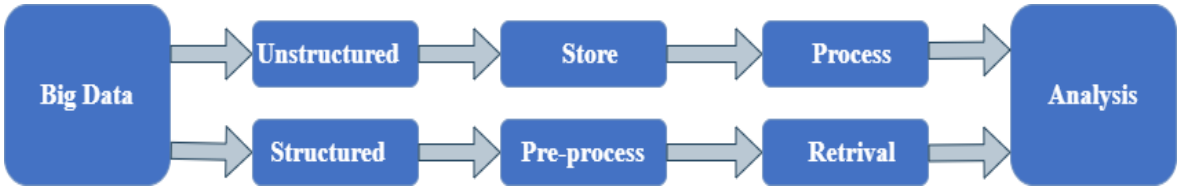


Figure 16. Transformation of Big data for data analysis (Adopted from Hashem et al., 2015)

Data integration process is the most time and resource-requiring process in business intelligence. It is estimated that approximately 70% of the time and effort is spent on the ETL process. ETL stands for extract, transform and load, it is a technique of incorporating data from multiple data sources into one consistent data store (Wikipedia). Usually, data conversion is straightforward and can be done either as an ETL transform or as part of the source to target mapping, as shown in the Figure 17. The data conversion is needed when the source and target data types do not have the same structure; different data sources with different data types may be used to populate a table or a column; a source column contains a specific code that is based on some combination of other values and that need to break down into separate components etc. (Sherman, 2015).

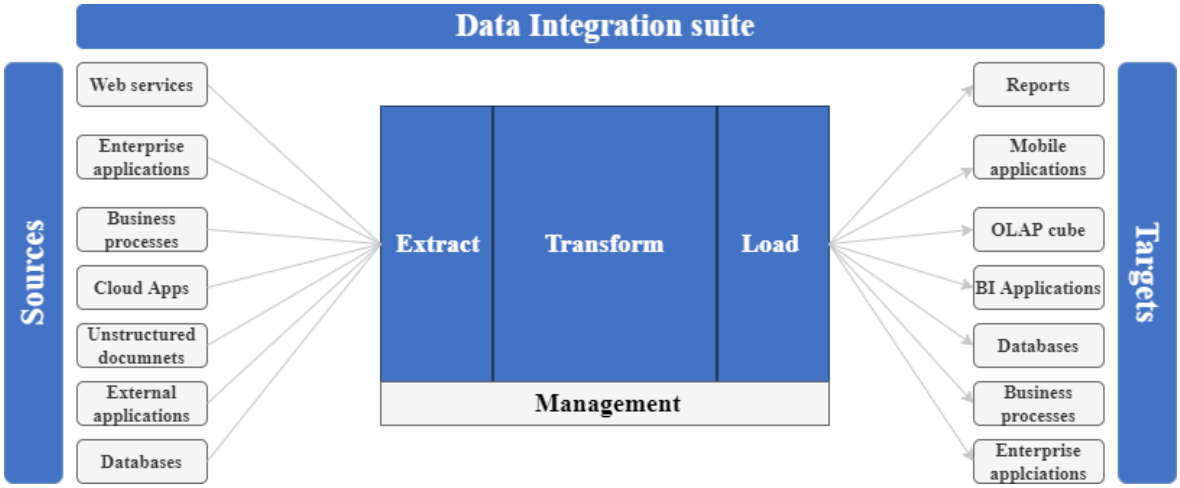


Figure 17. Data conversion from source to target (Adapted from Sherman, 2015)

2.5 Machine learning methods

After data pre-processing, applying a machine learning algorithm on the data might be useful to solve a specific problem. There are a number of different machine learning algorithms, and when choosing a machine learning algorithm, the most common question is *"which algorithm to select to solve the given problem."* The selection of a machine learning algorithm is based on multiple factors such as (i) data size, data quality, and the problem domain, (ii) computational time, (iii) urgency and importance of the task, and (iv) loss function to be minimized (Nguyen et al., 2019).

Machine learning is a subset of artificial intelligence techniques which allows computers to learn from prior information using historical data and enhance their performance in solving a specific task (Ge et al., 2017; Nguyen et al., 2019). However, probability and statistics theories also play an essential part in modern machine learning algorithms. ML techniques include regression, classification, clustering, decision trees, SVM, NN, decision trees, Bayes learning, etc. Machine learning algorithms are based on four types supervised, unsupervised, semi-supervised, and reinforcement learning. The first three types are typically used for data mining and analytics, whereas reinforcement learning is mainly used in developing robots, games, and navigation applications (Ge et al., 2017).

Unsupervised learning is a technique for extracting information from training data without a ground-truth label (Ge et al., 2017; Nguyen et al., 2019). Its primary goal is to study data and reveal hidden information. Unsupervised learning methods (K-means, PCA, Clustering etc.) and applications (Feature extraction, Dimensionality reduction, Process monitoring etc.) of unsupervised machine learning are listed in Figure 18 (Ge et al., 2017).

Supervised learning techniques are utilized when the samples in the data consist of a set (X_i, y_i) where X_i is the input value to be given to the predictor and y_i is the label (Ge et al., 2017; Nguyen et al., 2019). In supervised learning, labelled data samples can be discrete or continuous. When the labels are of discrete data type, supervised learning can be used to classify the process data, such as operation mode classification, quality classification, or fault classification. If the labels are continuous, regression models can be implemented to predict and estimate the label. Methods (PCR, SVM, Random Forest etc.) and applications (Feature extraction, Dimensionality reduction, Process monitoring etc.) of unsupervised machine learning are listed in Figure 18.

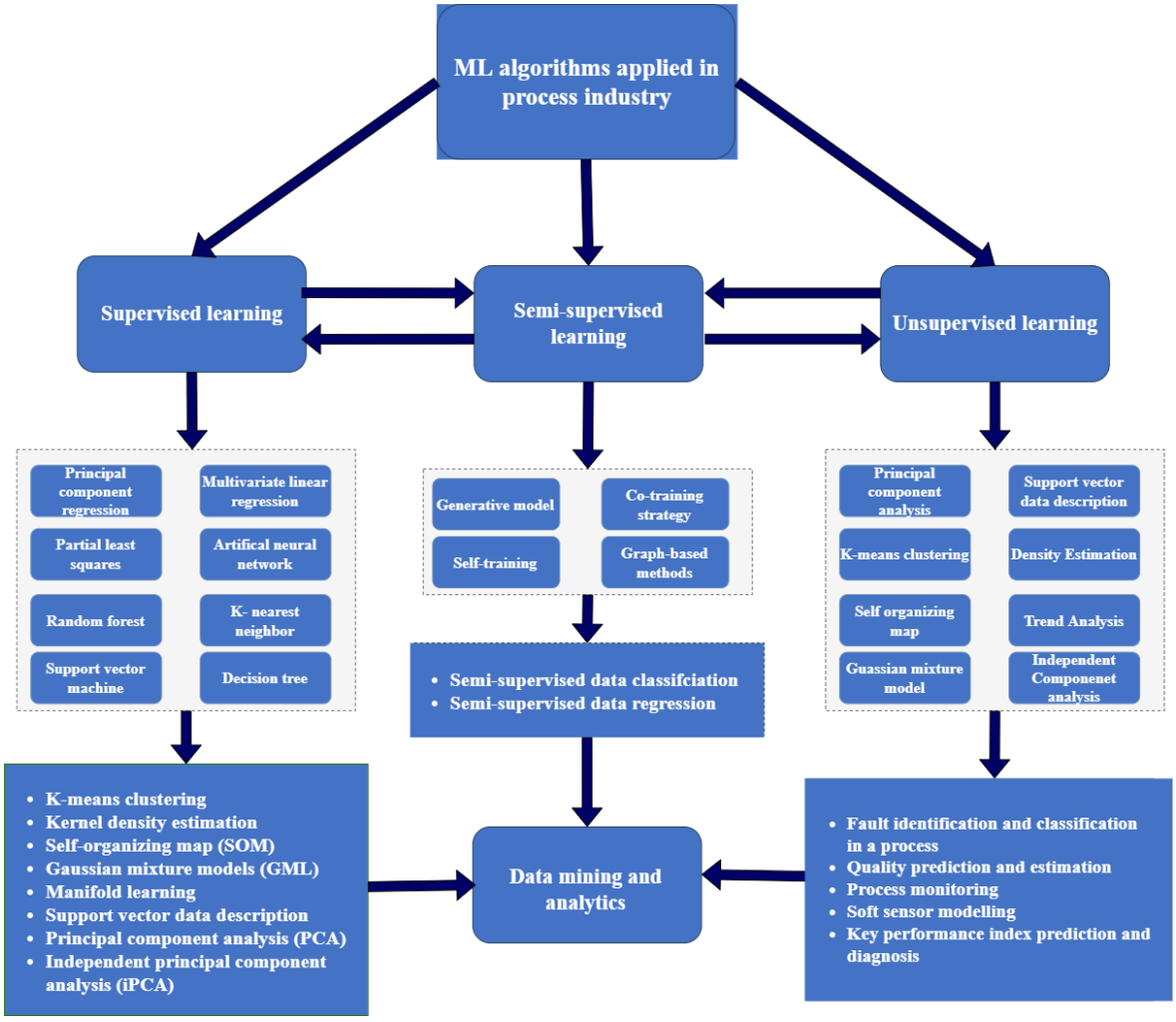


Figure 18. ML algorithm types and usage in the process industry (Adapted from Ge et al., 2017)

Semi-supervised learning techniques are effective in the process industry applications, where the labelling of data is costly and time-consuming. For example, labelling the form of discovered defects in data is a complicated task that may need engineers' process expertise and experience, which can be costly and time-consuming. A subset of machine learning called reinforcement learning is used to design algorithms that decide which actions to perform in a given situation in order to maximize reward. In contrast to supervised learning, where the predictor is assigned by data labels, the model in reinforcement learning learns by trial and error to determine the optimum output (Ge et al., 2017).

Recently, research in machine learning has been focused on deep learning, which is a method of learning multiple layers of abstractions to interpret the given data more accurately(Gubbi et al., 2013). Figure 19 shows the commonly used deep learning architectures.

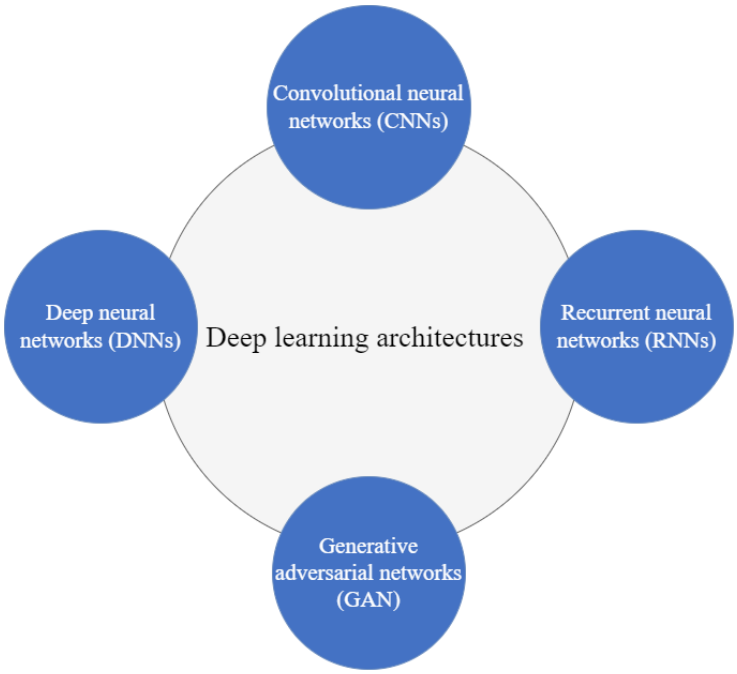


Figure 19. Deep learning architectures (Adapted from Nguyen et al., 2019)

3 Literature review

This chapter consists of the systematic literature review on I4.0, focusing on veneer/LVL industry. The first sub-chapter describes the followed review methodology for the literature review and the second sub-chapter focuses on keyword search on scientific databases to find out machine learning methodologies applied in optimization of veneer/LVL industry.

The aim of the literature review is to, first, evaluate and compile the current research that has been done related to the peeling and drying of veneer/LVL industry for smart manufacturing. Secondly, which machine learning methods has been applied in existing research papers to optimize the peeling and drying processes?

3.1 Review methodology

To ensure the literature review's robustness and reliability, the study follows the concept-centric structured approach recommended by (Webster and Watson, 2002) to determine the source material for the review. To conduct the current literature review in an organized manner, a series of processes is followed: Preparation, Search, Selection, and Analysis process. Figure 20 shows the research processes, sub-processes, and its outcomes.

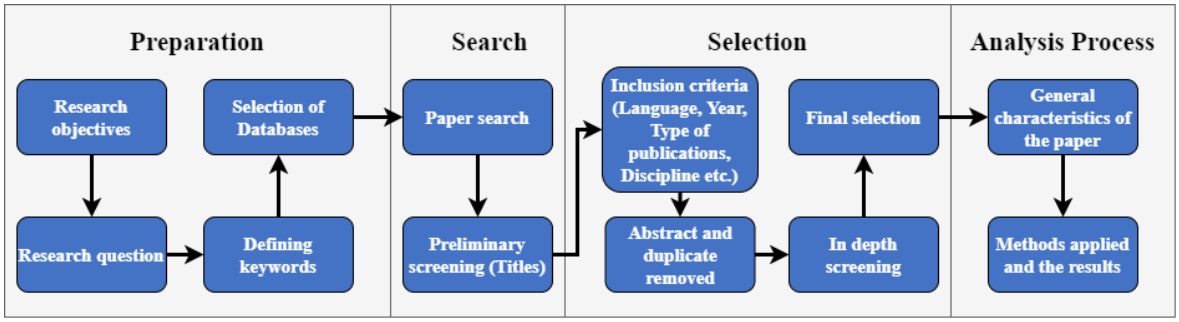


Figure 20. Research processes, sub-processes, and its outcomes

3.1.1 Selection of research databases

After evaluating the relevant data sources, the search methodology is developed to access the wide range of existing research related to the research objectives and question. To begin the research related to the literature review, five scientific research publication platforms have been identified as in Table 3 and a search engine (Google scholar) to extract the papers related to the research objectives. These databases provided complete and organized access to the scientific publications from journals, along with the combination of search keywords helpful in performing a systematic literature review.

Google Scholar is not a database; instead, it is a search engine to search the papers, which does not allow detailed information and characteristics about the research. By using Harzing's Publish or Perish tool, Google scholar is used as a database, which was

significantly helpful in searching the papers according to the search criteria by applying specific filters.

Table 3. Research publication platforms

Research Publication Platform	Database
EBSCO	Academic Search Elite
Elsevier	Scopus
IEEE	IEEE Xplore Digital Library
ProQuest	ProQuest Technology Collection
Springer	Springer Link

3.1.2 Keyword selection

For the literature review, it is important to limit the research in the direction of research objectives. For this, relevant and applicable research journals related to the research topic are selected to build a reproducible, thorough, and impartial literature review process. At the first stage, *"Veneer peeling" AND/OR "LVL peeling", "Veneer drying" AND/OR "LVL drying", "Industry 4.0", "Machine Learning" OR "Optimization", "Wood Defects", "Image processing", "Feature extraction", "GLCM" and "CBIR" AND/OR "Similarity measure"* are chosen as the keywords which are the most focused, common, and representative terms in the literature review related to the research question to search published papers collected primarily from LUT Primo and Secondarily from Google Scholar. These selected keywords were used as pairwise queries in the databases listed in Table 4 with the search criteria in article, abstract and keyword.

Table 4 shows the bibliographical analysis on veneer drying and peeling keywords in general from the selected databases. Only articles originated from journals were consider in this research and the language for the article is English and there is no range applied on the number of years. In five databases term veneer drying produced total 5624 articles having largest number of articles in springer database and lowest in IEEE Xplore whereas for term veneer peeling total number of articles were produced are 935 having largest number of articles in springer and least in IEEE Xplore similar to the term veneer drying. Mainly in

general springer database presented high number of articles for both keywords. Figure 21 and Figure 22 shows the top six authors having research publications in veneer peeling and veneer drying.

Table 4. Numbers of papers related to the keyword search in each database.

Keyword	EBSCO	Scopus	IEEE Xplore	ProQuest	Springer	Total
Veneer drying	140	164	61	2102	3157	5624
Veneer peeling	51	81	11	319	473	935

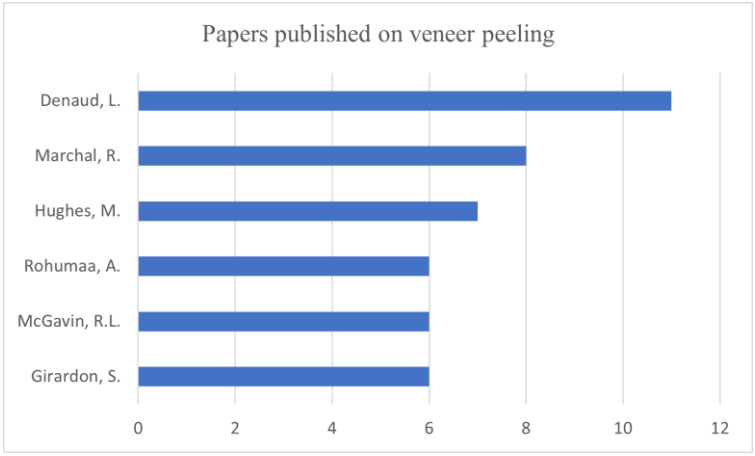


Figure 21. Papers published by authors on veneer peeling (Data source: Scopus)

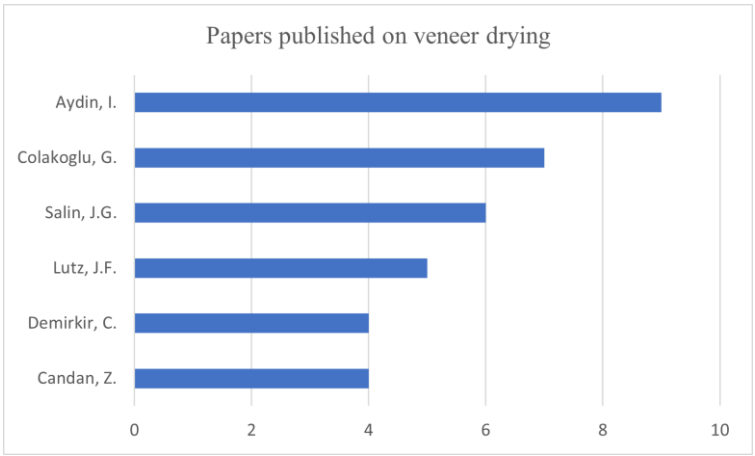


Figure 22. Papers published by authors on veneer drying (Data source: Scopus)

To limit the research in the right direction pairwise queries are used to search the relevant literature. For this along with the veneer drying and veneer peeling other keywords are also used to create combinations of keywords and filters are applied as in Table 5.

Table 5. Combination of search queries

Pairwise query	**Database**
(TITLE-ABS-KEY (veneer) AND TITLE-ABS-KEY (peeling)) AND (LIMIT TO (LANGUAGE , “English”)) AND (LIMIT-TO (SRCTYPE , “j”))	Scopus
(TITLE-ABS-KEY (wood AND defects) AND TITLE-ABS-KEY (image AND processing)) AND (LIMIT-TO (LANGUAGE , “English”)) AND (LIMIT-TO (SRCTYPE , “j”))	Scopus
(TITLE-ABS-KEY (veneer) AND TITLE-ABS-KEY (drying AND temperature)) AND (LIMIT-TO (LANGUAGE , “English”)) AND (LIMIT-TO (SRCTYPE , “j”)) Filters: Moisture and Plywood	Scopus
(TITLE-ABS-KEY (veneer AND drying) AND TITLE-ABS-KEY (lvl AND drying)) AND (LIMIT-TO (LANGUAGE , "English")) AND (LIMIT-TO (SRCTYPE , "j"))	Scopus
(TITLE-ABS-KEY (veneer) OR TITLE-ABS-KEY (lvl) AND TITLE-ABS-KEY (feature AND extraction)) AND (LIMIT-TO (LANGUAGE , “English”)) AND (LIMIT-TO (SRCTYPE , “j”))	Scopus
(TITLE-ABS-KEY (cbir) AND TITLE-ABS-KEY (similarity AND measure))	Scopus
(TITLE-ABS-KEY (glcm) AND TITLE-ABS-KEY (texture))	Scopus
Veneer AND Drying Filters: Manufacturing, English	ProQuest Technology Collection
(Image processing) AND veneer OR LVL Filters: Scholarly journals, English, Journal of wood science	ProQuest Technology Collection
veneer AND drying AND process AND optimization Filters: Manufacturing, English	Springer Link
veneer AND peeling AND drying	EBSCO - Academic Search Elite

3.2 Results of the literature review

Papers has been filtered out on the basis of inclusion criteria (keywords, search boundaries, and language) and exclusion criteria (similar titles, abstract, and relevancy towards research objective). Following papers have been short listed to analysed further.

(Kamal et al., 2017) implemented a feed-forward back-propagation neural network (BPNN) to identify the wood knot defects (leaf, dry, sound, horn, and edge). The gray level co-occurrence matrix (GLCM) and laws' texture energy measures (LTEM) are applied to extract texture features from the images. GLCM is one of the popular approaches applied in textural feature analysis for image classification invented in 1973 by (Haralick et al., 1973). The inspiration behind using GLCM by (Kamal et al., 2017) is to gauge the results of LTEM against GLCM. The technique is applied using distinct features: Contrast, Correlation, Energy and Smoothness for a BPNN applied as a classifier. University of Oulu dataset for the wood knot defects has been used, which contains 395 samples of two feature sets first set is for the GLCM-based features, and the second one is for LTEM-based features. Most of the samples were related to dry knot, sound knot, and edge knot defects. Dataset is divided for both feature sets (GLCM, LTEM) into 70% to train the neural network, 15% for testing and 15% for validation. For GLCM based feature set, BPNN's best performance results are obtained by 15 hidden layers of neurons, with the 0.1072 MSE and 84.3% accuracy. For LTEM based feature set, BPNN's best performance results are obtained by 30 hidden layers of neurons, with the 0.0718 MSE and 90.4% accuracy. Furthermore, it has been observed that by combining both GLCM and LTEM features, classification accuracy can be improved.

(Haryanto et al., 2020) presented a method for texture feature extraction using a multi-patch image pixel approach with sliding windows to reduce computational time for features extraction as high-resolution images take longer due to the enormous volume of data available in the images. A mean shift filter removes noise from the images before calculating the GLCM texture features. In this study, texture features for histopathology images (used for diagnosing cancer) are calculated using GLCM contrast, energy, ASM, dissimilarity, correlation, and homogeneity with the pixel distance d=5 and angles = (0°, 45°, 90°, 135°). The presented GLCM method is then trained using DNN with ReLu as an activation function during the training process and compared to other classification methods for performance

benchmarking. This research has observed that DNN with GLCM has higher accuracy of 96.72% with four CVs compared to the other classifiers.

(Alsmadi, 2020) presented a content-based image retrieval technique for similar candidate images from a database. The proposed system is based on GLCM for extracting texture features, Canny edge histogram for extracting colour features, Canny edge method for extracting shape features, YCbCr with discrete wavelet transform and neutrosophic clustering method. YCbCr is a colour space family used in video and digital photography systems as part of the colour image pipeline where Y represents the luma component, and CB and CR define blue and red difference chroma components (Wikipedia). A metaheuristic algorithm is used to select the image with the highest similarity to the queried image. Squared Euclidean distance is used for calculating the similarities between the query and the list of images in the database. During the image pre-processing different types of noises (salt, speckle, and pepper) have been removed by applying the median filter. The image dataset is based on thousand different types of images from the Corel dataset in which each image has a resolution of 384x256. Before applying GLCM on the image set, each image is converted to a grayscale image and a 5x5 Gaussian filter is applied to the image. The filtered image is then sliced into 4x4 blocks, and GLCM texture features energy, mean value, standard deviation, contrast, and homogeneity are calculated for four angles θ = (0°,45°,90°,135°). Results show that similarity measure calculated through meta-heuristics has proven an image's solid capacity for discriminating texture, shape, and colour.

(Wang and Ren, 2014) proposed a recognition method based KPCA and generalized learning vector (GLVQ) for rotary kiln combustion working conditions, either complete or incomplete combustion, by extracting textural data of flame image using GLCM. All 14 GLCM texture features have been calculated for the images, and KPCA with Gaussian function as kernel function is used on texture features to reduce the input vector's dimensionality. The first five principal components explaining more than 85% of variance are used for the GLVQ model. The model input layer is based on five neurons as five PCs are used as input, and the output layer is based on two neurons representing the condition of the kiln combustion. The model gives 95.83% accuracy with the hyperparameters 5000 iterations, 10^{-3} as a learning error and the learning step to 0.5.

(Surabaya et al., n.d.) proposed a method for auto-colouring the grayscale image by matching image block similarity based on GLCM texture features using the sum of absolute difference.

Blocks of the colour image are used as a template, and grayscale images are used as the target image. GLCM texture features (ASM, IDM, Contrast, entropy, and correlation) are calculated with distance L=1 and only one angle = 0°. Similarities of the two image blocks are calculated by subtracting the GLCM texture values and comparing the GLCM of the target with the template value to get the smallest value in the colour block. Block having the smallest colour image block is a pair of grayscale blocks due to similarity in the texture features. Lastly, transfer colour to the grayscale image until all areas having similarities are coloured.

(Jalonen et al., 2021) applied a visual method for tracking manufacturing products and presented its usage in plywood manufacturing. For the visual and positional transformations of the veneer, the authors have applied Siamese neural network. For the model training, validation and testing publicly available dataset (Veneer21) provided by the Raute corporation is used. Dataset is based on 2579 paired image samples of wet and dry wood veneer. Dimensions of the RGB images of wet veneers are roughly 3300 x 3300, and dry veneers are around 4100 x 4100. The images are resized to 224 x 224 and normalized to [0 1] pixel values. Image pairs are randomly divided into train, validation, and test sets. Training set contain 1879 image pairs, validation 200 pairs, and for testing 500 pairs. Model is trained with TensorFlow Keras using 200 epochs, cross-entropy is used as a loss function and Adam as an optimizer. The model has been trained sixteen times for training and with three batch sizes (16, 32, and 64), and four learning rates 10^{-3}, 10^{-4}, 10^{-5}, and 10^{-6}). Model's performance is evaluated by a matching pair of a wet veneer with five dry veneers out of which only one is the matching dry veneer pair and it has been repeated 1000 times for a set of 500 wet images resulting in a 500,000 clusters test. Hungarian and Greedy based decision rule approaches have been used to match dry and wet veneers. Hypermeters combination that gives the highest validation accuracy of 92% are the batch size of 64, learning rate of 10^{-5}, and validation loss of 0.200. Furthermore, the model has been tested for the real test scenario by creating a random cluster size 100 times with a cluster size of five dry and wet veneers. The average accuracy of the Hungarian decision rule is 98.54%, with a standard deviation of ± 0.71% and the greedy decision rule average accuracy is 91.41 ±1.66%.

(Urbonas et al., 2019) applied a method for the visual analysis for the positioning and classification of defects (Split, core, branch, and stain) on the veneer surface by applying a faster region-based convolutional neural network (faster R-CNN). Dataset is obtained by

scanning 250 veneer sheets of 1525 x 1525 mm into 300 x 300mm batches at a 4000 x 3000 pixels resolution in monochrome single-channel images for training and testing sets. A total of 353 (300 x 300mm) samples there is at least one defect present in 285 samples, and six samples had no defects, categorized as background. Out of 353, 291 samples are used for training, and 62 samples are used for testing. Testing samples contain the most defective veneer sheets. Furthermore, to improve the training of the neural network, the original dataset has been augmented (flip, resize, and rotation transformations), and the image size is reduced heuristically to 800 x 600 pixels to improve the training of the neural network. For the annotation and labelling of the defects, VoTT (Microsoft opensource application) is used to determine the defects' class, position, and size. The authors used transfer learning to achieve a better result by applying pre-trained neural networks BN-Inception, AlexNet, ResNet152 and VGG16. With the ResNet152 neural network, the model gives 80.6% of overall accuracy, and by combining all the defect classes into one type gives 96.1% of accuracy.

(Ahmed et al., 2020) stated that the temperature of the veneer sheet is a key component that could be utilized to assess the quality of the final product. In order to manage the temperature and humidity inside the sections of a drying machine, several process parameters must be controlled, including airflow, gas usage, drying speeds (time), dryer zone temperatures, and chain side temperature. Besides process parameters, product parameters such as wood species and veneer sheet thickness also influence the drying process. The authors applied the regression tree method to simplify the complexity between the process and product features and determine the essential elements for drying veneer and reaching the desired range of veneer temperature (output variable). The study shows that the temperature of the veneer is determined on the three parameters average temperature in zone 1 of the dryer (C1), average temperature in zone 3 of the dryer (C3) and mean daily temperature climatic variable (MDT).

(Demirkir et al., 2013) presented a method by applying ANN to predict the intermediate bonding strength values based on the peeling and drying temperature of the veneer sheets. The proposed approach has two objectives first is to predict optimum manufacturing parameters using ANN without compromising bonding strength and losing time. The second objective is to use the ANN model and values obtained from experimental research to determine the significant proportions of these parameters on plywood panel bonding strength. It has been found that the bonding strength of the plywood panels with phenol-

formaldehyde (PF) resin has a positive correlation with the drying temperature of the veneer, whereas plywood panels with melamine urea-formaldehyde (MUF) resin have a negative correlation. Optimum drying temperatures for higher bonding strength for PF and MUF have been calculated for different wood species veneer panels. The results obtained in this study can help to decide the optimum temperatures for drying and steam conditioning of wood logs according to the type of wood.

(Yuce et al., 2014) has presented ANN-based quality control to identify defects in wood veneer with lower assessment benchmarks and improve the response time during quality control. PCA has been applied for feature selection and to reduce the dimensionality of input variables for the ANN, which minimized the model training time and improved the method's accuracy. It has been observed that the model's performance has been improved by 56% with PCA compared to the training and testing of ANN with all features. To further improve the model's performance and benchmark the results of ANN with PCA, the Taguchi method has been applied, which shows that with only one hidden layer, a high number of neurons in the hidden layer, and fewer PCs, ANN performed well as compared to PCA.

Veneer drying process consumes a large amount of energy which is approximately fifty percent of the mill’s energy in the plywood manufacturing. (Han et al., 2015) presented an approach by using non-linear programming and operational research theory to design manufacturing conditions to optimize the energy consumption during veneer drying, the manufacturing conditions such as temperature, pressure, and the number of veneers are investigated as a function of energy consumption (Q), modulus of elasticity (MOE) and contact angle (CA).

(Çolak et al., 2007) investigated the log steaming and veneer drying conditions on technical properties and durability of LVL and solid sawn lumber. To evaluate the effects of treatment on different properties of the veneer, authors proposed a chemometric modelling and multivariate data analysis of near-infrared spectroscopy. As a result, PLS models were effective for the quality control of veneer treated thermally.

As a summary of the literature review, there are multiple ML methods that are used in optimizing the drying process, as in Figure 23. Secondly, GLCM seems to be one of the most used methods to extract features from the images. Most of the authors have used this method and applied different methodologies on extracted GLCM texture feature values in order to

match the images, which gives promising results. The most important papers covered in this literature review are summarized in Table 6.

Figure 23. Methods frequently used in optimizing the drying process according to current literature

Table 6. Papers according to the concept centric approach

Article	Drying	Peeling	Image processing / Feature extraction (Defects)	Approach
(Kamal et al., 2017)			x	GLCM, LTEM, BPNN
(Urbonas et al., 2019)			x	Data augmentation, transfer learning, R-CNN, Deep learning
(Han et al., 2015)	x			Non-linear programming and operational research theory
(Jalonen et al., 2021)			x	Siamese neural network, Hungarian and Greedy decision rules
(Demirkir et al., 2013)	x	x		Artificial neural network (ANN)

(Ahmed et al., 2020)	x			Regression tree, Data mining, ANOVA, Cross-validation
(Çolak et al., 2007)	x			Partial least square (PLS), Multivariate data analysis
(Yuce et al., 2014)			x	ANN, PCA, Taguchi, Feature selection
(Haryanto et al., 2020)			x	GLCM, DNN
(Alsmadi, 2020)			x	GLCM, Canny edge, YcbCr, Canny edge histogram, median filter, MA, SED
(Wang and Ren, 2014)			x	GLCM, KPCA, GLVQ,
(Surabaya et al., n.d.)			x	GLCM, SAD

4 Methodology and Data

This chapter describes the data and methodologies applied in the veneer/LVL manufacturing process. The production process section briefly explains the current manufacturing process from inputting of woodblock, peeling, quality evaluation, drying process, and final output of the dried veneer sheets. The second sub-chapter thoroughly explains the data and methodology applied. The complete process understanding from log selection to the final product was made possible with the learning package provided by the Raute Corporation.

4.1 Production process – Peeling and drying

Plywood and LVL is manufactured by gluing wood veneer sheets together. Plywood structure is based on cross-bonded structure, meaning in each layer, the wood grain directions of sheets vary by 90 degrees, making it strong in all directions. It can be produced in various dimensions according to the requirements, but the main limitation is the length of woodblock in the peeling process which limits the veneer length and final product length. Commonly veneer sheets are peeled from 4 ft or 8 ft long woodblocks, and the main plywood panel sizes are 4 ft x 8 ft and 8 ft x 4 ft.

Typically, plywood is composed of mainly two dimensions: long-grain plywood and short-grain plywood. In long-grain plywood, grain direction is in the larger dimensions, whereas in short-grain plywood, grain direction is in the smaller dimensions. Plywood panels can be composed of odd or even numbers of veneer sheets (piles). There are different types of veneer sheets on which the structure of plywood is based; these sheets are mainly top face veneer, long core veneer, cross core veneer and back face veneer.

LVL is made up of thick veneer sheets that are placed in the same grain directions with billet length. This type of LVL sheet is used in load-bearing structures such as roof rafters, curved roof components, interior wall posts, prefabricated floor units and concrete casting moulds. LVL sometimes has veneer sheets glued together similar to the plywood in the crosswise direction perpendicular to the panel length to get more stability in the width direction. The thickness of the LVL depends on the number of 3 mm veneer sheets glued together; usually, LVL thickness range from 21 to 90 mm. Maximum width ranges are 1200 mm, 1800 mm or 2500 mm, whereas the length of LVL can be up to 25 meters. LVL is dimensionally straight and a stable product due to its laminated structure and drying process. Like plywood, it has the best weight-strength ratio, which can compete with steel and concrete as building materials.

Production of plywood and LVL is based on multiple production stages. As in the Figure 24, these production stages are required to make logs into veneer sheets and from veneer sheets to plywood or LVL.

Figure 24. Overview of plywood and LVL manufacturing process

Wood is a renewable material, and it is ready for industrial harvesting when it has a required diameter of approx.150 mm. Different wood species are utilized to manufacture plywood and LVL, such as birch, pine species, eucalyptus, poplar, maritime pine, tropical hardwood, etc. The first step in the veneer peeling is selecting the raw material, which is a log. Specific

requirements should be fulfilled to select a log for further processing and to get a higher yield in the form of veneer ribbon.

After raw material selection and storage, logs have to be passed through the conditioning process. The purpose of the conditioning process is to heat the log to make it suitable for the peeling process. The logs are usually conditioned by soaking in a hot water pool and spraying water over the bundles for about 12 to 24 hours.

After log conditioning, the bark has to remove from the log, length of the log has to be measured to properly cut the log into the required length for the peeling process. Figure 25 shows the pre-processing of the log before the peeling process.

Figure 25. Pre-processing of the log before peeling

Logs are fed onto the log conveyor table and run through the metal detector, which removes any metal spots from the peeling block after this bark is removed from the log. A 3D scanner is used to create a 3D model of the logs to identify the log's exact volume for cutting. According to the scanner's inputs, a log is cut into the peeling block within tolerance limits. After cutting logs into specified peeling block lengths, the logs are inputted to the peeling line to be peeled into veneer ribbons and then cut into the desired length of veneer sheets, as shown in the Figure 26.

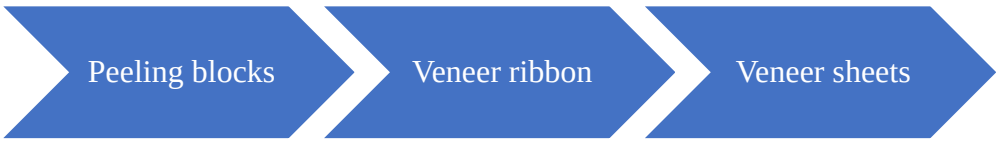


Figure 26. Overview of blocks to veneer sheet process

Veneer sheets manufacturing is based on multiple phases as in Figure 27. Peeled veneer ribbons get scanned to detect any defects in the veneer ribbon; if there is any bark or defect in the veneer ribbon, the scanner sends instructions to the clipping line to clip the veneer accordingly. The moisture of the veneer ribbon is also measured during the scanning of the defects on the veneer ribbon. The clipper then cuts the veneer ribbon into veneer sheets according to the instructions received from the scanner.

Figure 27. Overview of veneer sheets manufacturing

After peeling the blocks into veneer ribbon and clipping veneer ribbon into veneer sheets, the next process is the drying of the veneer sheets. The quality of the final product is based on the effectiveness of the drying process. The purpose of drying is to get the veneer sheet into target moisture content for gluing and hot pressing. Figure 28 shows the process of removing moisture from the peeled wet veneer sheets.

Figure 28. Overview of drying process

Veneer sheets are fed one by one on the rollers of the drying machine. The temperature of the drying machine is determined by the moisture content in the wet veneer and the wood species. The temperature inside the dryer is up to 200 °C.

The drying speed is adjusted according to the thickness, temperature, and the initial moisture content of wet sheets and the moisture content after the dryer. Veneer sheets are fed into the dryer rollers in the grain direction; while passing through the rollers, hot and humid air is

blown on the surface of the wet veneers. After the cooling process, veneer sheets are sorted, graded and stacked for further processing according to the automated visual analysis and moisture content available in the dried veneer sheets.

After the drying process, the width of the veneer gets shrinked up to 10 cm because of the different shrinking properties of wood areas. Visual defects appear according to log diameter; the smaller logs have better grade veneers because the knots are still very small in the heartwood, and sometimes even without knots. After the drying process, full-size sheets are stacked and transported directly to the lay-up or scarf jointing. Broken or defective veneers are stacked for the composing process.

4.2 Data and Methodology

Traditionally, in the wood industry, one of the most widely used ways to evaluate the quality of wood veneer has been random testing. It is done by selecting random processed sheets and evaluating the quality. If the quality is not met, then the production line has to stop to identify the problems in the peeling and drying processes which cause losses in productivity and revenue. Due to this, veneer/LVL manufacturing industries are investing in the automation of their processes, digital twins of the manufacturing line and crowdsourcing to increase their product quality standards, reliability, and efficiency (Urbonas et al., 2019).

The veneer sheet images dataset (5019 peeling and 3363 drying images) is provided by the Raute Corporation and comprises two types of grayscale veneer sheet images. The first set is of wet peeled veneer sheets, and the second is of dried sheets after drying process. Images of peeled and dried veneer sheets are captured and stored in a timestamp after peeling from the wood log and drying process.

A second dataset based on the fingerprint images for each peeled and dried veneer sheet image is also provided by the Raute Corporation. These fingerprint images are created by applying Canny edge detector, Gabor filter-based enhancement, and skeletonizing each image. Algorithms 1 and 2 provide more details on the Canny edge detector and how fingerprint images of veneer sheets were created. Furthermore, the provided dataset was unlabelled, and the aim was to develop an exploratory analysis of the dataset to find matching sheets.

Steps for extracting insights from big data mentioned in Figure 6 have been followed in this research, from capturing data (veneer sheets images) to interpreting the results. Considering different big data analytics techniques mentioned in Figure 8, descriptive analytics techniques, including data mining, statistical analysis, and multimedia (images) analytics, have been performed during the implementation of the model.

Firstly, veneer sheets images after peeling and drying processes have been captured. From each image, GLCM textural features have been calculated in an *n x m* matrix with 20 features representing texture properties calculated on neighbouring pixel distance and for angles. Data transformation methods have been applied to the data in order to transform it into a uniform format required for the data mining; for this, data normalization is applied to scale down calculated GLCM texture features into the smaller range [0,1] by applying L2 (square root of the sum of the squared vector) normalization.

After normalizing the calculated GLCM texture feature values, dimensionality reduction has been applied to the *n x m* data matrix by applying PCA to the normalized texture features of the veneer images. Feature extraction plays an integral part in the thesis in order to identify the most unique, explainable, and limited group of features to match the veneer sheet images with candidate images having high similarities. The cosine similarity method is used with a threshold of 90% of similarity to calculate the similarities between the transformed values of the veneer sheets. The results of similar veneer sheets are evaluated based on expert knowledge.

In image processing, texture analysis of the image is one of the main characteristics in content-based image retrieval, region of interest, comparison, or object detection in images. (Haralick et al., 1973) states, *"Texture is one of the important characteristics used in identifying objects or regions of interest in an image, whether the image is a photomicrograph, an aerial photograph, or a satellite image."* Veneer sheet also contains textures and grain patterns that can be used to compare veneer sheets to obtain matching candidates in a set of images. A set of candidate sheets after drying wet peeled sheets can be developed by utilizing statistical characteristics of the veneer sheet image. For extracting the textual features from each veneer peeled and dried sheet image, Gray Level Co-occurrence Matrices (GLCM) method is applied in this thesis. GLCM is a method of extracting second-order statistical texture features from an image by considering the relationship between pair of pixels (reference and neighbour pixels) in the image. It is introduced by the (Haralick et

al., 1973) to extract features based on gray-tone spatial dependencies and then take the mean and range of each measure (14 texture features; see Table 7) and use these values in the classification.

GLCM is a statistical method that counts the relative frequencies of relationship values for pairs of pixels $p(i,j)$ in an image separated by distance d, angle θ = 0, π/4, π/2, 3π/4 and normalized to probabilities. These descriptors utilize the spatial relationship between grayscale from distinct points in an image to describe the texture. Such matrices of gray-tone spatial dependence frequencies are based on the angular relationship in vertical (0°), horizontal (90°), and diagonal (45° and 135°) direction and the neighbouring distance, as illustrated in the Figure 29. Generally, the distance between the pair of pixels is one pixel, but it can be increased to more pixels and can be selected based on results, whereas the size of the GLCM matrix is dependent on the number of pixels in the analysed texture area that can be selected on different window sizes by dividing the image into multiple segments.

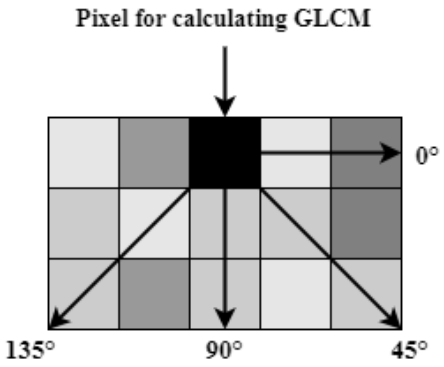


Figure 29. GLCM w.r.t four directions of edge detection

Co-occurrence matrix mathematical equation for size of G × G from a *N x M* image region, shown in (1)

$$CM_{d,\theta}(i,j) = \left|\{(n,m),(n+d_y,m+d_x); I(n,m)=i, l(n+d_y,m+d_x)=j\}\right| \qquad (1)$$

where:

$(n, m), (n + d_y, m + d_x) \in NxM$

i = Value of pixel at position (n, m)

j = Value of pixel at position $(n + d_y, m + d_x)$

G = Value of minimum and maximum difference in pair of pixels in an image.

This thesis focuses on comparing similarities between veneer sheets textures and grains to group similar candidate images peeled and dried from the same wood log. Initially, different combinations of GLCM parameters (14 features, pixel distances, and angles) are calculated for the veneer sheet images and then similarities between GLCM texture features of images.

4.2.1 GLCM texture features used in this research

In this research, five GLCM features have been applied based on the selection from edge texture and interior texture groups. The edge texture group generates high values when there are rapid changes in values between the neighbouring pixels, i.e. pixels in the neighbourhood contain visual edges. Whereas the interior texture group generates high values when there are few consistent edges but contain multiple uneven and subtle variations between the neighbouring pixels (Hall-Beyer, 2017). Dissimilarity and contrast have been selected from the edge texture group, whereas energy, correlation and homogeneity are selected from the interior texture group. Table 7 shows the remaining GLCM texture features that are not used in this research.

Contrast refers to measuring local variations or intensity of the reference pixel and its neighbouring pixels with the specified angle and distance. If there is significant amount of variation in an image, the contrast will be high (Haralick et al., 1973; Indra et al., 2022).

$$Contrast = \sum_{n=0}^{N_{g-1}} n^2 \left\{ \sum_{\substack{i=1 \\ |i-j|=n}}^{N_g} \sum_{j=1}^{N_g} p(i,j) \right\} \qquad (2)$$

Correlation shows the linear connectivity of the gray level value of one pixel relative to the other pixel in the GLCM, indicating the local gray-level dependency on the texture image.

Similar gray-level areas in the image give higher values of the correlation (Haralick et al., 1973; Indra et al., 2022).

$$Correlation = \frac{\sum_i \sum_j (ij) p(i,j) - \mu_x \mu_y}{\sigma_x \sigma_y} \quad (3)$$

Distance between the pair of pixels in the region of interest (distance and angle) is measured by dissimilarity. A higher value indicates a more significant disparity in the intensity levels of adjoining pixels (Haralick et al., 1973; Indra et al., 2022).

$$Dissimilarity = \sum_{i,j-0}^{N_{g-1}} p(i,j)|i-j| \quad (4)$$

The energy in GLCM is calculated from the angular second moment (ASM), which represents the homogeneity of an image, as shown in (5). Energy is used to calculate the local homogeneity and uniformity levels in an image's texture, and it is the inverse of the entropy. A higher level of texture homogeneity in an image gives higher energy values as homogeneous images contain few gray levels, which results in few GLCM values but relatively higher values of a pixel. Values of energy are in the range of [0,1], where 1 represents the maximum levels of homogeneity in an image (Haralick et al., 1973; Indra et al., 2022).

$$ASM = \sum_i \sum_j \{P(i,j)\}^2 \quad (5)$$

$$Energy = \sqrt{ASM} \quad (6)$$

Homogeneity refers to how close a pixel's distribution is in a GLCM. It is inversely proportional to the contrast; if there is a significant amount of contrast in an image, then homogeneity decreases. There will be high levels of homogeneity in an image if the values of the co-occurrence matrix are concentrated along the diagonal. If the homogeneity values

are greater and closer to 1, the texture has an ideal repeating structure, and if the values are low or near to 0, the texture element has numerous variations and is not uniformly distributed over the texture region (Haralick et al., 1973; Indra et al., 2022).

$$Homogenity = \sum_{i=1}^{N_g}\sum_{j=1}^{N_g}\frac{p(i,j)}{1+(i-j)^2} \qquad (7)$$

Table 7. GLCM texture features (Alsmadi, 2020; Hall-Beyer, 2017; Haralick et al., 1973; Indra et al., 2022; Wang and Ren, 2014)

GLCM Feature	Equation
Sum of squares (SS) : Variance	$f_4 = \sum_i\sum_j (i-u)^2 p(i,j)$
Sum average (SA)	$f_6 = \sum_{i=2}^{2N_g} i p_{x+y}(i)$
Sum variance (SV)	$f_7 = \sum_{i=2}^{2N_g} (i-f_8)^2 p_{x+y}(i)$
Sum entropy (SE)	$f_8 = -\sum_{i=2}^{2N_g} p_{x+y}(i)\log\{p_{x+y}(i)\}$
Entropy	$f_9 = -\sum_i\sum_j p(i,j)\log(p(i,j))$
Difference variance (DV)	$f_{10} = variance\ of\ p_{x-y}$
Difference entropy (DE)	$f_{11} = -\sum_{i=0}^{N_{g-1}} p_{x-y}(i)\log\{p_{x-y}(i)\}$
Maximum correlation coefficient (MCC)	$f_{12} = (second\ largest\ eigenvalue\ of\ Q)^{1/2}$ Where: $Q(i,j) = \sum_k \frac{p(i,k)p(j,k)}{p_x(i)p_y(k)}$
Information measures of correlation (IOC)	$f_{13} = \frac{HXY - HXY1}{\max\{HX, HY\}}$ $f_{14} = (1-\exp[-2.0(HXY2) - HXY])^{1/2}$ $HXY = -\sum_i\sum_j p(i,j)\log(p(i,j))$

	Where HX and HY are entropies of p_x and p_y $HXY1 = -\sum_i \sum_j p(i,j) \log\{P_X(i)p_y(j)\}$ $HXY2 = -\sum_i \sum_j p_x(i)p_y(j) \log\{P_X(i)p_y(j)\}$

GLCM texture features energy, correlation, homogeneity, contrast, and dissimilarity are calculated with pixel distance d = 1 and at angles θ = (0°,45°,90°,135°). Figure 30 shows the proposed system for extracting and comparing image similarity presented in a block diagram.

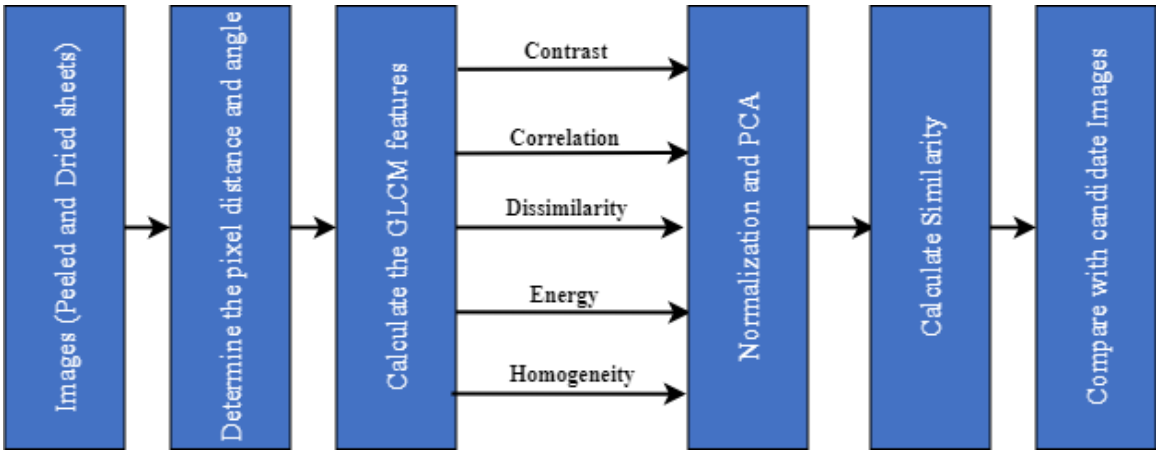


Figure 30. Proposed system for extracting and comparing image similarity presented in a block diagram

Figure 31 shows the GLCM feature texture values calculated on different angles with distance d = 1 for a set of hundred consecutive dry veneer sheet images. The figure shows that there are high variations in the values of all five calculated GLCM texture values at angle 0°, but there are only minor variations in angles 45°,90°, and 135°. From all four angles, it can be observed that there are high levels of contrast in the images, which indicates that a significant number of local variations are present between the reference and neighbouring pixels. It can also observed from the figure that homogeneity is inversely proportional to the contrast and have values closer to 0, which indicates that texture does not

have an ideal repeating structure. Except angle 0, there are no high values in dissimilarity, which indicates there is no significant disparity in the intensity levels of adjoining pixels.

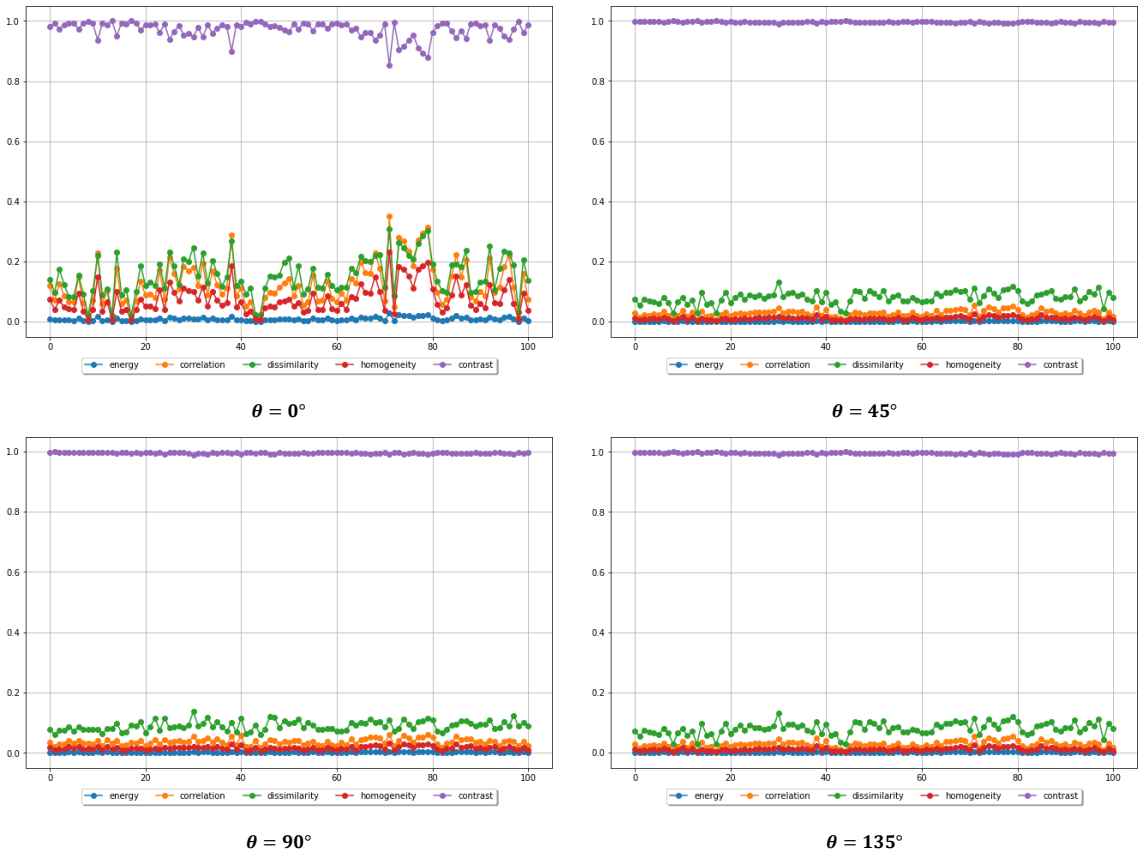


Figure 31. GLCM texture on different angle with distance of one pixel

After calculating the five GLCM texture features for each image with pixel distance d = 1 at four angles, a total of twenty features has been created. The calculated features are normalized by the L2 (square root of the sum of the squared vector) method, and lastly, PCA is applied to normalized values of GLCM texture features before further processing.

At first, a query image is selected from the set of sheets for which GLCM texture features are calculated. To calculate the similarity between the query image and the set of images cosine similarity method has been used. Cosine similarity computes similarity as a normalized dot product of X and Y on L2-normalized data as in the (8).

$$Cosine\ similarity = S_c(A,B) \coloneqq \cos(\theta) = \frac{A.B}{||A||.||B||} = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2}\sqrt{\sum_{i=1}^{n} B_i^2}} \quad (8)$$

Cosine similarity returns the similarity values for each image in the dataset with the queried image. To get the highest similar candidates related to queried image, cosine similarities of the images having similarity values equal to or larger than 90% have been selected, out of which the top 3 candidates are selected.

4.2.2 Edge detection and Fingerprints

During this research, it has been observed that the GLCM model seems to work well for highly textured images, but it does not give promising results if the veneer image contains low texture. To optimize the model for the low texture images, fingerprint images have been created by applying the Canny edge detector, Gabor filter-based enhancement and skeletonizing on each image. Canny edge detection is a multi-stage algorithm with five steps shown in algorithm 1, whereas algorithm 2 shows the process of creating fingerprint images using the original veneer sheet images.

Algorithm 1. Canny edge algorithm (Adapted from Wikipedia)

Begin

1. To smooth the image and remove the noise apply Gaussian filter.
2. Calculate the intensity gradients.
3. Gradient magnitude thresholding.
4. Double threshold for determining the potential edges.
5. Apply hysteresis for edge tracking.

End

Using canny edge detector method fingerprint image of each veneer sheet has been generated. Algorithm 2 shows the steps for generating fingerprint image of original veneer sheet image and Figure 32 shows fingerprint image of veneer sheet image.

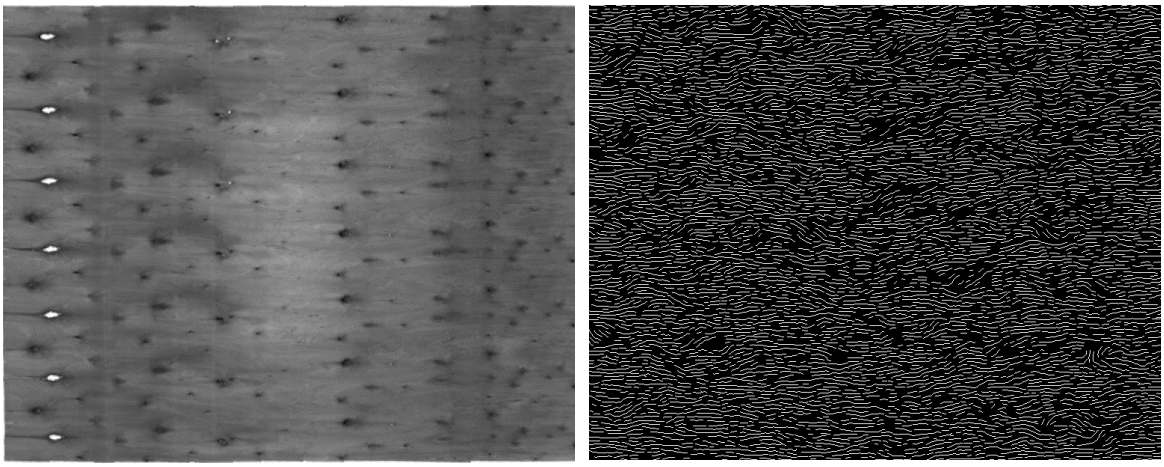


Figure 32. Original veneer sheet image (left) and Fingerprint image of the same veneer sheet (right)

Algorithm 2. Creating fingerprint image of a veneer sheet

Begin

1. Calculate the diagonal size of the input image as the image size changes from peeling to drying and absolute pixel lengths are not applicable.
2. Calculate canny edges within the given minimum and maximum threshold values that proportional to the diagonal image size (from (1))
3. Enhance the calculated edges by applying Gabor filter-based enhancement.
4. To remove noise from the image, all bright structures in an image are removed whose surface is less than the threshold value (e.g., 0.005 x Diagonal size of the input image).
5. Skeletonize enhanced edges.
6. Calculate the bitwise OR of the input image and skeletonize enhanced edges.

End

Instead of using original veneer sheets images, fingerprint images of veneer sheets are used to calculate the GLCM texture features (energy, correlation, homogeneity, contrast, and dissimilarity) with pixel distance d = 1 and angle θ = (0°, 90°, 45° and 135°).

Peeling and drying images of veneer sheets are captured and stored in order of their timestamp. Subsequent peeling images always form logs for clarity, but drying sheets are mixed up during the process. To sort the drying sheets according to their similarity and group them in one set so that it can be identified that the sheets belong to the same peeled wood log. For this purpose, a batch of consecutive fifty images have been used to test how the model works as some images have high texture and some have low texture. Algorithm 3 shows the steps to find similar images in a group of consecutive images.

Algorithm 3. Consecutive images to one log sheet according to texture and grain similarity

Begin

1. A timestamp-based batch of fingerprint images.
2. Calculate GLCM texture with pixel distance d = 1 at angle θ = (0°, 90°, 45° and 135°).
3. Normalize the GLCM feature values.
4. Apply Principal component analysis (PCA).
5. Use transformed values of GLCM.
6. Select the first image from the stack of the fingerprint images.
7. Calculate the cosine similarity between the transformed GLCM values of the first selected fingerprint image in the stack and the rest of the fingerprint images.
8. Sort similarity values in descending order and select top images having a similarity of 90% or greater.
9. Store the selected consecutive candidates from the stack and remove the fingerprint images from the existing stack.
10. Perform again step 7, step 8 and step 9 till there is no image left in the stack.
11. Group the stored selected consecutive candidates from step 9 and output the consecutive stacked result.

End

5 Results

In this thesis, two approaches have been applied for selecting similar candidates' veneer sheets. In the first approach, GLCM texture features are calculated for the original images of veneer sheets with the distance of d =1 and angles θ = (0°, 90°, 45° and 135°). GLCM texture values have been normalized, and PCA is applied to the normalized values of the GLCM texture features. The cosine similarity method calculates the similarities between the query image and other images in the set. Veneer sheets having similarity of GLCM texture features of 90% or greater with the query image are selected; out of these selected candidates top 3 images have been selected as the most similar. The Figure 33 shows the stack of similar candidates for the query image (1st from the left)

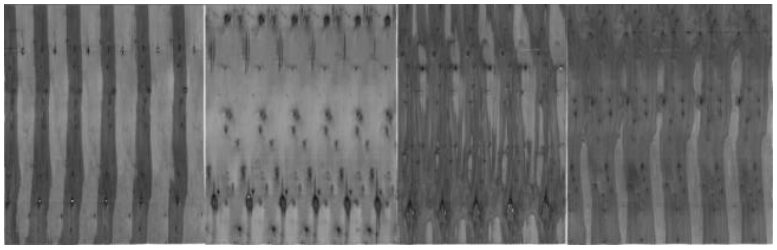


Figure 33. Candidates by using GLCM texture features of original sheet image

It is observed that when using algorithm 3 for sheets with high texture, the returned candidates are at least similar looking, yet, probably originating from different logs. With lower texture in the veneer sheets as in Figure 34 the model performance is poor. To optimize the model, fingerprint image of each veneer sheet image has been created as mentioned in algorithms 1 and 2, and Figure 31 shows the fingerprint image for the original veneer sheet image.

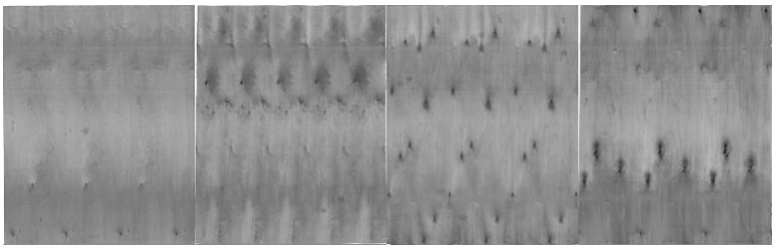


Figure 34. Low texture candidates by using GLCM texture features of original sheet image

Next, the fingerprint images were used to calculate the same GLCM texture features, pixel distance and angles used in calculating GLCM features for the original image in the first approach. The reason behind using fingerprint images is that GLCM works well for images with high texture, but it did not provide promising results where images have low textures. Using fingerprint images instead of original images also allows the ability to match low-textured images. The Figure 35 shows the stack of similar candidates for the query image (1st from the left) using fingerprint images.

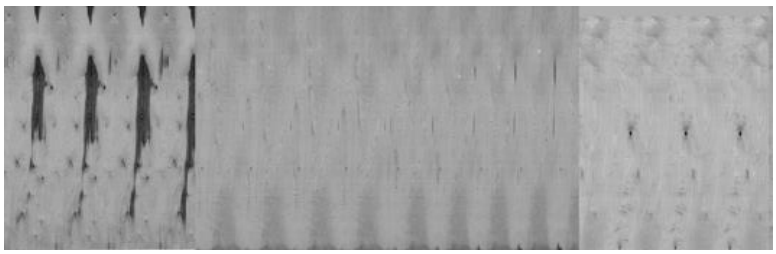


Figure 35. Candidates by using GLCM texture features of fingerprint image

A set of 50 consecutive drying images has been selected based on timestamps to sort them according to their similarity to identify whether drying images belong to the same peeled wood log. Figure 36 shows the veneer sheets sorted according to similarity on texture and grains. The first image is selected from the stack, and the similarity of the selected fingerprint image has been calculated with all other fingerprint images. Fingerprints having similarities equal to or greater than 90% have been removed and stored separately from the existing stack of the images. This process is repeated until all consecutive images have been sorted, as in algorithm 3.

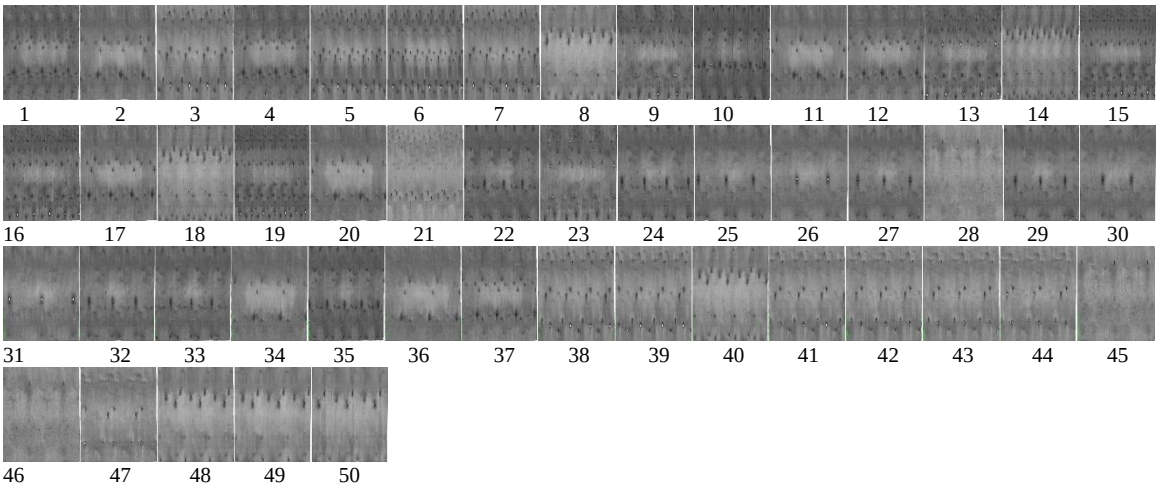


Figure 36. Consecutive veneer sheets similarity using fingerprints

It has been observed that fingerprint images match images with high textures and also those with low textures.

5.1 Result analysis

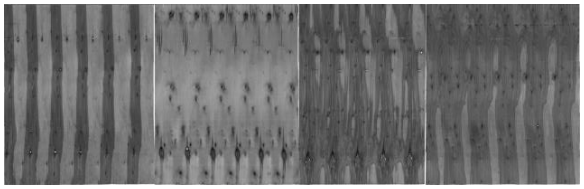
As it has been discussed in the theoretical background of the thesis that intelligent manufacturing is a multidimensional concept to optimize production processes through advanced data analysis, manufacturing technologies, and system engineering. Different challenges arise with multidimensionality processes, such as volume, variety, and veracity, as shown in Figure 4. The data used in this research is based on images of veneer sheets which also reflects multiple data challenges.

It has been found that using GLCM texture features of original veneer sheets works well for the images with high texture, but it does not work well where texture is low. In order to get similar sheets with low textures, fingerprint images have been created for each veneer sheet image as mentioned in algorithms 1 and 2, and then again, all data pre-processing methods discussed earlier have been applied to the texture features of fingerprint images. Using GLCM texture features of fingerprint images gives promising results; the model identifies similar candidates for both high- and low-texture veneer sheets as in Figure 35.

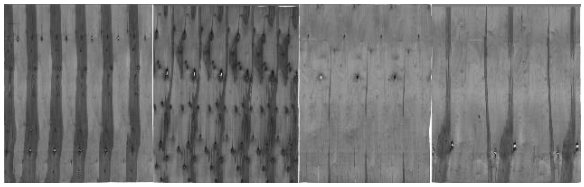
According to the current literature review, (Alsmadi, 2020; Haryanto et al., 2020; Surabaya et al., n.d.; Wang and Ren, 2014; Zhang et al., 2021) applied GLCM to calculate the texture features of the images. Some authors have calculated all GLCM texture features, and some have calculated for fewer texture features. These texture features have been grouped into the edge and interior texture groups. (Hall-Beyer, 2017) presented guidelines in the selection of GLCM texture features which have been followed in this thesis for the selection of texture features. Contrast, correlation, dissimilarity, energy, and homogeneity GLCM texture features have been calculated for both available veneer sheet original and fingerprint images with pixel distance d=1 and angle θ = (0°, 45°, 90° and 135°).

In current literature, (Haryanto et al., 2020) calculated ASM, contrast, correlation, dissimilarity, energy, and homogeneity with pixel distance d =5 and angles θ = (0°, 45°, 90° and 135°) for histopathology images. (Surabaya et al., n.d.) calculated ASM, IDM, contrast, entropy, and correlation with pixel distance d = 1 and for only one angle θ = 0°. (Kamal et al., 2017) calculated contrast, correlation, energy, and homogeneity GLCM texture feature for the wood images containing different types of wood knots defects. (Alsmadi, 2020) calculated contrast, energy, homogeneity, standard deviation, and mean for each 4 x 4 block in the image at angles θ = (0°, 45°, 90° and 135°).

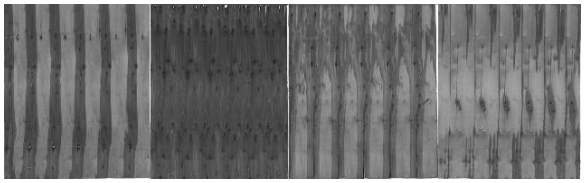
When it comes to the implementation of GLCM, in the current literature, the maximum value of pixel distance is usually d = 5, whereas most of the authors have used pixel distance d =1 to calculate the GLCM texture features in angles θ = (0°, 45°, 90° and 135°). It has been justified by (Haryanto et al., 2020) that too far pixel distance cause information between pixel to be irrelevant. Furthermore, in this thesis, initially, GLCM texture features for different pixels distance d = (1,5, 7, 9) have been calculated, which are shown in Figure 37, and in the proposed model pixel distance d =1 has been used to calculate GLCM texture features.



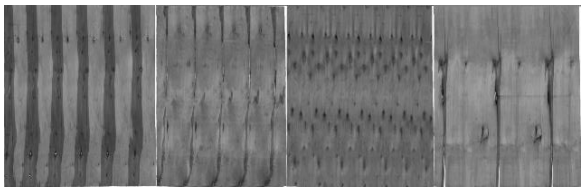
d = 1



d = 5



d = 7



d = 9

Figure 37. Candidate images with distance d = (1,5,7,9) and angles (0°, 45°, 90° and 135°)

In this thesis, fingerprint images have been created using the canny edge detector to get candidate sheets with low texture. (Alsmadi, 2020) also used the canny edge detector to extract the shape features, and by applying canny edge, diverse shapes existed in the image were obtained to improve the similarities between the query image and the images in the database. In this thesis, the canny edge method improves the results by selecting images as candidate images that have low texture.

Veneer sheets have variations in the texture; some have low texture, and some have high. It is easy to determine the similar sheets after the peeling process, but after the drying process, it is complicated to match similar sheets to form logs for clarity, as the veneer sheets get mixed during the drying process. The proposed method in this thesis shown in the Figure 38 can be used to form logs for clarity from a set of consecutive dried veneer sheet images by using fingerprint images to calculate GLCM texture features (Contrast, correlation,

Dissimilarity, Energy, and Homogeneity) with distance d=1 and angles = (0°,45°,90°,135°). Normalizing the GLCM texture feature values with L2, applying PCA on normalized values, and calculating the cosine similarity on the transformed values resulted in promising results for both low and high texture veneer sheet images. This model can be used to optimize the manufacturing process of the veneer/LVL industry by forming a log of similar dried veneer sheets for further processing.

6 Conclusion and discussion

The first objective of this thesis was to examine how machine learning algorithms have been applied in the veneer/LVL industry to optimize the peeling and drying process of veneer sheets. The second objective was to optimize the veneer/LVL manufacturing process to develop a data-based method to identify similar veneer sheets after the drying process according to their textural features and grains.

In order to comprehensively obtain the research objective, two research questions have been identified at the beginning of the thesis. The answer to the first research question is below, based on understanding the theoretical background and literature review.

RQ1. According to the current literature, how Machine Learning algorithms have been applied in the data-based optimization of drying process specifically in the veneer/LVL industry?

In general, multiple methodologies are currently being utilized in the veneer/LVL industry to optimize peeling and drying processes using the data generated from sensors and image processing of veneer sheets. In the current perspective, most of the research has been done on classification methods to identify the wood knot defects (leaf, dry, sound, horn and edge) and quality analysis of the veneer sheets before and after the drying process. Another research area in the drying process is to optimize the process in such a way to reduce the energy consumption in the drying process, which is approximately fifty percent of the mill's energy in the plywood manufacturing and lastly.

In this thesis, three approaches have been studied and evaluated using the texture features data from the veneer sheets to answer the second research question.

RQ2. How to enhance the peeling-drying process through data-based optimization in the given LVL-manufacturing case example through the use of data mining and machine learning?

The first approach was using GLCM texture features and calculating the similarities of the query image based on texture similarity. The second approach was implemented using GLCM texture features, normalizing the features, applying PCA and using the transformed values to calculate the similarities. The last approach was that instead of original veneer sheet images, fingerprints of the images were used to calculate the GLCM texture features. During the study, it was observed that extracting GLCM texture features on fingerprint images gives promising results compared to the original veneer sheet images used in the previous two approaches. However, the preliminary results show some similarities with the candidate sheets as the data was unlabelled, which provides direction to further analysis with fingerprint images.

6.1 Future research

This thesis provides a solid foundation for future studies. The research related to the thesis's topic shows several new ideas that can be investigated further. A significant topic of interest of this thesis may be to widen the scope and experiment with other image edge detecting techniques such as Sobel edge detection and Prewitt edge detection on the original images and evaluate the results on different similarity calculating methods.

In general, this thesis evaluates the use of GLCM texture features results in the veneer sheets and fingerprint images. The overall results are based on the research based on the literature review related to the veneer sheet peeling and drying methodologies. Furthermore, literature reviews of the papers from other disciplines other than the wood industry, like biomedical, have also been done. It is observed that applying image processing methods from biomedical imaging field in the veneer industry will be worth trying.

References

Adams, W.R., 2017. High-accuracy detection of early Parkinson's Disease using multiple characteristics of finger movement while typing. PLoS ONE 12. https://doi.org/10.1371/journal.pone.0188226

Adeyeri, M.K., Mpofu, K., Adenuga Olukorede, T., 2015. Integration of agent technology into manufacturing enterprise: A review and platform for industry 4.0. IEOM 2015 - 5th International Conference on Industrial Engineering and Operations Management, Proceeding. https://doi.org/10.1109/IEOM.2015.7093910

Agarwal, S., 2014. Data mining: Data mining concepts and techniques. Proceedings - 2013 International Conference on Machine Intelligence Research and Advancement, ICMIRA 2013 203–207. https://doi.org/10.1109/ICMIRA.2013.45

Ahmed, S.S., Cool, J., Karim, M.E., 2020. Application of decision tree-based techniques to veneer processing. Journal of Wood Science 66. https://doi.org/10.1186/S10086-020-01904-0

Alcácer, V., Cruz-Machado, V., 2019. Scanning the Industry 4.0: A Literature Review on Technologies for Manufacturing Systems. Engineering Science and Technology, an International Journal 22, 899–919. https://doi.org/10.1016/J.JESTCH.2019.01.006

Alsmadi, M.K., 2020. Content-Based Image Retrieval Using Color, Shape and Texture Descriptors and Features. Arabian Journal for Science and Engineering 45, 3317–3330. https://doi.org/10.1007/s13369-020-04384-y

Atzori, L., Iera, A., Morabito, G., 2010. The Internet of Things: A survey. Computer Networks 54, 2787–2805. https://doi.org/10.1016/j.comnet.2010.05.010

Belhadi, A., Zkik, K., Cherrafi, A., Yusof, S.M., el fezazi, S., 2019. Understanding Big Data Analytics for Manufacturing Processes: Insights from Literature Review and Multiple Case Studies. Computers and Industrial Engineering 137. https://doi.org/10.1016/j.cie.2019.106099

Bersimis, S., Panaretos, J., Psarakis, S., 2009. Multivariate Statistical Process Control Charts and the Problem of Interpretation: A Short Overview and Some Applications in Industry. https://doi.org/10.48550/arxiv.0901.2880

Bortolini, M., Ferrari, E., Gamberi, M., Pilati, F., Faccio, M., 2017. Assembly system design in the Industry 4.0 era: a general framework. IFAC-PapersOnLine 50, 5700–5705. https://doi.org/10.1016/J.IFACOL.2017.08.1121

Camizuli, E., Carranza, E.J., 2018. Exploratory Data Analysis (EDA), in: The Encyclopedia of Archaeological Sciences. John Wiley & Sons, Inc., Hoboken, NJ, USA, pp. 1–7. https://doi.org/10.1002/9781119188230.saseas0271

Castelo-Branco, I., Oliveira, T., Simões-Coelho, P., Portugal, J., Filipe, I., 2022. Measuring the fourth industrial revolution through the Industry 4.0 lens: The relevance of resources, capabilities and the value chain. Computers in Industry 138, 103639. https://doi.org/10.1016/J.COMPIND.2022.103639

Cheng, Y., Chen, K., Sun, H., Zhang, Y., Tao, F., 2018. Data and knowledge mining with big data towards smart production. J Ind Inf Integr 9, 1–13. https://doi.org/10.1016/J.JII.2017.08.001

Çolak, S., Çolakoğlu, G., Aydin, I., 2007. Effects of logs steaming, veneer drying and aging on the mechanical properties of laminated veneer lumber (LVL). Building and Environment 42, 93–98. https://doi.org/10.1016/J.BUILDENV.2005.08.008

Demirkir, C., Özsahin, Ş., Aydin, I., Colakoglu, G., 2013. Optimization of some panel manufacturing parameters for the best bonding strength of plywood. International Journal of Adhesion and Adhesives 46, 14–20. https://doi.org/10.1016/J.IJADHADH.2013.05.007

Gandomi, A., Haider, M., 2015. Beyond the hype: Big data concepts, methods, and analytics. International Journal of Information Management 35, 137–144. https://doi.org/10.1016/j.ijinfomgt.2014.10.007

Ge, Z., Song, Z., Ding, S.X., Huang, B., 2017. Data Mining and Analytics in the Process Industry: The Role of Machine Learning. IEEE Access 5, 20590–20616. https://doi.org/10.1109/ACCESS.2017.2756872

Giusto, D., Iera, A., Morabito, G., Atzori, L. (Eds.), 2010. The Internet of Things. Springer New York, New York, NY. https://doi.org/10.1007/978-1-4419-1674-7

GTAI, n.d. Industrie 4.0. [WWW Document]. URL https://www.gtai.de/en/invest/industries/industrial-production/industrie-4-0 (accessed 4.11.22).

Gubbi, J., Buyya, R., Marusic, S., Palaniswami, M., 2013. Internet of Things (IoT): A vision, architectural elements, and future directions. Future Generation Computer Systems 29, 1645–1660. https://doi.org/10.1016/J.FUTURE.2013.01.010

Hall-Beyer, M., 2017. Practical guidelines for choosing GLCM textures to use in landscape classification tasks over a range of moderate spatial scales. International Journal of Remote Sensing 38, 1312–1338. https://doi.org/10.1080/01431161.2016.1278314

Han, C., Zhan, T., Xu, J., Jiang, J., Lu, J., 2015. Process Optimization for Multi-Veneer Hot-Press Drying. Drying Technology 33, 735–741. https://doi.org/10.1080/07373937.2014.983243

Haralick, R.M., Dinstein, I., Shanmugam, K., 1973. Textural Features for Image Classification. IEEE Transactions on Systems, Man and Cybernetics SMC-3, 610–621. https://doi.org/10.1109/TSMC.1973.4309314

Haryanto, T., Pratama, A., Suhartanto, H., Murni, A., Kusmardi, K., Pidanic, J., 2020. Multipatch-GLCM for texture feature extraction on classification of the colon histopathology images using deep neural network with GPU acceleration. Journal of Computer Science 16, 280–294. https://doi.org/10.3844/JCSSP.2020.280.294

Hashem, I.A.T., Yaqoob, I., Anuar, N.B., Mokhtar, S., Gani, A., Ullah Khan, S., 2015. The rise of “big data” on cloud computing: Review and open research issues. Information Systems 47, 98–115. https://doi.org/10.1016/J.IS.2014.07.006

Houari, R., Bounceur, A., Kechadi, M.T., Tari, A.K., Euler, R., 2016. Dimensionality reduction in data mining: A Copula approach. Expert Systems with Applications 64, 247–260. https://doi.org/10.1016/J.ESWA.2016.07.041

Indra, D., Fadlillah, H.M., Kasman, Ilmawan, L.B., 2022. Rice Texture Analysis Using GLCM Features. Institute of Electrical and Electronics Engineers (IEEE), pp. 1–5. https://doi.org/10.1109/icecet52533.2021.9698594

Institut Teknologi 10 Nopember (Surabaya, I.F.T.E., Institute of Electrical and Electronics Engineers. Indonesia Section, Institute of Electrical and Electronics Engineers, n.d. 2017 International Seminar on Intelligent Technology and Its Application (ISITIA) : proceeding : Surabaya, Indonesia, August, 28-29, 2017.

Ishwarappa, Anuradha, J., 2015. A Brief Introduction on Big Data 5Vs Characteristics and Hadoop Technology. Procedia Computer Science 48, 319–324. https://doi.org/10.1016/J.PROCS.2015.04.188

Jagadish, H. v., 2015. Big Data and Science: Myths and Reality. Big Data Research. https://doi.org/10.1016/j.bdr.2015.01.005

Jalonen, T., Laakom, F., Gabbouj, M., Puoskari, T., 2021. Visual Product Tracking System Using Siamese Neural Networks. IEEE Access 9, 76796–76805. https://doi.org/10.1109/ACCESS.2021.3082934

Kamal, K., Qayyum, R., Mathavan, S., Zafar, T., 2017. Wood defects classification using laws texture energy measures and supervised learning approach. Advanced Engineering Informatics 34, 125–135. https://doi.org/10.1016/J.AEI.2017.09.007

Kerin, M., Pham, D.T., 2019. A review of emerging industry 4.0 technologies in remanufacturing. Journal of Cleaner Production 237, 117805. https://doi.org/10.1016/J.JCLEPRO.2019.117805

Komorowski, M., Marshall, D.C., Salciccioli, J.D., Crutain, Y., 2016. Exploratory data analysis. Secondary Analysis of Electronic Health Records 185–203. https://doi.org/10.1007/978-3-319-43742-2_15/FIGURES/18

Kourti, T., Lee, J., Macgregor, J.F., 1996. Experiences with industrial applications of projection methods for multivariate statistical process control. Computers & Chemical Engineering 20, S745–S750. https://doi.org/10.1016/0098-1354(96)00132-9

Lasi, H., Fettke, P., Kemper, H.G., Feld, T., Hoffmann, M., 2014. Industry 4.0. Business and Information Systems Engineering 6, 239–242. https://doi.org/10.1007/S12599-014-0334-4

Lestander, T.A., Holmberg, C., Stenberg, L., Lehtonen, R., 2012. Towards multivariate statistical process control in the wood pellet industry. Biomass and Bioenergy 45, 152–158. https://doi.org/10.1016/J.BIOMBIOE.2012.05.027

Lu, Y., 2017. Industry 4.0: A survey on technologies, applications and open research issues. J Ind Inf Integr 6, 1–10. https://doi.org/10.1016/J.JII.2017.04.005

Luo, Q., 2008. Advancing knowledge discovery and data mining. Proceedings - 1st International Workshop on Knowledge Discovery and Data Mining, WKDD 3–5. https://doi.org/10.1109/WKDD.2008.153

Nascimento, D.L.M., Alencastro, V., Quelhas, O.L.G., Caiado, R.G.G., Garza-Reyes, J.A., Rocha-Lona, L., Tortorella, G., 2019. Exploring Industry 4.0 technologies to enable circular economy practices in a manufacturing context. Journal of Manufacturing Technology Management 30, 607–627. https://doi.org/10.1108/JMTM-03-2018-0071

Nguyen, G., Dlugolinsky, S., Bobák, M., Tran, V., López García, Á., Heredia, I., Malík, P., Hluchý, L., 2019. Machine Learning and Deep Learning frameworks and libraries for large-scale data mining: a survey. Artificial Intelligence Review 52, 77–124. https://doi.org/10.1007/S10462-018-09679-Z

Park, S., 2016. Development of Innovative Strategies for the Korean Manufacturing Industry by Use of the Connected Smart Factory (CSF). Procedia Computer Science 91, 744–750. https://doi.org/10.1016/J.PROCS.2016.07.067

Peruzzini, M., Grandi, F., Pellicciari, M., 2017. Benchmarking of Tools for User Experience Analysis in Industry 4.0. Procedia Manufacturing 11, 806–813. https://doi.org/10.1016/J.PROMFG.2017.07.182

Philip Chen, C.L., Zhang, C.Y., 2014. Data-intensive applications, challenges, techniques and technologies: A survey on Big Data. Information Sciences 275, 314–347. https://doi.org/10.1016/j.ins.2014.01.015

PWC, 2016. PWC.2016.Industry 4.0: Building the Digital Enterprise. London: PWC.

Rajnai, Z., Kocsis, I., 2018. Assessing industry 4.0 readiness of enterprises, in: SAMI 2018 - IEEE 16th World Symposium on Applied Machine Intelligence and Informatics Dedicated to the Memory of Pioneer of Robotics Antal (Tony) K. Bejczy, Proceedings. https://doi.org/10.1109/SAMI.2018.8324844

Raschka, S., 2018. Model Evaluation, Model Selection, and Algorithm Selection in Machine Learning.

Reimann, C., Filzmoser, P., Garrett, R.G., Dutter, R., 2008. Statistical Distribution Measures. Statistical Data Analysis Explained 51–62. https://doi.org/10.1002/9780470987605.CH4

Roblek, V., Meško, M., Krapež, A., 2016. A Complex View of Industry 4.0. SAGE Open 6, 215824401665398. https://doi.org/10.1177/2158244016653987

Sezer, O.B., Dogdu, E., Ozbayoglu, A.M., 2018. Context-Aware Computing, Learning, and Big Data in Internet of Things: A Survey. IEEE Internet of Things Journal 5, 1–27. https://doi.org/10.1109/JIOT.2017.2773600

Shafiq, S.I., Sanin, C., Szczerbicki, E., Toro, C., 2015. Virtual Engineering Object / Virtual Engineering Process: A specialized form of Cyber Physical System for Industrie 4.0. Procedia Computer Science 60, 1146–1155. https://doi.org/10.1016/J.PROCS.2015.08.166

Sherman, R., 2015. Data Integration Processes. Business Intelligence Guidebook 301–333. https://doi.org/10.1016/B978-0-12-411461-6.00012-5

Simmert, B., Ebel, P.A., Peters, C., Bittner, E.A.C., Leimeister, J.M., 2018. Conquering the Challenge of Continuous Business Model Improvement. Business & Information Systems Engineering 2018 61:4 61, 451–468. https://doi.org/10.1007/S12599-018-0556-Y

Sivarajah, U., Kamal, M.M., Irani, Z., Weerakkody, V., 2017. Critical analysis of Big Data challenges and analytical methods. Journal of Business Research 70, 263–286. https://doi.org/10.1016/J.JBUSRES.2016.08.001

Sun, W., Cai, Z., Li, Y., Liu, F., Fang, S., Wang, G., 2018. Data processing and text mining technologies on electronic medical records: A review. Journal of Healthcare Engineering 2018. https://doi.org/10.1155/2018/4302425

Tang, H., Li, D., Wang, S., Dong, Z., 2017. CASOA: An Architecture for Agent-Based Manufacturing System in the Context of Industry 4.0. IEEE Access 6, 12746–12754. https://doi.org/10.1109/ACCESS.2017.2758160

Thames, L., Schaefer, D., 2016. Software-defined Cloud Manufacturing for Industry 4.0. Procedia CIRP 52, 12–17. https://doi.org/10.1016/J.PROCIR.2016.07.041

Tiryaki, S., Aydin, A., 2022. Multivariate Hotelling T2 Control Chart for Monitoring Some Quality Characteristics in Medium Density Fiberboard Manufacturing Process. Drvna industrija 73, 35–46. https://doi.org/10.5552/DRVIND.2022.2046

Trappey, A.J.C., Trappey, C. v, Govindarajan, U.H., Chuang, A.C., Sun, J.J., 2016. A review of essential standards and patent landscapes for the Internet of Things: A key enabler for Industry 4.0. https://doi.org/10.1016/j.aei.2016.11.007

Urbonas, A., Raudonis, V., Maskeliunas, R., Damaševičius, R., 2019a. Automated Identification of Wood Veneer Surface Defects Using Faster Region-Based Convolutional Neural Network with Data Augmentation and Transfer Learning. Applied Sciences 2019, Vol. 9, Page 4898 9, 4898. https://doi.org/10.3390/APP9224898

Urbonas, A., Raudonis, V., Maskeliunas, R., Damaševičius, R., 2019b. Automated Identification of Wood Veneer Surface Defects Using Faster Region-Based Convolutional Neural Network with Data Augmentation and Transfer Learning. Applied Sciences 2019, Vol. 9, Page 4898 9, 4898. https://doi.org/10.3390/APP9224898

Vabalasid, A., Gowen, E., Poliakoff, E., Casson, A.J., 2019. Machine learning algorithm validation with a limited sample size. https://doi.org/10.1371/journal.pone.0224365

Velliangiri, S., Alagumuthukrishnan, S., Thankumar Joseph, S.I., 2019. A Review of Dimensionality Reduction Techniques for Efficient Computation. Procedia Computer Science 165, 104–111. https://doi.org/10.1016/J.PROCS.2020.01.079

Veza, I., Mladineo, M., Gjeldum, N., 2015. Managing innovative production network of smart factories. IFAC-PapersOnLine 28, 555–560. https://doi.org/10.1016/J.IFACOL.2015.06.139

Vialetto, G., Noro, M., 2020. An innovative approach to design cogeneration systems based on big data analysis and use of clustering methods. Energy Conversion and Management 214, 112901. https://doi.org/10.1016/J.ENCONMAN.2020.112901

Wang, J., Xu, C., Zhang, J., Zhong, R., 2022. Big data analytics for intelligent manufacturing systems: A review. Journal of Manufacturing Systems 62, 738–752. https://doi.org/10.1016/J.JMSY.2021.03.005

Wang, J.-S., Ren, X.-D., 2014. GLCM Based Extraction of Flame Image Texture Features and KPCA-GLVQ Recognition Method for Rotary Kiln Combustion Working Conditions. International Journal of Automation and Computing 11, 72–77. https://doi.org/10.1007/s11633-014-0767-8

Ward, L., Dunn, A., Faghaninia, A., Zimmermann, N.E.R., Bajaj, S., Wang, Q., Montoya, J., Chen, J., Bystrom, K., Dylla, M., Chard, K., Asta, M., Persson, K.A., Snyder, G.J., Foster, I., Jain, A., 2018. Matminer: An open source toolkit for materials data mining. Computational Materials Science 152, 60–69. https://doi.org/10.1016/J.COMMATSCI.2018.05.018

Webster, J., Watson, R.T., 2002. Analyzing the Past to Prepare for the Future: Writing a Literature Review, Source: MIS Quarterly.

Xu, L. da, Xu, E.L., Li, L., 2018. Industry 4.0: State of the art and future trends. International Journal of Production Research 56, 2941–2962. https://doi.org/10.1080/00207543.2018.1444806

Xu, S., Lu, B., Baldea, M., Edgar, T.F., Wojsznis, W., Blevins, T., Nixon, M., 2015. Data cleaning in the process industries. Reviews in Chemical Engineering 31, 453–490. https://doi.org/10.1515/REVCE-2015-0022

Yuce, B., Mastrocinque, E., Packianather, M.S., Pham, D., Lambiase, A., Fruggiero, F., 2014. Neural network design and feature selection using principal component analysis and Taguchi method for identifying wood veneer defects. Production and Manufacturing Research 2, 291–308. https://doi.org/10.1080/21693277.2014.892442

Zebari, R.R., Mohsin Abdulazeez, A., Zeebaree, D.Q., Zebari, D.A., Saeed, J.N., 2020. A Comprehensive Review of Dimensionality Reduction Techniques for Feature Selection and Feature Extraction. Journal of Applied Science and Technology Trends 1, 56–70. https://doi.org/10.38094/jastt1224

Zhang, C., Cao, L., Romagnoli, A., 2018. On the feature engineering of building energy data mining. Sustainable Cities and Society 39, 508–518. https://doi.org/10.1016/J.SCS.2018.02.016

Zhang, F., Wang, Y., Wu, L., Liu, M., Hu, S., Li, M., 2021. GLM-Net: A multi-scale image segmentation network for brain abnormalities based on GLCM. Proceedings - 2021 14th International Congress on Image and Signal Processing, BioMedical Engineering and Informatics, CISP-BMEI 2021. https://doi.org/10.1109/CISP-BMEI53629.2021.9624341

Zheng, T., Ardolino, M., Bacchetti, A., Perona, M., 2020. The applications of Industry 4.0 technologies in manufacturing context: a systematic literature review. https://doi.org/10.1080/00207543.2020.1824085 59, 1922–1954. https://doi.org/10.1080/00207543.2020.1824085

Zhong, R.Y., Xu, X., Klotz, E., Newman, S.T., 2017. Intelligent Manufacturing in the Context of Industry 4.0: A Review. Engineering 3, 616–630. https://doi.org/10.1016/J.ENG.2017.05.015