DEVELOPING OPERATION AND ANALYTICS MODEL TO PREDICT AND OPTIMIZE SELECTIVE MANUFACTURING IN SEMICONDUCTOR SUPPLY CHAIN

MUHAMMAD RAZIN BIN SALIM

SCHOOL OF MECHANICAL ENGINEERING
UNIVERSITI SAINS MALAYSIA
2018

DEVELOPING OPERATION AND ANALYTICS MODEL TO PREDICT AND OPTIMIZE SELECTIVE MANUFACTURING IN SEMICONDUCTOR SUPPLY CHAIN

By:

MUHAMMAD RAZIN BIN SALIM

(MATRIX no. : 137847)

Supervisor:

Chin Jeng Feng,Assoc.Prof.Ir.Dr.

JULY 2021

This dissertation is submitted to

UNIVERSITI SAINS MALAYSIA

As partial fulfillment of the requirement to graduate with honors degree in

BACHELOR OF ENGINEERING (MECHANICAL ENGINEERING)

School of Mechanical Engineering

Engineering Campus

Universiti Sains Malaysia

ACKNOWLEDGEMENT

In the name of Allah SWT, the Most Gracious and the Most Merciful, on whom inevitably we lean for strength and wisdom. I express my humble gratitude to You for providing me with the perseverance necessary to accomplish this thesis during this pandemic. To begin, I would like to express my heartfelt gratitude and appreciation to my supervisor, Chin Jeng Feng, Assoc.Prof.Ir.Dr for his supervision, consistent support, and constructive comments. His prompt and efficient contribution aided me in shaping this into its final form, and he effectively walked me through some tense occasions, always willing to help me better comprehend this thesis. I would also like to express my gratitude to the academic and technical staff at the USM School of Mechanical Engineering for providing me with an academic foundation that enabled me to pursue this research. I would want to pay tribute to my late father, Salim Bin Abu Omar, who was vital throughout this triumph. It would be my greatest joy to experience this moment with him. I would also like to express my gratitude to my mother, Kamariah Binti Abdul Rashid, and my siblings and brothers for their prayers, encouragement, and unwavering compassion and moral support throughout my studies. To everyone who contributed directly to my study and to whose names I did not disclose, I want to convey my gratitude for their patience, love, support, and inspiration in assisting me in finishing this tough endeavour.

Muhammad Razin Bin Salim
July 2021

TABLE OF CONTENTS

LIST OF TABLES

Page

LIST OF FIGURES

Page

LIST OF ABBREVIATIONS

BD	Big Data
SHA	Sucessive Halving Algorithm
KNN	K-Nearest Neighbour Regressor
MLP	Multi-layer Perceptron Regressor Model
HT	Hoeffding Tree Regressor
HAT	Hoeffding Adaptive Tree
MAE	Mean Absolute Error
RMSE	Root mean squared error
SMAPE	Symmetric mean absolute percentage error
UCB	Upper Confidence Bound Regressor
EWA	Exponentially Weighted Average
USM	Universiti Sains Malaysia

ABSTRAK

Ketersediaan data besar (BD) adalah faktor kritikal digunakan oleh pembuat semikonduktor untuk meningkatkan ketepatan ramalan mereka. Pengilang dapat menggunakan data ini untuk meramalkan permintaan pengguna dengan tepat dalam pembuatan semikonduktor dan membuat tekaan terbaik mengenai jumlah setiap variasi produk yang akan dibuat melalui pembuatan terpilih berdasarkan risiko yang dikira. Potensi lima pembelajaran mesin yang berbeza telah dikaji dalam kajian ini: Linear Model Regression, Multi-layer Perceptron Regressor Model (MLP), K-Nearest Neighbour Regressor Model (KNN), Hoeffding Tree Regressor Model (HT), and Hoeffding Adaptive Tree Regressor (HAT). Tiga metrik statistik digunakan untuk menilai ketepatan model yang dibina: mean absolute error (MAE), root mean squared error (RMSE), and residual distribution. Dalam eksperimen Meta, Imblearn, Expert, dan Ensemble telah digunakan untuk meningkatkan prestasi. Seiring dengan modelnya, kaedah pengesanan diaplikasikan seperti ADWIN ke dalam eksperimen kami untuk melihat perubahan pesat dalam menghasilkan BD. Perbandingan kajian menunjukkan bahawa model Linear Model Regression menggunakan kaedah peningkatan Box Cox mengungguli MLP, KNN, HT, dan HAT. Dalam kajian ini, model hubungan dikembangkan menggunakan modul River perlombongan data tambahan dalam perisian sumber terbuka Python. Hasil keseluruhan menunjukkan bahawa model Linear Model Regresssion (Box Cox mendekati) dapat berhasil diterapkan dalam pilihan pembuatan untuk rantai bekalan semikonduktor dengan menggunakan BD yang dihasilkan.

ABSTRACT

The availability of Big Data (BD) is a critical factor that semiconductor makers can leverage to increase their forecast accuracy. Manufacturers can use this data to better forecast future consumer demand in semiconductor manufacturing and make best guesses about the quantities of each product variant to create via selective manufacturing based on calculated risk. The potential of five distinct machine learning models was investigated in this study: Linear Model Regression, Multi-layer Perceptron Regressor Model (MLP), K-Nearest Neighbour Regressor Model (KNN), Hoeffding Tree Regressor Model (HT), and Hoeffding Adaptive Tree Regressor (HAT). Three statistical metrics were used to assess the accuracy of constructed models: mean absolute error (MAE), root mean squared error (RMSE), and residual distribution. Meta, Imblearn, Expert, and Ensemble were employed in this experiment to enhance the machine learning technique. Along with the model, a drift detection method has been implemented such as ADWIN into our experiments to observe rapid changes in the generated BD. The comparison of research showed that the Linear Model Regression model using the Box-Cox improvement method outperformed MLP, KNN, HT, and HAT. In this study, the relationship model has been developed using the incremental data mining River's library in open source software Python. The overall results indicated that the Linear Model Regression model (Box-Cox approached) model could be successfully applied in the semiconductor supply chain by using the generated BD.

CHAPTER 1
INTRODUCTION

1.1 Overview of Data Stream Mining in Semiconductor Manufacturing

The semiconductor manufacturing industry is highly complex. The major processes in most semiconductor industries are in the following sequence: production of silicon wafers from pure silicon material, fabrication of integrated circuits onto the raw bare silicon wafers, assembly by putting the integrated circuit inside a package to form a ready-to-use product, and testing of the finished products. The number of process steps in wafer fabrication, front end-stage to semiconductor manufacturing is typically over 500, and each wafer is individually tracked because while wafer to wafer (WTW) variation can be incredibly small and elusive, it still impacts device yield (Munirathinam and Ramadoss 2016).

The semiconductor fabrication is highly automated, with the system generating large amounts of data and often on the order of a few terabytes per day, for instance, there are ~500 steps in semiconductor chip processing that generates terabytes of data daily (Munirathinam and Ramadoss 2015). Mining is a crucial method to extract and analyse useful information since the semiconductor fabrication process is a very complex manufacturing process composed of hundreds of steps, and interaction between different varieties can be difficult to fully understand, and data mining help to emphasize such relationships. Advanced statistical modelling techniques also help to analyse and build the relationship that explains such variation and its performance. This research studies and compares various stream data mining models using simulated data. The primary objective of the research is to look into suitable methods and programming platforms allowing stream data mining. Due to the exploratory nature of the work, the research is considered preliminary technical investigation, rather than industry application which will be addressed as future work.

1.2 Overview Project Background

The advancement of semiconductor integrated circuit (IC) in smaller technological nodes coupled with the complex module packaging enabled multi-chip integration in a single package with a wide variety of features. For instance, an IC can be marketed at different prices with options on needed features thus given birth to numerous variants. Furthermore, a customer will pay higher for an IC variant with faster transceiver speed and lower power where manufacturers can be shown in electrical testing. Subsequently, the manufacturer can command higher profits by binning out IC into different variants and selling better performance IC at a higher price while the lower performance IC at a more economical price and produced a wide range of marketing. This different variant of IC manufacturing is called selective manufacturing. This kind of strategy is allowed outsourcing companies to collaborate with other companies to produce and distribute a product. For example, a fabless company able to produce semiconductors and come out with different variants of semiconductors manufacturing. The manufacturing industry is currently in the midst of a data-driven revolution, which promises to transform traditional manufacturing facilities into highly optimized smart manufacturing facilities. These smart facilities are focused on creating manufacturing intelligence from real-time data to support accurate and timely decision-making that can have a positive impact across the entire organization. Therefore, manufacturing facilities must be able to manage the demands of the exponential increase in data production, as well as possessing the analytical techniques needed to extract meaning from these large datasets. More specifically, organizations must be able to work with BD technologies to meet the demands of smart manufacturing. The availability of BD enables the manufacturer to better predict the future demand outlook of the customer and do the best guesstimate quantities for each product variant to manufacture with selective manufacturing based on calculated risk.

1.3 Problem Statement

Advancement in semiconductor integrated circuit (IC) in smaller technological nodes coupled with complex module packaging enabled multi-chip integration in a single package with a wide variety of features (Mönch et al, 2018a). For example, a field-programmable gate array with unprecedented logic density hosts a wide variety of features such as embedded processors, DSP blocks, clocking, and high-speed serial. Such an IC can be marketed at different price points with options on needed features hence giving birth to numerous variants. Customers will pay higher for IC variants with faster transceiver speed and lower power where manufacturers can bin out during electrical testing. This allows manufacturers to command higher profits by binning out

IC into different variants, selling better performance IC at a higher price while lower performance IC at a more economical price and capturing broader market applications. Such a strategy to manufacture different variants out of a single complex IC is called selective manufacturing. The strategy is particularly important for semiconductor outsourcing, which allows fabless supply chains or a range of virtual enterprises in which different firms collaborate to produce and distribute a product (Mönch et al, 2018).

However, there are several challenges on this. Firstly, it is not easy for a manufacturer to control process variations to consistently produce a given percentage of higher performance variants out of an IC. In addition, yield may vary significantly over time, across facilities, and across different manufacturing technologies and products, especially the introduction of new products or processes into high-volume production (Mönch et al, 2018). Secondly, manufacturers cannot really predict future demand outlook of customers to decide which variant to produce more for cost optimization. Short product life cycles and different drivers of product demand limit which statistical forecasting approaches can be applied (Uzsoy et al., 2018). Customers might order more standard performance variants while manufacturers produce more low-performance variants. In addition, the lead time to manufacture an IC generally takes about 20 to 25 weeks so manufacturers must look ahead and estimate what to produce today for orders that are coming a few months later. This requires manufacturers to take a calculated risk when selective manufacturing is performed against specifics predictions from customer ordering patterns and market demand outlook.

The manufacturing industry currently undergoes a data-driven revolution to become highly optimized smart manufacturing facilities (O’Donovan et al., 2015). One key thing semiconductor manufacturer can rely on to improve their prediction accuracy is the availability of Big Data (BD). Big Data is a combination of structured, semi-structured, and unstructured data collected by organizations that can be mined for information and used in machine learning projects, predictive modelling, and other advanced analytics applications. The entire semiconductor supply chain which includes manufacturers themselves, vendors, and customers generates a huge amount of BD constantly every day. The information systems support business and manufacturing intelligence by storing increasingly BD (O’Donovan et al., 2015). Such data can be used by manufacturers to better predict the future demand outlook of customers and do best guesstimate quantities for each product variant to manufacture with selective manufacturing based on calculated risk. This research focuses on developing an operation and analytics model to predict and optimize selective manufacturing in the semiconductor supply chain based on the availability of BD. This research aims at constructing a relationship model to mathematically explain how leading indicators and BD contribute to selective manufacturing decision making and develop an algorithm to predict and optimize selective manufacturing options for the semiconductor supply chain.

1.4 Objectives

The research objective of this research is to examine the performance of various stream data analytics models onto simulated data with a structure similar to selective manufacturing.

1.5 Scope of Project

This project is based on the River library in Python and a simulated data stream. Several machine learning methods include Neural Network, Linear Model Regression, Hoeffding Tree Regressor, Hoeffding Adaptive Tree Regressor, and K-Nearest Neighbour. All the models from the algorithms have been extracted and are compared to which the model produced the best prediction by comparing the performance metrics. The research also implements several drift detection algorithms to track sudden changes in the generated data. Furthermore, the Meta, Imblearn, Expert, and Ensemble were used to develop an improved technique for the algorithm model, which was then tested. Then, compare the performance of the various models and select the one that exhibits the best MAE, RMSE, SMAPE, and residual distribution performance.

CHAPTER 2
LITERATURE REVIEW

2.1 Introduction of Data Stream Mining

Data stream mining is the process of analysing and extracting useful information from continuous and rapid data streams (Zhang et al. 2012). The process involves many technical areas such as classification, detection, and clustering. Today, huge volumes of sensory, transactional, and web data are continuously generated as streams of data. These change continuously with the updates of information where the forthcoming data is combined along with existing data. Streaming data can be considered as one of the main sources of what is called BD. More than half of the global population was already using email in 2019 with 3,93 billion users and the number is predicted to reach 4.37 billion users in 2023 (Altaylar 2020). These users would send out 347 billion emails per day by the end of 2023. Data stream mining is crucial for the big organization since they constantly generate large amounts of data. For example, a big organization like Google handles more than 3.5 billion searches on daily basis, NASA satellites generate around 4TB images and Walmart records more than 20 million transactions (Nguyen, Woon, and Ng 2015).

Three common challenges in BD are velocity, volume, and variety (Fong, Wong, and Vasilakos 2016). Velocity challenge means that the huge amount of data to be handled at an escalating high speed, Variety is a problem that makes data processing and integration difficult due to the data come from various sources and they are formatted differently and for the Volume problem, it means that the storing, processing and analysis over them both computational and archiving challenging. In contrast to stationary data in traditional cluster learning scenarios, streaming data can be extracted just a single time and is unbound in size. traditional data mining approaches are not suitable for streaming data (Finlay, Pears, and Connor 2014), as they results in unstable models with limited applicability for the construction of predictive models to determine to build outcomes in the advance of the build attempt.

BD transverse enterprise data processing pipelines in a streaming manner due to the ubiquitous use of sensors and network monitoring software. As a result, organizations are turning to process systems for data or event sources and wish to develop complicated analytics online (Ari et al., 2012). Furthermore, to handle and analyse the massive data that may have a high level of uncertainty, there is an additional difficulty in choosing a suitable systems architecture (Bodyanskiy et al., 2016). This was how they constructed a cascade neo-fuzzy system with a pooling of extended neo-fuzzy neurons. This method allows stationary and stochastic non-linear chaotic signals, such as data streams, to have both filtering and tracking capabilities in online mode.

2.1.1 Type of Data

In traditional data mining, this data is often referred to as "batch processing" data. In other words, all data is immediately available and stored in memory (Wares, Isaacs, and Elyan 2019). Batch-based processing involves the extraction of information from large amounts of data, and the outcome is expected once the entire process has been completed (Benjelloun et al. 2020). Each block is processed individually, and each block is separated by a period, whereas batch processing operations are frequently executed simultaneously and in sequential order. Its primary advantage is the ability to break down huge jobs into smaller pieces in order to maximize efficiency (Benjelloun et al. 2020). Because of this, Benjelloun et al. (2020) review the MapReduce paradigm, which is the most well-known processing paradigm for this type of processing. The following is the formal definition of the term MapReduce: When it comes to commercial applications, MapReduce is a programming model that may be applied to a wide variety of scenarios. It is intended for the processing of enormous amounts of data in parallel, and it accomplishes this by dividing the work into several separate jobs.

Unlike batch data, the data stream is an infinite series of data points identified either by time stamps or by an index. In data sources, they can also see data in an integer, categorical, graphical vector of data in hierarchical or unstructured formats in organized or structured data (PhridviRaj and GuruRao 2014). These data streams are either static or changing and are graded based on their core distribution. Particularly, data streams generate dynamically with unbound size (Alothali, Alashwal, and Harous 2019). Although batch data is different from the data stream, there is a certain technique that handles data stream as a batch which is called micro-batching. They (Shahrivari 2014), (Ksieniewicz and Zyblewski 2020) treated the stream as a sequence of small batches of data. The amount of data being generated is increasing, and streams are continuous and extremely fast. It is critical to approach this data differently since it demands rapid acquisition and processing. Stream processing (Benjelloun et al. 2020) is utilized when data needs to be analysed as soon as it is received, which is often the case. When this type of processing is not used when it is required, it can result in a variety of problems, including data value loss over time as a result of excessive latency.

2.1.2 Regression Method for Data Stream Mining

In context, data stream mining, method or technique is needed to effectively analyze, integrate, acquire and transform data, extract valuable patterns in real-time with only one scan and maintain the continuity of the process (Alothali et al., 2019). Generally, there are two method types that have been well known used, which are the classification and regression techniques. In these techniques, there are a hundred kinds of algorithms that used different kinds of methods or approaches. However, in this context, the objective is to specified this such a problem to be as precise as possible in approximating the mapping function (f), so that if fresh input data (x) is added to the dataset, the output variable (y) can be predicted (Friedman, 2012). Both classification and regression techniques are based on the concept of making predictions using known datasets (referred to as training datasets). However, because the output variables are both numerical and continuous, the regression is chosen as the experiment strategy. This part will discuss a review article on a few regression algorithms that have will be employed in this study and experimentation.

2.1.2(a) Linear Regression

Consider the regression issue (Su et al., 2012), which entails fitting a continuous response Y to a set of predictors X_1,..., X_p. By fitting a linear equation to observed data, linear regression seeks to model the relationship between two variables. One variable is thought of as an explanatory variable, while the other is thought of as a dependent variable. The simplest model type is linear regression, which represents the regression function as a linear collection of components. As follows is the definition of the linear model:

$$y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip} + \varepsilon_i$$

In matrix form,

$$y = X\beta + \varepsilon$$

where as

1. (Linearity) $\mu = [\mathrm{E}(y_i|x_i)]_{nx1} = X\beta$
2. (Independence) ε_i are independent of each other;
3. (Homoscedasticity), ε_i's have equal variance σ^2 ;
4. (Normality) ε_i's are normally distributed.

2.1.2(b) KNN Regressor

In KNN-based regression (Barrash, Shen, and Giannakis 2019), $\hat{f}$ is first generating a graph in which each node represents a data sample and is connected to k additional training samples that are "near" to x in some way. KNN regression is a non-parametric technique that approximates the relationship between independent variables and continuous outcomes intuitively by averaging observations in the same neighbourhood. The following is a simplified representation of the KNN regressor that is best for the linear Gaussian data model:

$$\hat{f}_{kNN}(X) = \frac{1}{k}\sum_{\mathrm{j}=1}^{k} \mathrm{y}_{(j)}(\mathrm{x})$$

where as:

1. Let $x_{(j)}(x)$ denote the jth closest training datum (neighbour) to x
2. Let $\mathrm{y}_{(j)}(\mathrm{x})$ be the y-value pertaining to $x_{(j)}(x)$

2.1.2(c) MLP Regressor

An artificial neural network (ANN) is a network made up of synthetic neurons or nodes that mimic biological neurons (Nassif, Ho, and Capretz 2013). The feed-forward method ANN layers are typically denoted by the terms input, hidden, and output. If no hidden layer exists, this sort of ANN is referred to as a perceptron. At least one hidden layer is present in an MLP, and each input vector is represented by a