DOKUZ EYLÜL UNIVERSITY

GRADUATE SCHOOL OF NATURAL AND APPLIED SCIENCES

DETECTING ANOMALIES IN PRODUCTION USING MACHINE LEARNING METHODS

by

Devrim Naz AKDAŞ

January, 2023

İZMİR

DETECTING ANOMALIES IN PRODUCTION USING MACHINE LEARNING METHODS

A Thesis Submitted to the
Graduate School of Natural and Applied Sciences of Dokuz Eylül University
In Partial Fulfillment of the Requirements for the Degree of
Master of Science in Computer Engineering

by
Devrim Naz AKDAŞ

January, 2023
İZMİR

M.SC THESIS EXAMINATION RESULT FORM

We have read the thesis entitled **"DETECTING ANOMALIES IN PRODUCTION USING MACHINE LEARNING METHODS"** completed by **DEVRİM NAZ AKDAŞ** under the supervision of **ASSOC**. **PROF. DR. DERYA BİRANT** and **ASSIST. PROF. DR. PELİN YILDIRIM TAŞER** and we certify that in our opinion it is fully adequate, in scope in quality, as a thesis for the degree of Master of Science.

Assoc. Prof. Dr. Derya Birant	Dr. Pelin Yıldırım Taşer
Supervisor	Co-Supervisor
Prof. Dr. Recep Alp Kut	Assist. Prof. Dr. Serhat Peker
(Jury Member)	(Jury Member)

Assoc.Prof. Dr. Vahid Akram

(Jury Member)

Prof. Dr. Okan FISTIKOĞLU

Director

Graduate School of Natural and Applied Sciences

ACKNOWLEDGEMENTS

To begin with, I would like to sincerely thank my supervisors, Assoc. Prof. Dr. Derya BİRANT and Assist. Prof. Dr. Pelin YILDIRIM TAŞER, for their great supervision, supports, and recommendations during the research and the writing of this thesis. The support and guidance which is they provided shed light on me throughout my thesis.

I am deeply grateful to my family and my fiancée for their unwavering support throughout my thesis work. They have been a constant source of encouragement during my master's education and my thesis writing. I could not have accomplished this without their help and support.

I would also like to express my appreciation to the company where I am employed, Proje Sistem Yazılım Limited Şirketi, for their support during my thesis work. Their understanding and flexibility have allowed me to focus on my studies and complete my thesis successfully. I am grateful for their confidence in me and their support. Also, I' am thankful to the Çelikiş Dişli Otomotiv San. ve Tic. A.Ş. which provided me the data that I used in my thesis.

Devrim Naz Akdaş

DETECTING ANOMALIES IN PRODUCTION USING MACHINE LEARNING METHODS

ABSTRACT

Discovering previously unknown anomalies that are rare and dramatically differ from the majority of the data is a critical need for the automotive industry. Rare Itemset Mining (RIM), one of the pattern-based methods, has been used for anomaly detection. However, several aspects still need to be explored, such as improving the mining process by identifying more targeted, valuable, and reliable rare itemsets.

Motivated by this fact, this study proposes a novel approach, named Ensemble of Rare Itemset Mining (ERIM), which investigates weak rare itemsets (WRIs) using different algorithms and aggregates these rules to obtain strong rare itemsets (SRIs). This study also combines four different RIM algorithms (Apriori Rare, Apriori Inverse, CORI, and RP-Growth) for the first time. The proposed ERIM approach applied to the automotive industry as a case study.

The ERIM approach was applied to a real-world dataset of gear manufacturing to identify anomalies in machine downtimes. The experimental results were evaluated based on the number and length of itemsets, and some examples were provided for illustration. According to the results, the ERIM approach gives more reliable common knowledge by jointly considering the relation between WRIs discovered by the base learners.

The experimental results showed that the proposed ERIM technique was successful in detecting anomalies. It is clear from the results that our method performed 43.37% better on average than the state-of-the-art methods on the same dataset. Based on its superiority, it is possible to say that it can be effectively used in future research projects.

Keywords: Anomaly detection, artificial intelligence, automotive industry, data mining, ensemble learning, and rare itemset mining.

ÜRETİMDEKİ ANOMALİLERİN MAKİNE ÖĞRENMESİ METOTLARI KULLANILARAK TESPİT EDİLMESİ

ÖZ

Nadir görülen ve verilerin çoğundan önemli ölçüde farklı olan ve önceden bilinmeyen anormallikleri keşfetmek otomotiv endüstrisi için kritik bir ihtiyaçtır. Model tabanlı yöntemlerden biri olan Seyrek Öge Seti Madenciliği (RIM), başarılı analiz sonuçları sağlaması nedeniyle bu çalışmada anomali tespiti için kullanılmıştır. Ancak, hedefe daha uygun, değerli ve güvenilir nadir öğe kümelerini belirleyerek madencilik sürecini iyileştirmek gibi bazı yönlerin hala keşfedilmesi gerekmektedir.

Bu yeni yaklaşım ile motive edilen bu çalışma, farklı algoritmalar kullanarak seyrek öğe setlerini (WRIs) araştıran ve güçlü öğe setleri (SRIs) elde etmek için bu kuralları bir araya getiren Seyrek Öğe Seti Madenciliği Topluluğu (ERIM) adlı yeni bir yaklaşım önermektedir. Bu çalışma aynı zamanda dört farklı Seyrek Öge Seti algoritmalarını (Apriori Rare, Apriori Inverse, CORI ve RP-Growth) ilk kez birleştirir. Önerilen ERIM yaklaşımı, bir vaka çalışması olarak otomotiv endüstrisine uygulanmıştır.

Yapılan deneylerde, ERIM, makine duruş sürelerindeki anormallikleri keşfetmek için gerçek dünyadaki bir dişli üretim veri setine uygulanmıştır. Deneysel sonuçlar, bazı örnekler de verilerek, öğe kümesi sayısı ve öğe kümesi uzunluğu açısından değerlendirilmiştir. Sonuçlar, önerilen ERIM yaklaşımının, temel öğrenenler tarafından keşfedilen seyrek öğe setleri arasındaki ilişkiyi birlikte ele alarak daha güvenilir ortak bilgi verdiğini göstermiştir.

Deneysel sonuçlar, önerilen ERIM tekniğinin anomalileri tespit etmede başarılı olduğunu göstermiştir. Yöntemimizin aynı veri setinde en son teknolojiye sahip yöntemlerden ortalama %43,37 daha iyi performans gösterdiği sonuçlardan açıkça görülmektedir. Önerilen yaklaşımın üstünlüğünden yola çıkarak gelecekteki araştırma projelerinde etkin bir şekilde kullanılabileceğini söylemek mümkündür.

Anahtar Kelimeler: Anomali tespiti, yapay zeka, otomotiv endüstrisi, veri madenciliği, topluluk öğrenimi ve nadir öğe kümesi madenciliği.

CONTENTS

Page

LIST OF FIGURES

Page

LIST OF TABLES

Page

CHAPTER ONE
INTRODUCTION

1.1 General

Nowadays, topics including machine learning, deep learning, and artificial intelligence have garnered significant attention and importance. Furthermore, with the increasing importance of data and big data, analyzing and processing these data has become a critical subject. Anomaly detection is the process of identifying observations that do not correspond to the anticipated, typical behavior (Chandola, Banerjee, & Kumar, 2009). Today, these inappropriate behavior patterns occur in many domains. Anomaly detection has numerous applications and is a topic of widespread research in various fields such as Cyber-Intrusion Detection, Machine Fault Detection, Fraud Detection, Medical Anomaly Detection, Industrial Damage Detection, Image Processing, Textual Anomaly Detection, and Sensor Networks (Quatrini, Costantino, Di Gravio, & Patriarca, 2020). The philosophy of Lean Manufacturing focuses on enhancing customer satisfaction by eliminating inefficiencies and maximizing the utilization of resources. It drives continuous improvement efforts to streamline processes, minimize waste, and higher performance, leading to the efficient production system. One of the ways to increase production efficiency and performance is to detect machine downtimes. Machine downtimes are unusual events in production. Analyzing the posture data and detecting these anomalies is important for the production industry. In the literature, pattern-based approaches are receiving increasing attention in the field of anomaly detection. Rare Itemset Mining (RIM) is a branch of data mining that focuses on revealing uncommon phenomena (known as rare itemset) and low-rank items from a large dataset. Rare Itemset Mining is the inverse variation of Frequent Itemset Mining (FIM). Usually, a large number of rare itemsets are generated by the RIM algorithms, and the challenge is to determine which itemsets are reliable and valuable for end-users. Our approach aims to solve this problem by taking advantage of the strengths of Ensemble Learning.

Ensemble Learning is an active subfield in machine learning in which a collection of separate learning models is combined, and the outputs from each model are

aggregated using a consensus function to generate a single final result (Yıldırım, Birant, & Birant, 2019). Although the ensemble technique was first suggested for supervised learning (classification), it has recently been adapted to unsupervised learning applications.

Therefore, this study leverages the advantages of all these methodologies and combines them to detection of anomalies in gear manufacturing data as a case study.

1.2 Purpose

The main motivation of this study to propose a new approach called Ensemble of Rare Itemset Mining (ERIM), which combines the discovery of rare itemsets from various algorithms and uses a voting mechanism to select the most common ones among them for the purpose of anomaly detection. This approach is proposed in response to the need for a reliable method for identifying anomalies. The goal behind the proposed approach is to attempt to make such weak rare itemsets (WRI) robust by converting them into strong rare itemsets (SRI) that indicate significantly important anomalies.

1.3 Main Contributions of the Thesis

The novelty and main contributions of this paper are highlighted as follows:

(i) It proposes a novel approach, named ERIM, that investigates rare weak itemsets using different algorithms and aggregates these rules to get strong (significantly important) rare itemsets.

(ii) It is the first attempt to combine RIM and ensemble learning methodologies.

(iii) It is the first study that combines four different RIM algorithms (Apriori Rare, Apriori Inverse, Correlated Rare Itemset (CORI), and Rare Pattern Growth (RP-Growth)) as base learners in an ensemble manner.

(iv) This study is also original in that it applies the ERIM approach to a real-world experimental dataset to detect rare itemsets (anomalies) in machine downtimes.

(v) It evaluates the experimental results regarding the number of rare itemsets and the length of rare itemsets by also giving samples.

(vi) The proposed method outperformed the state-of-the-art methods by 43.37% on average on the same dataset.

The proposed ERIM approach is a general methodology that can be applied to various fields. As a case study, here, it was used in the automotive industry for detecting anomalies in gear manufacturing downtime of earth-moving machinery. The automotive sector, one of the world's major industries in terms of revenue, includes a diverse variety of businesses and organizations engaged in design, manufacturing, development, and marketing. The automobile sector is under pressure to increase productivity because numerous businesses and organizations compete in global markets for the greatest profitability index and market share (Soltanali, Rohani, Tabasizadeh, Abbaspour-Fard, & Aditya, 2018). Downtime in manufacturing is a significant factor impeding production efficiency. Therefore, analyzing machine downtimes and detecting anomalies, if any, are critical for optimizing lean manufacturing operations. The experiments conducted on a real-world dataset showed that the proposed ERIM approach is an effective method for detecting anomalies.

1.4 Organization of the Thesis

The thesis consists of six chapters. The remainder of this thesis is structured as follows:

In the Chapter 2, the relevant literature about Rare Itemset Mining Studies and Machine Downtime Studies are reviewed. This chapter also presents the comparison of our study with the previous studies.

In the Chapter 3, general background information about Anomaly Detection, Rare Itemset Mining, and Ensemble Learning paradigms are described in detail.

In the Chapter 4, the proposed approach, which is named ERIM, is explained with its formal definition. This chapter also mentions the advantages of the novel approach.

In the Chapter 5, describes the real-world dataset containing gear manufacturing downtimes of earth-moving machinery and presents the experimental results with discussions.

In the Chapter 6, conclusions and future work are presented.

CHAPTER TWO
LITERATURE REVIEW

2.1 Review of Previous Studies

This chapter includes a comprehensive and comparative review of the literature on the topics that the thesis focuses on. In this thesis, Frequent Itemset Mining and Machine Downtime studies are considered together as a new approach for the first time. A literature review of previous studies on both these issues is included in this section.

2.1.1 Review of Rare Itemset Mining Studies

RIM technique has been studied by researchers in many areas, such as education (Luna, Romero, Romero, & Ventura, 2015; Weng, 2011), healthcare (Wulandari, Ou-Yang, & Wang, 2019;Borah & Nath, 2018; Reps & Aickelin, 2015; Ji et al., 2013; Shrivastava & Jotwani, 2020; Bouasker, Inoubli, Yahia, & Diallo, 2021), marketing (Jeyakarthic & Singaram, 2019), network (Nithya & Jayakumar, 2016), and text mining (Zhu, Wang, Wu, Hu, & Wang, 2016). In the study of (Bouasker, Inoubli, Yahia, & Diallo, 2021), the author aims to discover students who need extra help in learning using Fuzzy-Apriori, Fuzzy-Apriori-Rare, and Fuzzy Apriori Rare Itemsets Mining (FARIM) algorithms. Another RIM algorithm, called Fuzzy Recognition-Primed Decision (RPD), was proposed by Ji et al. for discovering adverse drug reaction signal pairs. In another study (Bouasker, Inoubli, Yahia, & Diallo, 2021), two novel frameworks Correlated Associations Mining (CORAM) and Distributed Correlated Associations Mining (DIST-CORAM) were proposed for uncovering rarely correlated association rules from pregnancy-associated breast cancer gene expressions. A detailed comparison of this study with the existing RIM studies is given in Table 2.1. From the reviewed literature, it is understood that the RIM method has never been applied before to the automotive sector.

Table 2.1 Comparison of this study with the previous RIM studies

Reference	Year	Objective	Method	Application Area	Ensemble Structure	Type of RI
(Luna et al., 2015)	2014	Identifying infrequent student behavior in Moodle repository	Apriori-Frequent, Apriori-Infrequent, Apriori-Inverse, Apriori-Rare, and Grammar Guided Genetic Programming (G3P)	Education	-	Weak
(Weng, 2011)	2011	Discovering students who need extra help in learning	Fuzzy-Apriori, Fuzzy-Apriori-Rare, and FARIM		-	Weak
(Wulandari et al., 2019)	2019	Identifying potential risk factors on stroke	Apriori-Rare and Rare-Unusual Association Rule (RUAR)	Healthcare	-	Weak
(Borah et al., 2018)	2018	Identifying risk factors for adverse diseases	FP-Growth, Fast Update (FUP), Fast Updated FP-Tree (FUFP-Tree), and Single Scan Pattern Tree (SSP-tree)		-	Weak
(Reps et. al, 2015)	2015	Refining adverse drug reaction signals	Apriori		-	Weak
(Ji et al., 2013)	2013	Discovering adverse drug reaction signal pairs	Fuzzy Recognition-Primed Decision (RPD)		-	Weak

Table 2.1 continues

(Shrivastava et al., 2020)	2020	Determining adverse drug diseases	Apriori-Inverse, Apriori-Rare, Relative Support Apriori Algorithm (RSAA), and Multiple supports Apriori (MsApriori)		-	Weak
(Bouasker et al., 2021)	2020	Discovering hidden relations among genes properties related to breast cancer	Correlated Associations Mining (CoRaM) and Distributed Correlated Associations Mining (DIST-CoRaM)		-	Weak
(Jeyakarthic et al., 2019)	2019	Forecasting client revenue	Apriori-Rare and Rare Itemsets Frequency Money (RIFM)	Marketing	-	Weak
(Nithya et al., 2016)	2016	Detecting network intrusion	Apriori and Multiple Minimum Support with Probability (MMSP)	Network	-	Weak
(Zhu et al., 2016)	2016	Detecting personalized and abnormal behaviors of Internet users	User-aware Rare Sequential Topic Patterns (URSTPs)	Text Mining	-	Weak
Our study		Detection anomalies in machine downtimes	Ensemble of rare itemset mining (ERIM)	Automotive	√	Strong

2.1.2 Review of Machine Downtime Studies

The machine learning-based studies that exist in the literature apply different tasks, such as classification (Mucchielli, Bhowmik, Ghosh, & Pakrashi, 2021; Garcés & Castrillón, 2017), regression (Nwanya, Udofia, & Ajayi, 2017; Wang, Liu, & Jin, 2019; Shafieezadeh, Desroches, Rix, & Werner, 2014), and time-series (Roosefert Mohan, Preetha Roselyn, Annie Uthra, Devaraj, & Umachandran, 2021) for detecting and reducing downtime in various fields. For example, Mucchielli et al. proposed a novel real-time downtime detection approach that implements Second Order Eigen-Perturbation using the k-Nearest Neighbor (SOEP-KNN) method. In the experiments, the proposed approach was tested on the Irish Wind Supervisory Control and Data Acquisition (SCADA) data and outperformed the traditional Artificial Neural Network (ANN) technique. In another study (Nwanya, Udofia, & Ajayi, 2017), the authors apply multiple regression for optimizing machine downtime in plastic manufacturing. Wang et al. also applied multiple regression analysis to evaluate downtimes and other equipment performance characteristics, such as cycle time, capacity, weight, and overall equipment effectiveness (OEE). Table 2.2 presents the comparison of this study with the previous downtime studies. It is clearly seen from this table that, to the best of our knowledge, a RIM paradigm has not been performed in the previous downtime studies until now. To bridge this gap, this study proposes a novel RIM approach (ERIM) that jointly discovers rare itemsets from different algorithms for detecting anomalies in gear manufacturing downtime of earth-moving machinery.

Table 2.2 Comparison of this study with the previous studies on machine downtime

Reference	**Year**	**Objective**	**Task**	**Algorithms**	**Application Area**
(Mucchielli et al.,2021)	2021	Detecting real-time wind turbine downtime	Classification	ANN and SOEP-KNN	Wind Turbine
(Garcés et al., 2017)	2017	Identifying and reducing downtime in a production system	Classification	Decision Tree (C4.5)	Production

Table 2.2 continues

(Nwanya et al., 2017)	2017	Optimizing machine downtime in the plastic manufacturing	Regression	Multiple Regression	Plastic Manufacturing
(Wang et al., 2019)	2019	Predicting schedule in response to machine breakdown	Regression	Support Vector Regression	Steel Manufacturing
(Shafieezadeh et al., 2014)	2013	Estimating downtime of geo-structures	Regression	Linear Regression and two Bilinear Regression	Geology
(Roosefert et al., 2021)	2021	Developing total productive maintenance for achieving zero downtime	Time-Series	Adaptive Autoregressive Integrated Moving Average (A-ARIMA)	Machine Industry
Our study		Anomaly detection in gear manufacturing downtime of earth-moving machinery	Ensemble of Rare Itemset Mining	Apriori Rare, Apriori Inverse, CORI, and RP-Growth	Automotive

There are many studies in the literature showing that ensemble learning gives more successful results than traditional (single) learning. While some of these studies (Jain, Semwal, & Kaushik, 2021; Gupta & Semwa, 2020; Semwal, Gupta, & Lalwani, 2021) combine deep learning methods, this study proposes an ensemble of rare itemset mining. In addition, while some studies use ensemble learning for a different purpose (i.e., human activity recognition (Jain et al., 2021; Semwal et al., 2021; Gupta et al., 2020)), our study focuses on outlier detection.

When Tables 2.1 and 2.2 are taken into account in general, this study differs from the previous studies in four different respects. First, it is the first study that proposes a novel approach, named ensemble of rare itemset mining (ERIM), that aims to discover

rare items obtained from different multiple RIM algorithms and combines these rules using a majority voting mechanism. Second, the previous downtime studies performed only regression, classification, and time-series tasks, while this study focuses on the application of RIM. Third, it is the first attempt to apply the RIM technique to the automotive industry for the purpose of anomaly detection. Lastly, unlike the previous studies, four different RIM algorithms (Apriori Rare, Apriori Inverse, CORI, and RP-Growth) were combined in an ensemble manner for the first time.

CHAPTER THREE
BACKGROUND INFORMATION

3.1 Anomaly Detection

Anomaly detection, an active study field in a variety of research communities, is the process of discovering data samples that significantly deviate from the norm (Böhmer & Rinderle-Ma, 2020). Anomaly detection is of capital importance as it provides the opportunity to identify important and critical anomalies that occur in the data in various application areas by converting them into significant information and preventing these inappropriate behaviors if it is possible. For example, a data anomaly on a credit card may indicate that credit card or identity thief, and also an anomaly in data from an industrial machine may indicate that the machine is about to break down or operating an incorrect product. These anomalies could be caused by either positive or negative events. For example, a rising temperature sensor in an engine could indicate an impending failure or change, while an unusually high number of clicks on a new product page could signal a sudden increase in demand. In both cases, these anomalies should be observed, and precautions should be taken. For this reason, anomaly detection makes it easier to take precautions against negative and positive situations that may occur in many areas or to plan possible actions that can be taken against these situations.

As in this study, it is critical to detect anomalies in the production area in the automotive industry. Detection of positive or negative changes in the production area reveals a work that is likely to be groundbreaking in the industry, as it increases production efficiency and leads to taking care of the situations that may occur.

3.1.1 Types of Anomalies

Anomalies are unusual points or patterns in specific data. The term anomaly is also referred to interchangeably with an outlier. One of the important points in the detection of anomalies is that the type of an anomaly which is observed in the data. An anomaly can be categorized in the following ways (Chandola et al., 2009).

• **Point anomalies:** Point anomalies are individual instances in a dataset that differ from the other instances with respect to their attributes. This type of anomaly is the most common type of anomaly. Many studies in the literature have focused on detecting such anomalies.

• **Contextual anomalies**: When a data point is unusual within a specific context, it is called a contextual anomaly. Without context, all data points may appear normal. Outliers of this type are common in time series data, as these datasets consist of records of specific quantities over a period of time. For instance, a high temperature in the summer is not considered anomalous. However, if the same temperature is observed in the winter, it would be considered an anomaly. In some cases, defining a context is straightforward, making it suitable to use a contextual anomaly detection technique. In other cases, defining a context is challenging and makes the use of such techniques difficult.

• **Collective anomalies**: When a group of data points within a larger set is unusual in comparison to the rest of the data, those values are referred to as collective anomalies. This subset of anomalies may not show an anomalous pattern by themselves, but when looking at the whole dataset, this subset is defined as an anomaly.

3.1.2 Representations of Anomalies

To detect anomalies, the data in the dataset must be labeled as normal or anomalous. This correctly labeled data is difficult to find, of course. Moreover, anomaly behavior is often dynamic in nature; for example, new types of anomalies that are not labeled can appear in the training data (Chandola et al., 2009). Generally, anomaly labeling is done manually by experts. Therefore, this process is time-consuming and difficult to implement. Based on these labeling, anomaly detection studies can be classified into 3 categories: Supervised Anomaly Detection, Unsupervised Anomaly Detection, and Semi-Supervised Anomaly Detection (Ramachandran & Sangaiah, 2018) .

In Supervised Anomaly Detection, the model is trained with the labeled input data as shown in the Figure 3.1 and then predictions are made based on this model in future instances. The supervised learning algorithm analyzes training data and generates an inference model for predicting new samples. The resulting detector is then used to assign class labels as an anomaly or normal to the testing instances.

In Unsupervised Anomaly Detection, Figure 3.1 shows that there is no labeled training data. There is also no distinction between a training and a test dataset. Therefore, it can be applied in more fields and data.

In Semi-Supervised Anomaly Detection, it uses training and test datasets, whereas training data only consists of normal data without any anomalies as shown in the Figure 3.1. The fundamental concept is that a model of the "normal" class is learned, and anomalies can subsequently be identified by their deviation from this model.

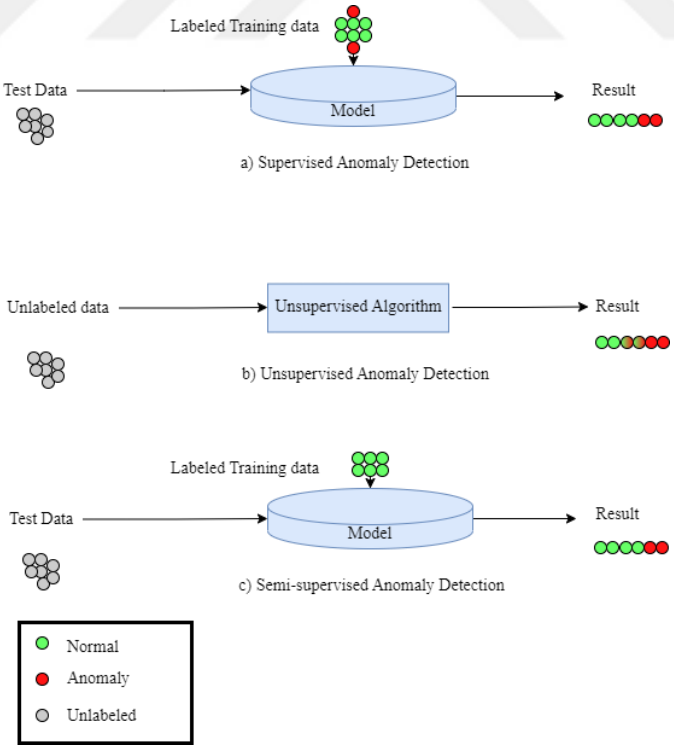

Figure 3.1 Anomaly detection types

3.2 Frequent Itemset Mining (FIM)

Frequent itemsets are patterns that occur with a frequency greater than or equal to a user-defined threshold in a dataset (Han, Cheng, Xin, & Yan, 2007; Abed, Abdelaal, Al-Shayeji, & Ahmad, 2020; Deng, 2013). For instance, a frequent itemset is a collection of items, such as bread and cheese, that frequently appear together in a transaction dataset.

Discovering frequent itemsets has emerged as a way to mine associations, correlations, and a variety of other interesting relations between data. Also, it aids in the indexing, categorization, and clustering of data, as well as other data mining techniques. Consequently, frequent itemset mining (FIM) has emerged as a prominent field of study and research topic within the realm of data mining.

The FIM is one of the well-known and widely-used data mining techniques, which was introduced by Agrawal (Agrawal, Imieliński, & Swami, 1993) for market basket analysis. It conducts research on the purchasing behavior of customers by identifying product sets that are commonly purchased together. Although it was originally intended for market basket analysis, it has commonly been preferred in a broad range of application areas, such as recommendation systems, healthcare, transportation, business, and multimedia. The FIM is the process of extracting any existing frequent itemsets whose frequency is greater than or equal to a certain threshold from the dataset (Borgelt, 2012).

While most of the mining studies in the literature have concentrated on frequent itemsets, rare ones can be more important in many practical scenarios (i.e., detecting network intrusions, discovering adverse drug reactions, etc.). Because of this reason, the rare itemset mining (RIM) technique was introduced in the literature.

3.3 Rare Itemset Mining (RIM)

In certain cases, it may be worthwhile to investigate rare itemsets, that is, itemsets that do not often occur in the dataset (infrequent itemsets) (Darrab, Broneske, & Saake,

2021). These relate to unanticipated events, which may contradict domain-specific assumptions. Thus, rare itemsets are associated with exceptions and may include information of considerable interest to specialists in fields like education, healthcare, and so on. In this study, we applied RIM in the automotive industry. Rare items are significant for detecting abnormal states since RIM discovers unusual events.

RIM seeks rare correlations between a group of objects in a large dataset. RIM is converse to FIM, as it tries to find rare sets that exist in a dataset, rather than frequent sets. An itemset *R* is referred to as a rare itemset if its support value over the dataset *D* is lower than the minimum support (Sup_{min}) threshold, as shown in Equation (3.1).

$$Sup_D(R) < Sup_{min} \tag{3.1}$$

In this study, the most common RIM algorithms (Apriori Rare, Apriori Inverse, CORI, and RP-Growth) are used to construct the ensemble structure of the proposed ERIM approach. After that, each algorithm in this approach discovers some rare itemsets, and the majority votes of these itemsets are chosen as final rare itemsets.

3.3.1 Apriori Rare

Apriori Rare, an Apriori-based algorithm, aims to discover minimal rare itemsets in a large database (Szathmary, Napoli, & Valtchev, 2007). A rare itemset is called a minimal rare itemset if it is not a frequent itemset, but all its proper subsets are frequent. A potential constraint of this research is the need to keep all rare itemsets, which might be rather costly in terms of storage capacity. Additionally, discovering rare association rules from all rare itemsets causes a massive collection of new association rules.

3.3.2 Apriori Inverse

Apriori Inverse is also a modified version of the Apriori algorithm that mines perfectly rare itemsets (also known as perfectly sporadic itemset) by disregarding all

candidate itemsets with support value greater than a certain threshold (Koh & Rountree, 2005). A perfectly rare itemset is not a frequent itemset, and all its proper subsets are also infrequent. The Apriori Inverse algorithm requires two parameters: minsup (minimum support) and maxsup (maximum support). The algorithm discovers perfectly sporadic itemsets whose support values are lower than maxsup and higher than minsup thresholds. The Apriori Inverse algorithm is more efficient than the Apriori Rare algorithm because it finds perfectly rare itemsets without creating all the itemsets that are unnecessarily frequent.

3.3.3 Correlation Rare Itemset (CORI)

The CORI algorithm, an extension of the ECLAT, aims to find rarely correlated itemsets (Bouasker & Yahia, 2015). A rarely correlated itemset is an itemset that is rare (its support value is higher than the user-defined minsup threshold) and correlated (its bond value is higher than the user-defined minbond threshold) in the transactional database. The bond of an itemset refers to the total number of transactions that include it divided by the total number of transactions that contain any of its items. The bond is a value in the range between 0 and 1. While a high bond value shows a highly correlated itemset, a low bound value means that the itemset is slightly correlated.

3.3.4 RP-Growth

The RP-Growth algorithm is a variation of the FP-Growth, which generates rare itemsets via a divide-and-conquer strategy (Tsang, Koh, & Dobbie, 2011). A rare itemset is one that occurs inside the range specified by the minraresup and minsup user-defined criteria. The algorithm builds a conditional tree for each rare item throughout the mining process in order to generate rare itemsets. The RP-Growth algorithm's primary benefits are its speed and memory efficiency.

3.4 Ensemble Learning

Ensemble learning is a technique that involves combining the predictions of multiple learning algorithms to create a stronger model. This approach typically

involves extracting features from a variety of data projections and then using a combination of voting mechanisms to combine the results of the individual models. Ensemble learning can often achieve better performance than any of the constituent algorithms could achieve on their own (Dong, Yu, Cao, Shi, & Ma, 2019; Dongdong, et al., 2021). The ensemble-learning strategy develops a strong output from a collection of individual learners. Numerous ensemble-based researches (Li & Chen, 2020; Chicco & Jurman, 2021) have shown that ensemble learners succeed more than traditional individual learners. Ensemble methods are commonly classified into four types in the literature: bagging, boosting, stacking, and voting. In this study, the voting technique was preferred for constructing the ensemble system.

The voting technique in ensemble learning aims to merge the individual outputs from each learner to get the final strong one. It does not need homogeneous base learners, so different algorithms can be used to generate an ensemble structure in this method. Majority voting is one of the simplest and most often utilized combination rules in learner ensembles (Onan, Korukoğlu, & Bulut, 2016). In that mechanism, each algorithm has an equal vote of 1, and the output with the highest number of votes is selected as the ensemble's output.

CHAPTER FOUR
PROPOSED APPROACH

4.1 Proposed Approach: Ensemble of Rare Itemset Mining (ERIM)

Analyzing anomalies is of great importance for manufacturing companies to enhance operational efficiency and lean manufacturing processes. RIM has gained popularity in recent years because of its versatility in applications such as anomaly detection, uncommon disease discovery, intrusion detection, and detecting seldom-bought items. For this reason, this study aims to improve the mining process by identifying more targeted, reliable, and valuable rare itemsets.

Meanwhile, the ensemble learning approach aims to employ multiple learners and then combine them via some strategies to get a consensus solution. It outperforms single individual algorithms in terms of accuracy, reliability, stability, and robustness because they entirely use the information offered by the learning members. This paper proposes a novel approach, called Ensemble of Rare Itemset Mining (ERIM), that constructs globally strong rare itemsets (SRIs) from weak rare itemsets (WRIs) locally discovered by different algorithms. In this study, the ERIM jointly discovers WRIs and chooses the majority vote of them (SRI) for detecting anomalies. Although the proposed approach in this study has been applied to the automotive industry, it is a general approach that can be applied to any field that aims to increase the efficiency of knowledge discovery systems.

Figure 4.1 shows the general overview of the proposed ERIM approach. First, the data is gathered from related sources. In the following phase, the dataset undergoes various data preprocessing steps, such as feature selection and feature extraction, to prepare it for further analysis. Therefore, the experimental data will be suitable for applying the ERIM approach. Then, a set of RIM algorithms are implemented on the dataset for discovering WRIs. Finally, the obtained WRIs from each individual RIM algorithm are combined, and the majority vote of these itemsets is chosen as SRIs.

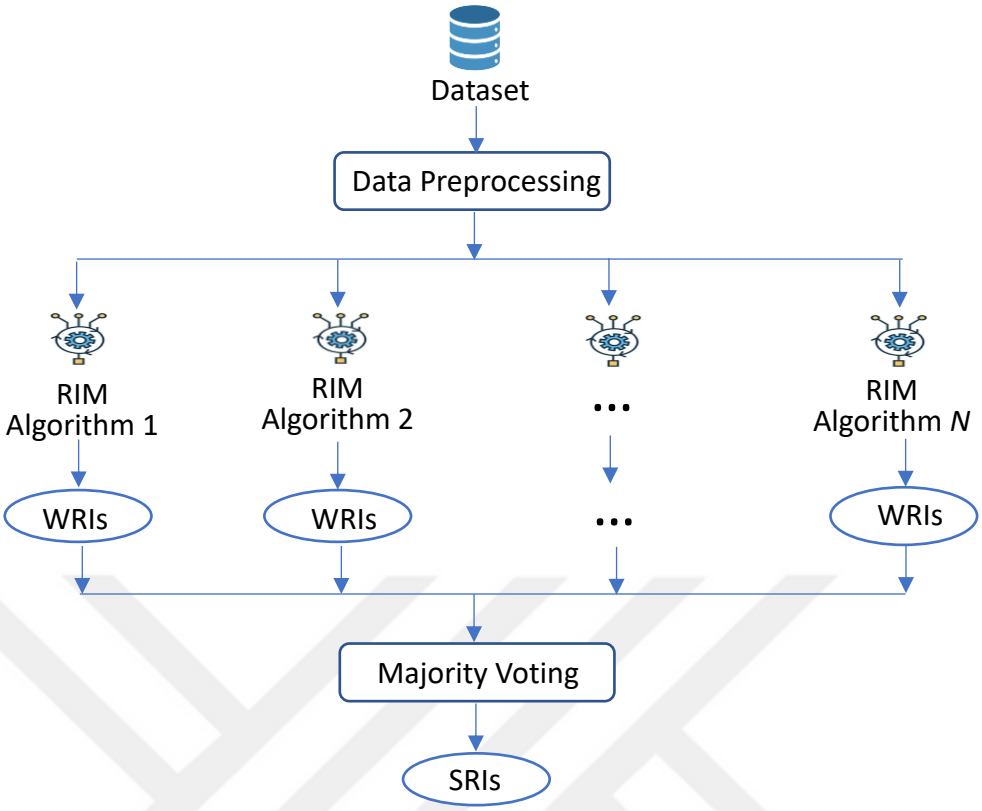

Figure 4.1 The general overview of the proposed ERIM approach

Table 4.1 shows an example to illustrate how the proposed ERIM approach determines SRIs from the WRIs when four different algorithms are used *(n=4)*. First, rare itemsets are discovered by each algorithm individually. After that, each candidate itemset is checked whether it is strong to be on the final list. An itemset is considered as strong if it appears more than *n/2* times in the outputs of algorithms. For example, the itemsets I1, I2, I3, and I5 are selected as a result of majority voting. However, the itemset I4 is filtered out because of occurring in less than *n/2* outputs.

Table 4.1 Example illustrating the proposed ERIM approach

Weak Rare Itemsets				**Strong Itemset**
Algorithm 1	**Algorithm 2**	**Algorithm 3**	**Algorithm 4**	
I_1	I_1	-	I_1	I_1
-	I_2	I_2	-	I_2
I_3	I_3	I_3	I_3	I_3
-	-	I_4	-	-
I_5	I_5	-	I_5	I_5

4.2 Formal Definition

Formally speaking, let an item base $B = \{i_1, i_2, \dots, i_n\}$ be a set of distinct items, and a database $D = \{t_1, t_2, \dots, t_m\}$ of transaction sets, where each transaction $t \subseteq B$. Typically, the item base B is implicitly defined as the union of all transactions t, i.e., $B = \cup_{k=1}^{m} t_k$. The itemset $I \subseteq B$'s cover $K_D(I) = \{k \in \{1, 2, \dots, m\} \mid I \subseteq t_k\}$ identifies the transactions it contains. The support $Sup_T(I)$ of I refers to the frequency of transactions that include it. An itemset I is a frequent iff $Sup_D(I) \geq Sup_{min}$,where a user-defined minimum support $Sup_{min} \in$ N. The FIM aims to discover all frequent itemsets $Freq_D(Sup_{min}) = \{I \subseteq B \mid Sup_D(I) \geq Sup_{min}\}$ in the database D. An itemset R is referred to as a rare itemset if its support value over the dataset D is lower than the minimum support (Sup_{min}) threshold.

The ERIM approach proposes two novel concepts: weak rare itemset (WRI) and strong rare itemset (SRI).

Definition 1 (Weak Rare Itemset) A weak rare itemset WRI is a rare itemset discovered using a particular RIM algorithm a from the database $D = \{t_1, t_2, \dots, t_m\}$ of transaction sets, where each transaction $t \subseteq B$ and its support value $Sup_D(WRI)$ is lower than the user-defined minimum support (Sup_{min}) threshold, as shown in Equation (4.1).

$$Sup_D(WRI_a) < Sup_{min} \tag{4.1}$$

where A is a set of algorithms such as $A = \{a_1, a_2, \dots, a_n\}$.

Definition 2 (Strong Rare Itemset) A strong rare itemset SRI is a rare itemset that occurs in more than half of the $WRIs$ obtained from n RIM algorithms $\{A_i\}_{i=1}^{n}$. An SRI is produced for each WRI detected by at least $n/2$ base algorithms in the ensemble structure using the global minimum support values gathered on the multiple algorithms where the itemset is determined to be rare. The SRI is defined as follows (Equation (4.2)):

$$Sup_D(SRI_A) < min_{a \in A}(Sup_{min})\ if\ Sup_D(WRI_a) \geq n/2 \quad (4.2)$$

4.3 Algorithm

Algorithm 1 presents the pseudocode of the proposed ERIM approach. First, each RIM algorithm in the ensemble structure discovers WRIs, whose support values are lower than the user-defined minimum support (Sup_{min}) threshold, from the same dataset D. Then, in the aggregation step, the proposed approach combines WRIs using a majority voting technique, which means itemsets are discovered by at least half of the RIM algorithms. In this way, WRIs become globally SRIs, which indicate more robust and reliable anomalies.

In the worst-case scenario, the time complexity of the proposed ERIM technique is defined as $O(T(m)*n + n^2 * l)$, where n is the number of RIM algorithms, l is the number of WRIs, m is the number of records, and T is the execution time of the used RIM algorithm. The experimental execution time of the ERIM approach varies according to the applied algorithms' execution times.

Algorithm: Ensemble of Rare Itemset Mining (ERIM)
Inputs:
D: the dataset consists of record sets such that $R = \{r_1, r_2, \dots, r_m\}$
m: the dataset D's record number
A: a set of RIM algorithms $\{A_i\}_{i=1}^{n}$
n: the number of RIM algorithms in the ensemble structure
Sup_{min}: the minimum support threshold
Output:
SRI: strong rare itemsets
Begin Algorithm:
//Discovering WRIs
for *i=1* **to** *n* **do**
  for *j=1* **to** *m* **do**
    $WRI_i^j = A_i\,(D_j, Sup_{min})$
    $WRI_i = WRI_i \cup WRI_i^j$
  end for
end for
//Aggregating WRIs to get SRIs
for *i=1* **to** *n* **do**
  for *j=1* **to** $WRI_i.Length$ **do**

```
        itemset = WRI_i
        count = 0
        for k=i+1 to n do
            if (WRI_k.Contains(itemset)) then
                count++;
            end if
        end for
        if (count >= n/2)
            if (!SRI.Contains(itemset)) then
                SRI.Add(itemset)
            end if
        end if
    end for
end for
End Algorithm
```

Figure 4.2 The pseudocode of the proposed ERIM approach

4.4 Advantages of ERIM

This paper introduces a new concept: "Ensemble of Rare Itemset Mining" and demonstrates its efficiency on a real-world dataset. The proposed approach discovers robust rare itemsets by combining the local sets obtained by different algorithms. It has many advantages, including leveraging the strengths of ensemble learning, easy implementation, suitability for parallel processing, and scalability.

- *Ensemble-based model:* The standard RIM algorithms have been used independently to discover rare itemsets from an entire dataset. Usually, a large number of rare itemsets are generated by a RIM algorithm, and the challenge is to determine which itemsets are reliable and valuable for end-users. The itemset obtained by a single algorithm can be weak to indicate uncommon items, events, or observations. A single algorithm can fail to discover a meaningful rare itemset. In contrast, our approach jointly discovers rare itemsets under the concept of ensemble learning. Therefore, it is helpful in discovering common and reliable information.

- *Easy implementation:* One of the key features of our approach is that the standard RIM algorithms do not require any modification in the mining phase. Therefore, the proposed ERIM approach can be implemented easily.

- *Running in parallel:* Another advantage of the proposed ERIM approach is that it can be executed in parallel. Discovering weak rare itemsets can be parallelized and distributed on several nodes in a grid platform, i.e., one node for each algorithm. In this way, ERIM can perform parallel processing by executing different algorithms at the same time. Therefore, significant gains may be achieved by using multiple-core processors.

- *Scalability:* One of the main advantages of the proposed ERIM approach is that a new algorithm can be easily added to the system without any modification. The results obtained by other algorithms remain the same.

CHAPTER FIVE
EXPERIMENTAL STUDIES

5.1 Experimental Settings

The proposed ERIM approach was applied to a real-world dataset, which consists of gear manufacturing downtime records, for discovering SRIs. These SRIs refer to the detected robust and reliable anomalies in gear manufacturing downtimes of earth-moving machinery. This proposed method used Apriori Rare, Apriori Inverse, CORI, and RP-Growth algorithms as base learners of the ensemble structure. The application was implemented utilizing an open-source data mining library written in the Java programming language, specialized in pattern mining (Fournier-Viger, et al., 2016). The experiments were conducted on a personal computer equipped with an Intel Core i5-7200U CPU running at 3.1 GHz and 8 GB of RAM. The experiments were evaluated regarding the number of itemsets, and the length of itemsets by also giving sample itemsets. The results obtained in this research are reported in tables and graphs.

5.2 Dataset Description

This research uses a real-world dataset that contains records of downtimes in the gear manufacturing of earth-moving equipment collected from an Izmir facility. The gear manufacturing downtime dataset contains 11,040 instances and eleven attributes informing machine downtimes.

5.3 Data Preprocessing

The dataset was preprocessed using feature selection, feature extraction, and data discretization techniques before implementing the ERIM approach.

- **Feature Selection:** The redundant and irrelevant attributes of the used dataset, such as downtime record ID, company name, operator name, operator surname, etc., were removed because they do not affect the knowledge discovery process.

- **Feature Extraction:** The new attributes, named month and day, were derived from the date attribute due to give useful information about downtimes.

- **Data Discretization:** Because the RIM algorithms can be executed on only categorical data, time and total production attributes, which have numerical values, were discretized into five levels. Equal Width Discretization, also known as rectangular binning, was used in this research as a method of data discretization where the data is divided into equal intervals or bins of the same size.

The attributes of the used dataset with their categories are given in Table 5.1.

Table 5.1 The attributes and their categories of the experimental dataset

Attributes	**Categorical values**
Month	January, February, March, April, May, June, July, August, September, October, November, December
Day	Monday, Tuesday, Wednesday, Thursday, Friday, Saturday, Sunday
WorkCenterCode	IM01, IM02, IM03, IM04, IM05, IM06, IM07, IM08, IM09, IM10, IM11, IM12, IM13, IM14, IM15, IM16, IM17, IM18, IM19
Downtime	Undefined, Failure / Maintenance, Launch, Initial adjustment, Settings correction, Changing tool holder, Changing cutting tool, Drill sharpening, Warming the bench, Assigning another task, Carrying material, Waiting for material, Personal need, Cleaning, No operator, Setting, Emending, Waiting for approval, Sampling
DowntimeGroupName	Breakdown, Quality, Personnel, Planned, Workbench preparation, Other
OperationName	CNC turning, Hard turning, Gear cutting, Finishing and grinding
DowntimeType	Planned, Unplanned
Department	CNC, Gear, Finishing
WorkCenterGroupCode	IMG110, IMG130, IMG131, IMG134, IMG141
Time	[0-3) = Very low, [3-15) = Low, [15-30) = Medium, [30-50) = High, [50-10000) = Very high
TotalProduction	[0-30) = Very low, [30-80) = Low, [80-130) = Medium, [130-200) = High, [200-50000) = Very high

5.4 Experimental Results of Individual RIM Algorithms

In this study, two different experiments were performed on the gear manufacturing downtime dataset for analyzing the followings:

(i) the maximum memory usage of the algorithms on different minimum support or maximum support values,

(ii) the execution time performances of the algorithms on different minimum support or maximum support values

In the following sections, the experimental results for each algorithm are given in detail.

5.4.1 Results of Apriori Rare Algorithm

In the first experiment, the Apriori Rare algorithm was applied with different minimum support threshold levels (%) ranging from 10 to 40 in increments of 10 for evaluating the memory usage of the algorithm. Figure 5.1 shows that the higher the minimum support values, the higher the memory usage obtained by the Apriori Rare algorithm. It was observed from the results that the minimum support value significantly affected memory usage.

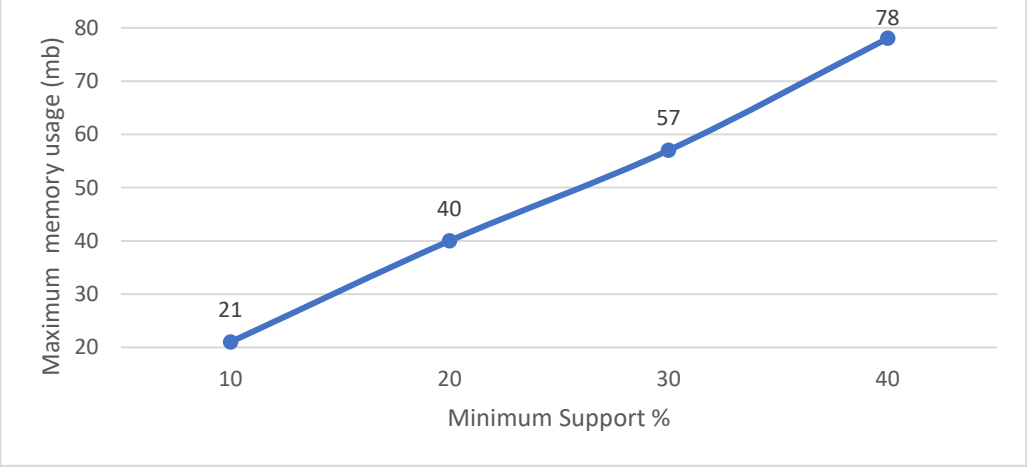

Figure 5.1 The maximum memory usage performance of the Apriori Rare algorithm with different minimum support threshold

In the second experiment evaluated execution times of the Apriori Rare algorithm on the dataset with various minimum support threshold levels (%) ranging from 10 to 50 in increments of 5. The obtained execution times are given in Figure 5.2 in milliseconds (ms).

Figure 5.2 shows that the higher minimum support values give lower execution times on average. In other words, as the minimum support value increases, the execution time decreases due to the decrease in the number of rules found by the algorithm.

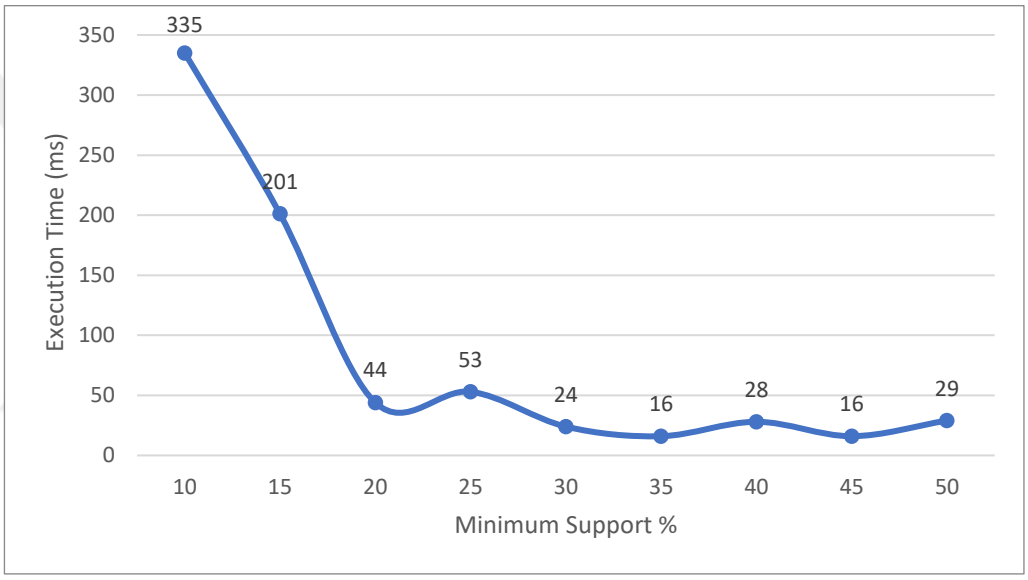

Figure 5.2 The execution time performance of the Apriori Rare algorithm with different minimum support thresholds

5.4.2 Results of Apriori Inverse Algorithm

The first experiment evaluated maximum memory usage in terms of megabit on the dataset with a constant minimum support threshold value (0.1%) and different maximum support threshold levels (%) ranging from 1 to 4 in increments of 1 for the Apriori Inverse algorithm. As shown in Figure 5.3 the Apriori Inverse algorithm provides efficient memory usage with acceptable levels.

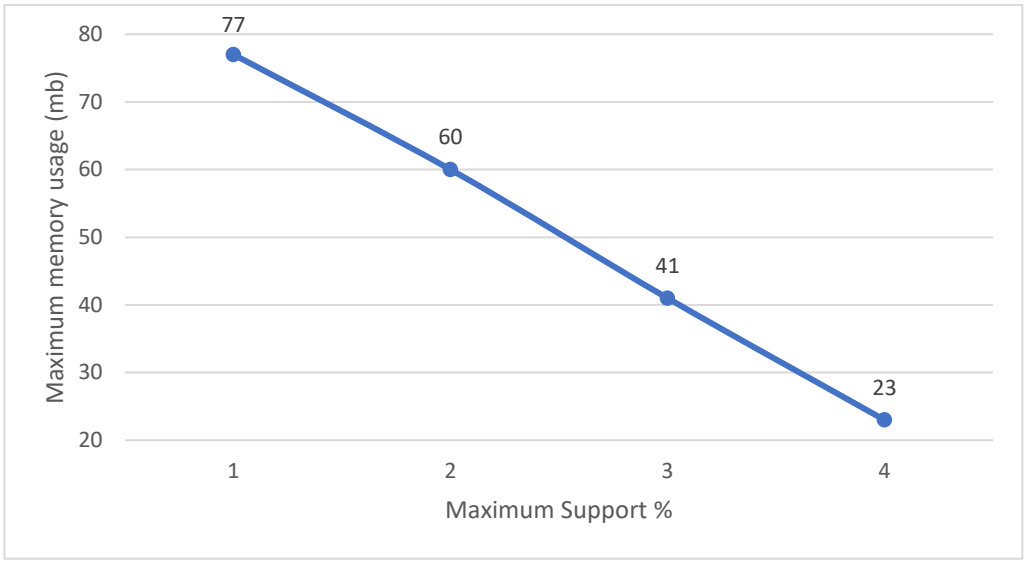

Figure 5.3 The maximum memory usage performance of the Apriori Inverse algorithm with different maximum support threshold

The second experiment evaluated execution times of the Apriori Inverse algorithm on the dataset with a constant minimum support threshold value (0.1%) and different maximum support threshold levels (%) ranging from 1 to 9 in increments of 1.

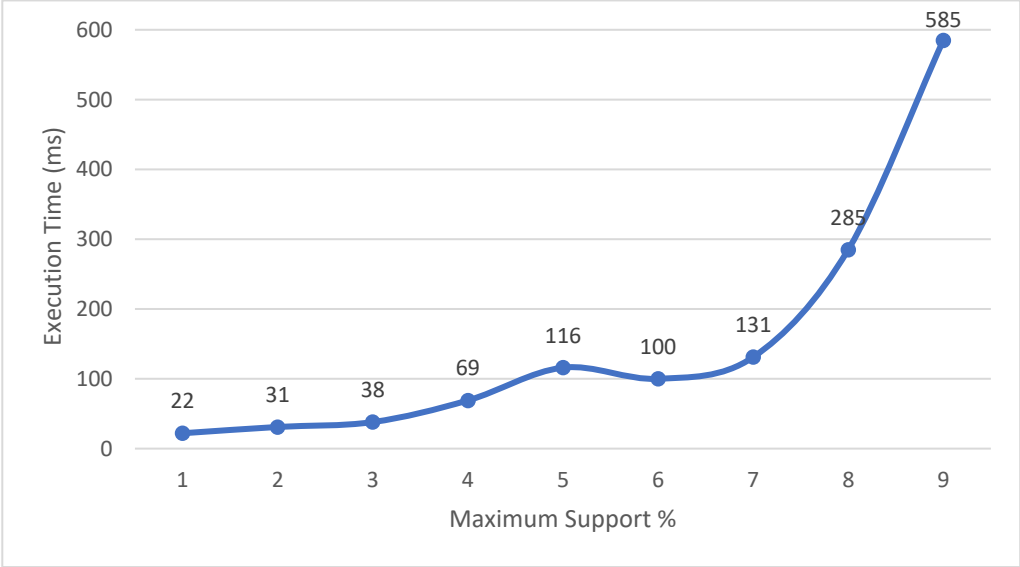

Figure 5.4 The execution time performance of the Apriori Inverse algorithm with different maximum support thresholds

The results are presented as a graph in Figure 5.4 in milliseconds (ms). This figure indicates that the execution time of the Apriori Inverse algorithm increases almost linearly as the maximum support value increases.

5.4.3 Results of CORI Algorithm

The first experiment evaluated maximum memory usage in terms of megabit on the dataset with a constant minimum bound threshold value (90%) and different maximum support threshold levels (%) ranging from 2 to 8 in increments of 2 for the CORI algorithm. The Figure 5.5 presents the relation between maximum support values and the memory usage of the algorithm. According to the graph, as the maximum support value increases, the memory usage decreases linearly.

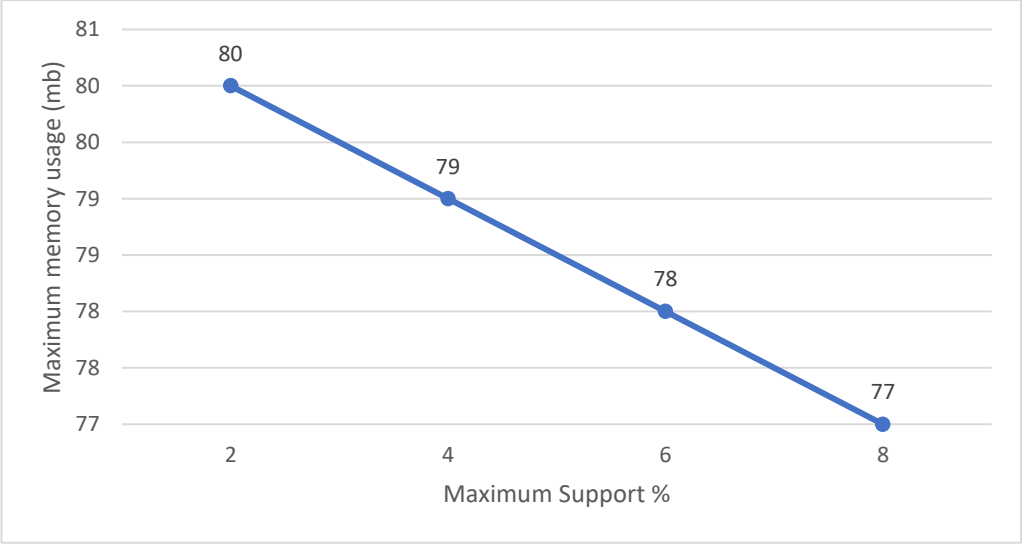

Figure 5.5 The maximum memory usage performance of the CORI algorithm with different maximum support threshold

The second experiment evaluated execution times of the CORI algorithm on the dataset a constant minimum bound threshold value (90%) and different maximum support threshold levels (%) ranging from 2 to 9 in increments of 1. The obtained execution times are given in Figure 5.6 in milliseconds (ms).

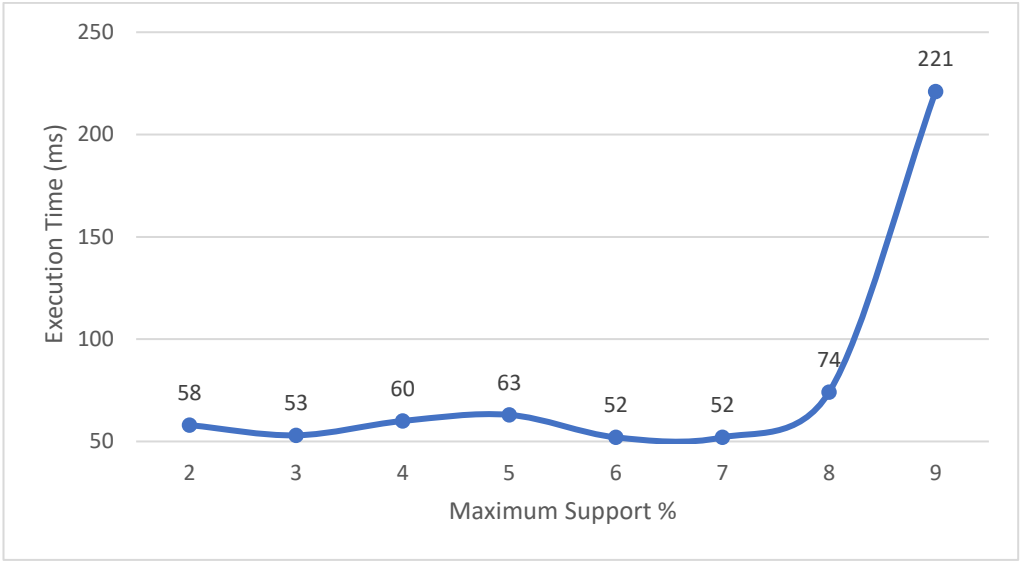

Figure 5.6 The execution time performance of the CORI algorithm with different maximum support thresholds

5.4.4 Results of RP-Growth Algorithm

In the first experiment, the maximum memory usage of the RP-Growth algorithm was calculated in terms of megabit (mb) with a constant minimum support threshold value (90%) and various minimum rare support threshold levels (%) ranging from 15 to 45 in increments of 10. The results are presented as a graph in Figure 5.7.

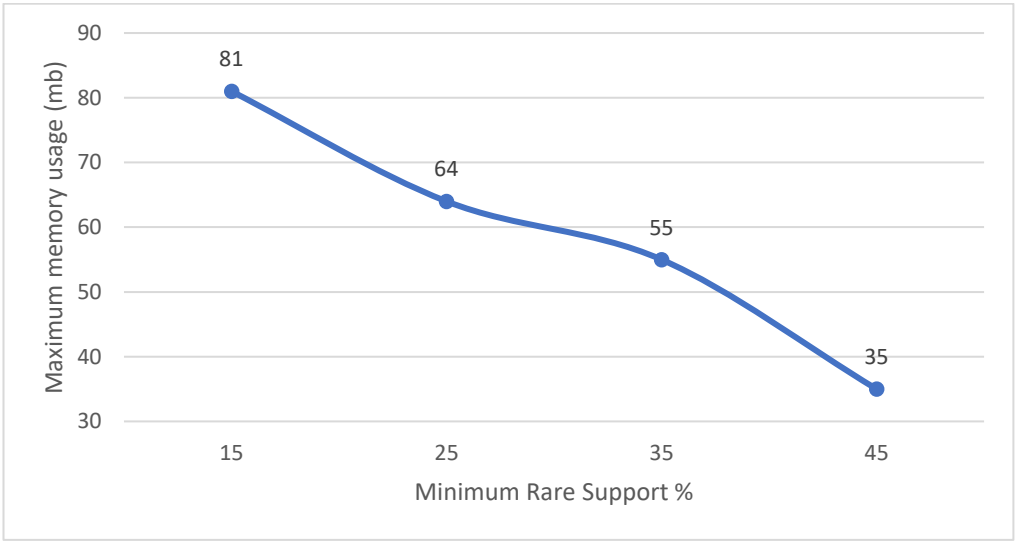

Figure 5.7 The maximum memory usage performance of the RP-Growth algorithm with different minimum rare support threshold

It is obviously seen from this graph which is given in Figure 5.7, The RP-Growth algorithm provides efficient memory usage with acceptable levels (Akdaş, Birant, & Yıldırım Taşer, 2023).

The second experiment evaluated execution times of the RP-Growth algorithm on the dataset with a constant minimum support threshold value (90%) and various minimum rare support threshold levels (%) ranging from 15 to 55 in increments of 5. The obtained execution times are given in Figure 5.8 in milliseconds (ms). The varying minimum rare support values significantly affect the execution times of the experiments. This figure indicates that the execution time of the RP-Growth algorithm increases almost linearly as the minimum rare support threshold value increases.

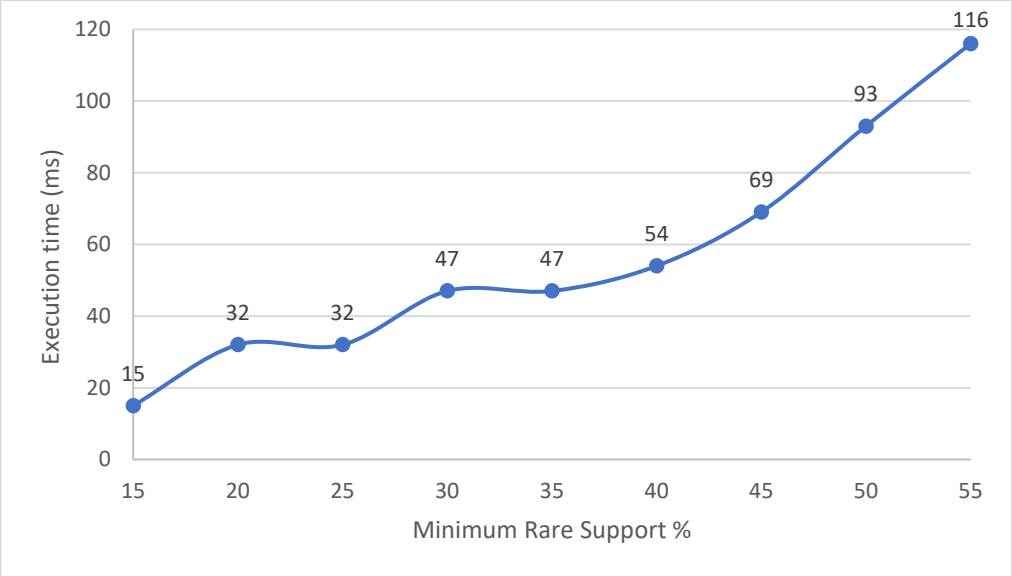

Figure 5.8 The execution time performance of the RP-Growth algorithm with different minimum rare support thresholds

5.5 Experimental Results of ERIM

Three different experiments were conducted on the gear manufacturing downtime dataset in order to analyze the followings:

(i) the WRIs obtained individually from the used RIM algorithms (Apriori Rare, Apriori Inverse, CORI, and RP-Growth) and their relationship with some varying measure thresholds,

(ii) the distribution of the number of SRIs obtained from the ERIM approach by their lengths,

(iii) examples of discovered SRIs obtained from the ERIM approach (downtime anomalies).

The first experiment applied the four different base algorithms, Apriori Rare, Apriori Inverse, CORI, and RP-Growth, individually on the gear manufacturing downtime dataset to obtain WRIs (Akdaş, Birant, & Yıldırım Taşer, 2022).

Each algorithm has different measurement parameters, and the threshold values of these parameters used in this experiment are given in Table 5.2.

Table 5.2 The parameters of the algorithms used to discover WRIs

Algorithm	**Parameter**	**Values (%)**
Apriori Rare	Minimum support	ranging from 10 to 50 in increments of 5
Apriori Inverse	Minimum support	0.1
	Maximum support	ranging from 1 to 9 in increments of 1
CORI	Minimum support	ranging from 1 to 9 in increments of 1
	Bound	9
RP-Growth	Minimum support	90
	Minimum rare support	ranging from 15 to 55 in increments of 5

Figure 5.9 plots the number of WRIs obtained from each algorithm and their relationship with varying support thresholds. This graph demonstrates that the higher the minimum support and minimum rare support values, the lower the number of WRIs obtained by Apriori Rare and RP-Growth algorithms, respectively. The situation is the opposite in Apriori Inverse and CORI algorithms, such that when maximum support and minimum support values decrease, the number of WRIs obtained by Apriori Inverse and CORI algorithms also decreases, respectively.

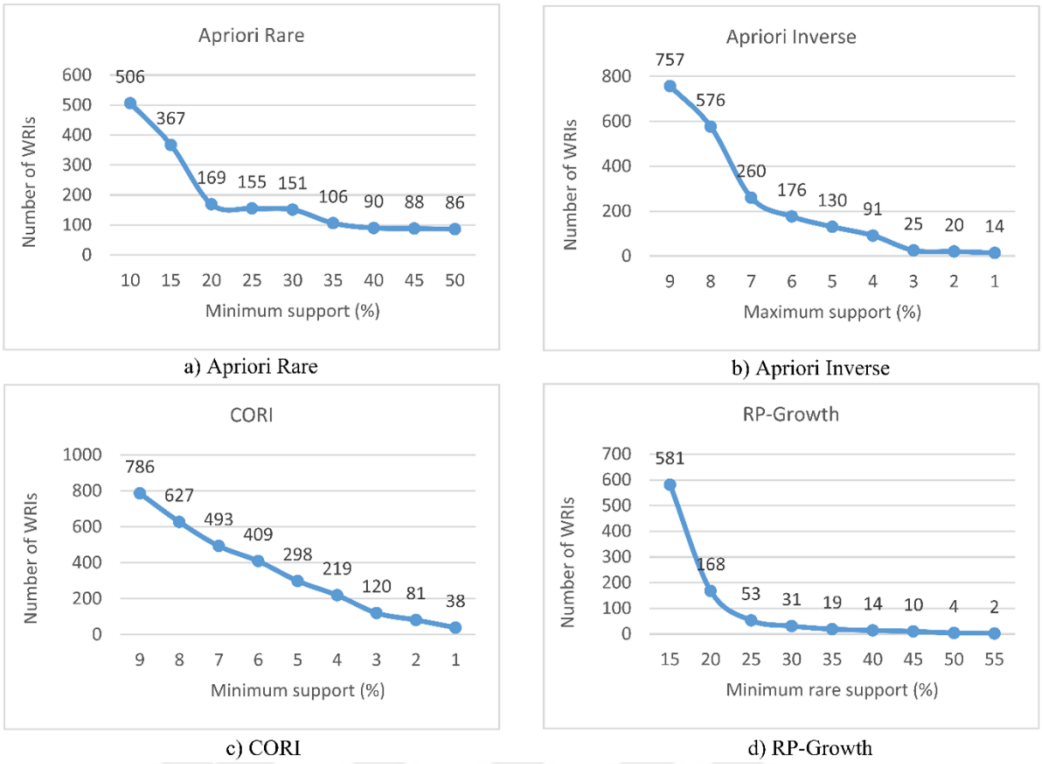

Figure 5.9 The number of WRIs obtained by a) Apriori Rare, b) Apriori Inverse, c) CORI, and d) RP-Growth algorithms with different support thresholds.

In the second and third experiments, the parameters of the base algorithms of the ERIM approach were chosen as follows:

- **Apriori Rare:** minimum support = 0.4%
- **Apriori Inverse:** minimum support = 0.1%, maximum support = 35%
- **CORI:** minimum support = 90%, bound = 9%
- **RP-Growth:** minimum support = 90%, minimum rare support = 0.5%

In the second experiment, we implemented the ERIM approach on the same dataset to discover SRIs that would indicate reliable and robust gear manufacturing downtimes. Figure 5.10 illustrates the number of SRIs based on their different lengths, such as 1-item, 2-itemset, and so on. It is clearly seen from this graph that the 3-itemset

has the highest number of discovered SRIs. Also, the number of 3-itemset and 4-itemset SRIs are rather close. It is understood from this figure that the ERIM technique generated a shape that was pretty comparable to the bell curve in general.

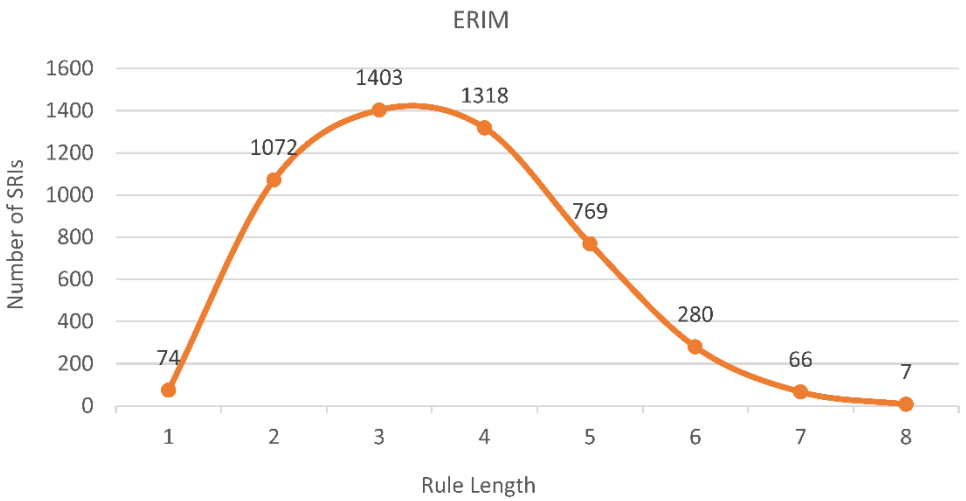

Figure 5.10 The distribution of the number of SRIs by their lengths

In the last experiment, the ERIM approach was applied to the same dataset, and some examples of the discovered SRIs were presented in Table 6. For example, the itemset "Day=Saturday, Month=October, OperationName=CNC turning, WorkCenterCode=IM02" has a 2.35% support value means that it is an anomaly (rare downtime), and it occurs only in 2.35% of total downtime data. It also implies that on Saturdays in October, the downtime in work center IM02, where the CNC turning operation is performed, is less frequent than the others. This table also provides a comparison of the ERIM approach with the state-of-the-art methodologies (Apriori Rare (Szathmary et al., 2007), Apriori Inverse (Koh et al., 2005), CORI (Bouasker et al., 2015), and RP-Growth (Tsang et al., 2011)). The table clearly shows which state-of-the-art methods were discovered for the SRIs. The results indicate that SRIs, which could not be discovered by some state-of-the-art methods, were successfully detected with the proposed ERIM approach. The results also demonstrated that our method outperformed the state-of-the-art methods by 43.37% on average on the same dataset. By analyzing these discovered anomalies (SRIs), it is possible to make decisions about efficient productivity.

Table 5.3 Examples of the SRIs obtained by the proposed ERIM approach

Length	Strong Rare Itemsets	Support (%)	Apriori Rare (Szathmary et.al, 2007)	Apriori Inverse (Koh et. al, 2005)	RP-Growth (Tsang et. al, 2011)	CORI (Bouaskeret. al, 2015)	ERIM (proposed)
1-Item	Downtime=Initial adjustment	4.26	-	√	√	√	√
	OperationName=Hard turning	6.81	-	√	√	√	√
	TotalProduction=Very low	7.12	-	-	√	√	√
2-Itemset	Month=January, Time=Very high	0.19	√	√	-	-	√
	Day=Sunday, WorkCenterCode= IM16	0.24	√	√	-	-	√
	DowntimeGroupName=Personnel, OperationName=CNC turning	2.00	-	√	√	-	√
3-Itemset	DowntimeGroupName=Personnel, OperationName= Finishing and grinding, WorkCenterGroupCode=IMG 134	0.71	-	√	√	-	√
	Downtime=Initial adjustment, DowntimeGroupName=Workbench preparation, Time=Very high	2.36	-	√	√	√	√
	Downtime= Waiting for approval, DowntimeGroupName=Quality, TotalProduction=Very low	1.09	-	√	√	√	√
4-Itemset	Day=Saturday, Month=October, OperationName=CNC turning, WorkCenterCode=IM02	2.35	-	√	√	-	√
	Month=April, WorkCenterGroupCode=IMG 130, OperationName=Gear cutting, Department=Gear	5.72	-	-	√	√	√
	Day=Wednesday, Downtime=Launch, DowntimeType=Planned, Time=Medium	2.32	-	√	√	-	√
5-Itemset	Downtime=Launch, DowntimeType=Planned, OperationName=CNC turning, Time=Very low, WorkCenterCode=IM03	2.45	-	√	√	-	√

Table 5.3 continues

	Month=October, OperationName=CNC turning, Time=Very low, TotalProduction=Low, WorkCenterCode=IM02	2.72	-	√	√	-	√
	Department= Finishing, Month=September, OperationName=Hard turning, WorkCenterCode=IM19, WorkCenterGroupCode=IMG 141	1.32	-	√	√	-	√

CHAPTER SIX
CONCLUSION AND FUTURE WORKS

6.1 Conclusion

The conventional RIM algorithms work independently to discover rare itemsets from an entire dataset. However, these itemsets can be weak to indicate uncommon items, events, or observations. Because of this reason, this study proposes an approach called ERIM, which uses multiple RIM algorithms as base learners and combines them to uncover SRIs to detect anomalies. Unlike typical RIM approaches, the ERIM approach can identify more targeted, reliable, and global SRIs from WRIs locally discovered by different algorithms. It is the first study that combines RIM and ensemble learning methodologies. The proposed ERIM approach is a general method; therefore, it can be effectively used in future research projects. In this paper, the proposed ERIM approach was applied to a real-world dataset to detect anomalies in gear manufacturing downtime of earth-moving machinery as a case study. For the first time, the ERIM approach was applied to detect rare itemsets (anomalies) in machine downtimes. The main improvement of this study is that the ERIM approach converted WRIs into SRIs that indicate significantly meaningful anomalies. The base learners of the ERIM approach were selected as Apriori Rare, Apriori Inverse, CORI, and RP-Growth because of their advantages, such as speed of analyses, easy implementation, and relevance accuracy.

The experimental results were evaluated regarding the number of itemsets, and the length of itemsets. The results showed that the proposed ERIM approach gives reliable, common knowledge by jointly considering the outputs of different algorithms in an ensemble manner. The results also showed that the proposed ERIM approach successfully detected the anomalies whose support values are below 7.12. Besides, it is easily understood from the experimental results that the ERIM discovered the highest number of SRIs with 1403 and each of them is a 3-itemset. Finally, the outcomes also demonstrated that our method outperformed the state-of-the-art methods by 43.37% on average on the same dataset.

6.2 Future Works

In future research, the proposed ERIM approach can be implemented for detecting anomalies in different domains, including finance, healthcare, law, and computer networks. Also, the ensemble structure can be modified for discovering frequent itemsets, instead of rare itemsets. In this way, weak frequent itemsets can be converted into strong frequent itemsets. Moreover, considering confidence threshold values, rare association rules can be combined in an ensemble manner.

REFERENCES

Abed, S., Abdelaal, A. A., Al-Shayeji, M. H., & Ahmad, I. (2020). SAT-based and CP-based declarative approaches for Top-Rank-K closed frequent itemset mining. *International Journal of Intelligent Systems, 36*(1), pp. 112–151. doi:10.1002/int.22294

Agrawal, R., Imieliński, T., & Swami, A. (1993). Mining association rules between sets of items in large databases. *ACM SIGMOD Record, 22*(2), pp. 207–216. doi:10.1145/170036.170072

Akdaş, D. N., Birant, D., & Yıldırım Taşer, P. (2022). ERIM: An ensemble of rare itemset mining and its application in the automotive industry. *Expert Systems*, pp. 1– 14. doi:10.1111/exsy.13122

Akdaş, D. N., Birant, D., & Yıldırım Taşer, P. (2023). Anomaly Detection for Gear Manufacturing Downtime in The Automotive Sector Using Rare Itemset Mining. *International Journal of Innovative Engineering Applications*. doi:10.46460/ijiea.1067365

Borah, A., & Nath, B. (2018). Identifying risk factors for adverse diseases using dynamic rare association rule mining. *Expert Systems with Applications, 113*, s. 233-263. doi:doi:10.1016/j.eswa.2018.07.010

Borgelt, C. (2012). Frequent item set mining. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 2*(6), pp. 437–456. doi:10.1002/widm.1074

Bouasker, S., & Yahia, S. B. (2015). Key correlation mining by simultaneous monotone and anti-monotone constraints checking. *Proceedings of the 30th Annual ACM Symposium on Applied Computing*, pp. 851-856. doi:10.1145/2695664.2695802

Bouasker, S., Inoubli, W., Yahia, S. B., & Diallo, G. (2021). Pregnancy Associated Breast Cancer Gene Expressions : New Insights on Their Regulation Based on Rare Correlated Patterns. *IEEE/ACM Trans Comput Biol Bioinform, 18*(3), pp. 1035-1048. doi:10.1109/TCBB.2020.3015236

Böhmer, K., & Rinderle-Ma, S. (2020). Mining association rules for anomaly detection in dynamic process runtime behavior and explaining the root cause to users. *Information Systems, 90*. doi:10.1016/j.is.2019.101438

Chandola, V., Banerjee, A., & Kumar, V. (2009, July). Anomaly detection: A survey. *ACM Computing Surveys*, pp. 1-58. doi:https://doi.org/10.1145/1541880.1541882

Chicco, D., & Jurman, G. (2021). An Ensemble Learning Approach for Enhanced Classification of Patients With Hepatitis and Cirrhosis. *IEEE Access, 9*, pp. 24485–24498. doi:10.1109/access.2021.3057196

Darrab, S., Broneske, D., & Saake, G. (2021). Modern Applications and Challenges for Rare Itemset Mining. *International Journal of Machine Learning and Computing, 11*(3), pp. 208–218. doi:10.18178/ijmlc.2021.11.3.1037

Deng, Z. (2013). Mining Top-Rank-k Erasable Itemsets by PID_lists. *International Journal of Intelligent Systems, 28*(4), pp. 366–379. doi:10.1002/int.21580

Dong, X., Yu, Z., Cao, W., Shi, Y., & Ma, Q. (2019). A survey on ensemble learning. *Frontiers of Computer Science, 14*(2), pp. 241–258. doi:10.1007/s11704-019-8208-z

Dongdong, L., Ziqiu, C., Bolu, W., Zhe, W., Hai, Y., & Wenli, D. (2021). Entropy-based hybrid sampling ensemble learning for imbalanced data. *International Journal of Intelligent Systems, 36*(7), pp. 3039–3067. doi:10.1002/int.22388

Fournier-Viger, P., Lin, J. C.-W., Gomariz, A., Gueniche, T., Soltani, A., Deng, Z., & Lam, H. T. (2016). The SPMF Open-Source Data Mining Library Version 2. *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pp. 36-40.

Garcés, D., & Castrillón, O. (2017). Diseño de una Técnica Inteligente para Identificar y Reducir los Tiempos Muertos en un Sistema de Producción. *Información Tecnológica, 28*(3), pp. 157-170. doi:10.4067/S0718-07642017000300017

Gupta, A., & Semwa, V. B. (2020). Multiple Task Human Gait Analysis and Identification: Ensemble Learning Approach. *Emotion and Information Processing*, pp. 185-197. doi:10.1007/978-3-030-48849-9_12

Han, J., Cheng, H., Xin, D., & Yan, X. (2007). Frequent pattern mining: current status and future directions. *Data Mining and Knowledge Discovery, 15*(1), pp. 55–86. doi:10.1007/s10618-006-0059-1

Jain, R., Semwal, V. B., & Kaushik, P. (2021, June). Deep ensemble learning approach for lower extremity activities recognition using wearable sensors. *Expert Systems, 36*(9). doi:10.1111/exsy.12743

Jeyakarthic, M., & Singaram, S. (2019, February). An efficient approach using FM-weight for Revenue Prediction on Rare Itemsets. *International Journal of Recent Technology and Engineering*, s. 226-232.

Ji, Y., Ying, H., Tran, J., Dews, P., Mansour, A., & Massanari, R. M. (2013, April). A Method for Mining Infrequent Causal Associations and Its Application in Finding Adverse Drug Reaction Signal Pairs. *IEEE Transactions on Knowledge and Data Engineering, 25*(4), pp. 721-733. doi:10.1109/TKDE.2012.28

Koh, Y. S., & Rountree, N. (2005). Finding Sporadic Rules Using Apriori-Inverse. *Advances in Knowledge Discovery and Data Mining*, pp. 97–106. doi:10.1007/11430919_13

Li, Y., & Chen, W. (2020). A Comparative Performance Assessment of Ensemble Learning for Credit Scoring. *Mathematics, 8*(10), p. 1756. doi:10.3390/math8101756

Luna, J., Romero, C., Romero, J., & Ventura, S. (2015, April). An evolutionary algorithm for the discovery of rare class association rules in learning management systems. *Applied Intelligence, 42*(3), pp. 501-513. doi:https://doi.org/10.1007/s10489-014-0603-4

Mucchielli, P., Bhowmik, B., Ghosh, B., & Pakrashi, V. (2021). Real-time accurate detection of wind turbine downtime - An Irish perspective. *Renewable Energy, 179*, pp. 1969-1989. doi:10.1016/j.renene.2021.07.139

Nithya, S., & Jayakumar, C. (2016). Automatic Firewall Rule Generator for Network Intrusion Detection System based on Multiple Minimum Support. *Indian Journal of Science and Technology, 9*(41), pp. 1-4. doi:10.17485/ijst/2016/v9i41/86987

Nwanya, C., Udofia, J., & Ajayi, O. (2017, May). Optimization of machine downtime in the plastic manufacturing. *Cogent Engineering, 4*(1). doi:10.1080/23311916.2017.1335444

Onan, A., Korukoğlu, S., & Bulut, H. (2016). A multiobjective weighted voting ensemble classifier based on differential evolution algorithm for text sentiment classification. *Expert Systems with Applications, 62*, pp. 1-16. doi:10.1016/j.eswa.2016.06.005

Quatrini, E., Costantino, F., Di Gravio, G., & Patriarca, R. (2020, May). Machine learning for anomaly detection and process phase classification to improve

safety and maintenance activities. *Journal of Manufacturing Systems*(56), pp. 117–132. doi:https://doi.org/10.1016/j.jmsy.2020.05.013

Quatrini, E., Costantino, F., Gravio, G. D., & Patriarca, R. (2020). Machine learning for anomaly detection and process phase classification to improve safety and maintenance activities. *Journal of Manufacturing Systems, 56*, s. 117-132. doi:10.1016/j.jmsy.2020.05.013

Ramachandran, A., & Sangaiah, A. K. (2018). Unsupervised Anomaly Detection for High Dimensional Data—an Exploratory Analysis. *Computational Intelligence for Multimedia Big Data on the Cloud With Engineering Applications*, s. 233–251. doi:10.1016/B978-0-12-813314-9.00011-6

Reps, J., & Aickelin, U. (2015, January). Refining Adverse Drug Reaction Signals by Incorporating Interaction Variables Identified Using Emergent Pattern Mining. *SSRN Electronic Journal*. doi:10.2139/ssrn.2822199

Roosefert Mohan, T., Preetha Roselyn, J., Annie Uthra, R., Devaraj, D., & Umachandran, K. (2021). Intelligent machine learning based total productive maintenance approach for achieving zero downtime in industrial machinery. *Comput. Ind. Eng., 157*. doi:10.1016/J.CIE.2021.107267

Semwal, V. B., Gupta, A., & Lalwani, P. (2021). An optimized hybrid deep learning model using ensemble learning approach for human walking activities recognition. *The Journal of Supercomputing, 77*(11), s. 12256–12279. doi:10.1007/s11227-021-03768-7

Shafieezadeh, A., Desroches, R., Rix, G., & Werner, S. (2014, April). A probabilistic framework for correlated seismic downtime and repair cost estimation of geo-structures. *Earthquake Engineering & Structural Dynamics, 43*(5). doi:10.1002/eqe.2369

Shrivastava, K., & Jotwani, V. (2020, June). Study to Determine Adverse Diseases Pattern using Rare Association Rule Mining. *International Journal of Scientific Research in Computer Science, Engineering and Information Technology*, pp. 519-526. doi:10.32628/CSEIT2063111

Soltanali, H., Rohani, A., Tabasizadeh, M., Abbaspour-Fard, M., & Aditya, P. (2018, July). Improving the performance measurement using Overall Equipment Effectiveness(OEE) in an automotive industry. *International Journal of Automotive Engineering, 8*(3), pp. 2781-2791. doi:10.22068/ijae.8.3.2781

Szathmary, L., Napoli, A., & Valtchev, P. (2007). Towards Rare Itemset Mining. *19th IEEE International Conference on Tools With Artificial Intelligence(ICTAI 2007)*, pp. 305-312. doi:10.1109/ictai.2007.30

Tsang, S., Koh, Y. S., & Dobbie, G. (2011). RP-Tree: Rare Pattern Tree Mining. *International Conference on Data Warehousing and Knowledge Discovery*, pp. 277–288. doi:10.1007/978-3-642-23544-3_21

Wang, D., Liu, F., & Jin, Y. (2019). A proactive scheduling approach to steel rolling process with stochastic machine breakdown. *Natural Computing, 18*(4), pp. 679-694. doi:10.1007/s11047-016-9599-5

Weng, C.-H. (2011, February). Mining fuzzy specific rare itemsets for education data. *Knowledge-Based Systems, 24*(5), pp. 697-708. doi:https://doi.org/10.1016/j.knosys.2011.02.010

Wulandari, C., Ou-Yang, C., & Wang, H. (2019, March). Applying mutual information for discretization to support the discovery of rare-unusual association rule in cerebrovascular examination dataset. *Expert Systems with Applications, 118*, pp. 52-64. doi:https://doi.org/10.1016/j.eswa.2018.09.044

Yıldırım, P., Birant, U., & Birant, D. (2019, March). EBOC: Ensemble-Based Ordinal Classification in Transportation. *Journal of Advanced Transportation*, pp. 1-17. doi: https://doi.org/10.1155/2019/7482138

Zhu, J., Wang, K., Wu, Y., Hu, Z., & Wang, H. (2016, July). Mining User-Aware Rare Sequential Topic Patterns in Document Streams. *IEEE Transactions on Knowledge and Data Engineering, 28*(7), s. 1790 - 1804. doi:10.1109/TKDE.2016.2541149