ESSAYS ON SUSTAINABLE SUPPLY CHAIN, GROUP DECISION MAKING, AND EXPERT-AUGMENTED FEATURE SELECTION

by

MEYSAM RABIEE

A DISSERTATION

Presented to the Department of Operations and Business Analytics
and the Division of Graduate Studies of the University of Oregon
in partial fulfillment of the requirements
for the degree of
Doctor of Philosophy

June 2022

DISSERTATION APPROVAL PAGE

Student: Meysam Rabiee

Title: Essays on Sustainable Supply Chain, Group Decision Making, and Expert-Augmented Feature Selection

This dissertation has been accepted and approved in partial fulfillment of the requirements for the Doctor of Philosophy degree in the Department of Operations and Business Analytics by:

Michael Pangburn	Chairperson
Dursun Delen	Core Member
Saeed Piri	Core Member
Edward Rubin	Institutional Representative

and

Krista Chronister	Vice Provost for Graduate Studies

Original approval signatures are on file with the University of Oregon Division of Graduate Studies.

Degree awarded June 2022

DISSERTATION ABSTRACT

Meysam Rabiee

Doctor of Philosophy

Department of Operations and Business Analytics

June 2022

Title: Essays on Sustainable Supply Chain, Group Decision Making, and Expert-Augmented Feature Selection

My dissertation consists of three essays on key areas including sustainable supply chain management, group decision making and expert-augmented feature selection. My first essay is a previously unpublished co-authored work with Prof. Nagesh Murthy and Dr. Hossein Rikhtehgar Berenji. In some supply chains, it is extraordinarily expensive for a buyer to audit all selected suppliers to guarantee compliance with the buyer's social and environmental responsibility code of conduct. In this work, we provide insight to help such a buyer profit from judicious audits, despite risk of revenue loss due to non-compliance. The possible benefits and issues of group decision making in multi-criteria decision making and feature selection problems are a central theme of the second and third essays in this dissertation. My second essay is a co-authored piece with Babak Aslani and Dr. Jafar Rezaei that was previously published. The purpose of this article is to investigate the detection and handling of biased decision-makers in group decision-making processes. We develop three algorithms including extreme, moderate, and soft versions to address this issue. The third essay is a previously unpublished co-authored work with Mohsen Mirhashemi, Michael Pangburn and Dursun Delen. In this study, we enrich the conventional feature selection method by incorporating the opinions of experts on the features, a technique we refer to as expert-augmented feature selection. To reflect the trade-off between explainability and prediction

accuracy, we develop two very similar models (one for classification and one for regression problems). Finally, we develop a posterior ensemble approach for quantifying each feature's accuracy contribution degree. This algorithm's output helps us to discover features that are consistently, under- or over-rated by experts.

This dissertation includes both previously published/unpublished and co-authored material.

CURRICULUM VITAE

NAME OF AUTHOR: Meysam Rabiee

GRADUATE AND UNDERGRADUATE SCHOOLS ATTENDED:

University of Oregon, Eugene, Oregon, USA
K. N. Toosi University of Technology, Tehran, Iran
Bu-Ali Sina University, Hamadan, Iran

DEGREES AWARDED:

Doctor of Philosophy, Operations and Business Analytics, 2022, University of Oregon
Master of Science, Industrial Engineering, 2010, K. N. Toosi University of Technology
Bachelor of Science, Industrial Engineering, 2008, Bu-Ali Sina University

AREAS OF SPECIAL INTEREST:

Healthcare Analytics
Sustainable Supply Chain
Large-scale Group Decision Making
Data Mining
Evolutionary Algorithms

PROFESSIONAL EXPERIENCE:

Full-time Lecturer, Bu-Ali Sina University, 2012-2016
Program Coordinator, Bu-Ali Sina University, 2015-2016

GRANTS, AWARDS, AND HONORS:

Robin and Roger Best Teaching Award, Lundquist College of Business, 2021
Robin and Roger Best Teaching Award, Lundquist College of Business, 2022
Robin and Roger Best Research Award, Lundquist College of Business, 2021

ACKNOWLEDGMENTS

I want to express my gratitude to my committee members, co-authors, mentors, and family members. I would not have been able to accomplish this without their assistance and support.

To begin, I want to express my deepest gratitude to Prof. Michael Pangburn, chair of my committee and advisor. I only got the opportunity to work with him for a year. However, from the start of our interactions, I learned a lot from him. He always put my interests first and was always available to guide and help me. As a caring and kind advisor, he pointed out my areas of weakness and provided me with some valuable advice that was immensely beneficial during my job search.

I also want to express my appreciation to all my co-authors and mentors who patiently worked with me and inspired me along this journey.

A special thanks goes to other members of my PhD committee, including Prof. Dursun Delen, Dr. Saeed Piri, and Dr. Edward Rubin, for their constructive feedback that was very useful in the advancement of my dissertation.

I also want to thank all of the PhD directors and employees at Lundquist College of Business for their great support and guidance throughout my PhD studies.

I want to thank all of my friends who helped me make a lot of wonderful memories during my time studying in the lovely state of Oregon.

To my dear wife, Mona.

It didn't matter how difficult PhD life sometimes get,

she was always there

to provide a helping hand

and a listening ear.

TABLE OF CONTENTS

LIST OF FIGURES

Figure Page

LIST OF TABLES

Chapter 1
Overview

My dissertation consists of an essay in sustainable supply chain management, an essay in group decision making, and an essay in expert-augmented feature selection. My first essay is an unpublished work co-authored with Prof. Nagesh Murthy and Dr. Hossein Rikhtehgar Berenji. In some supply chains, it is extraordinarily expensive for a buyer to audit all selected suppliers to guarantee compliance with the buyer's code of conduct for social and environmental responsibility. In this work, we provide insight to help such a buyer profit from judicious audits, despite the risk of revenue loss due to non-compliance.

My second essay is co-authored with Babak Aslani and Dr. Jafar Rezaei; it has been published. The purpose of this article is to investigate the detection and handling of biased decision-makers in group decision-making processes. To address this issue, we developed three algorithms including extreme, moderate, and soft versions.

The third essay is an unpublished work co-authored with Mohsen Mirhashemi, Prof. Michael Pangburn, Prof. Dursun Delen, and Dr. Saeed Piri. In this study, we enrich the conventional feature-selection method by incorporating the opinions of experts on the features, a technique we refer to as expert-augmented feature selection. To reflect the trade-off between explainability and prediction accuracy, we develop two very similar models: one for classification problems, and one for regression problems. Finally, we develop a posterior ensemble approach for quantifying each feature's accuracy-contribution degree. This algorithm's output helps us to discover features that are consistently, under-rated or over-rated by experts.

1.1 Essay 1: Judicious Audits for Managing Social and Environmental Compliance in the Supplier Base

Recent years have seen a surge of interest in research on green-supplier selection and order allocation. The increasingly competitive complexity resulting from social and environmental considerations is the primary driver of these studies, which examine the buyer's experience when confronted with customers who are socially and environmentally conscious. In certain supply chains, it is prohibitively expensive to audit all selected suppliers to ensure compliance with the buyer's code of conduct for social and environmental responsibility. If the buyer chooses to inspect all its suppliers to ensure conformity, it prevents market loss due to a having a supplier with a lack of sustainability. On the other hand, if the buyer chooses any supplier without auditing, there is a risk of business loss due to the possibility of non-compliance in procurement. When a supplier chosen without auditing is actually non-compliant, the media or non-governmental organizations expose social or environmental violations. Additionally, if any supplier fails an audit, the buyer would require the supplier to rectify the issue by performing corrective actions. After incurring the costs of corrective action, the supplier will only agree to do so if the expected benefit is greater than its minimum expectation. The buyer's margin for the item is determined by whether the chosen suppliers follow the code of conduct. In this study, we develop mixed-integer models to mimic the aforementioned issues in two separate environments, bridging major gaps in the literature.

In the first setting, we consider that the margin loss associated with non-compliance in procurement (referred to as market loss) is negligible, meaning that it refers only to the portion of the market sourced by non-compliant suppliers. A technology such as Blockchain enables

identification of the source of non-compliant supply as well as the corresponding quantity sourced, so this information can be delineated in a transparent manner for consumers.

In the second setting, the market loss is considered to be uniform (i.e., for the entire supply), even when not all chosen suppliers are non-compliant. As a result, when transparency is lacking or customers are very socially and environmentally aware, customers demand complete compliance in the supplier base; thus, with a non-compliant supplier, the buyer will lose part of its total market (and therefore part of its profit). To provide insight into this issue, we expanded on the design of the experiment: ten instances were created at random in each of sixteen separate extreme scenarios, yielding a total of 160 problems. Along with profit maximization as a primary objective function, we developed two additional performance metrics for further analysis: Portion of Suppliers Selected with Auditing (PSSA), and a Sustainability Index (SI). Then, we used a multiple linear regression model for each combination of performance measure and model setting, to estimate the marginal effects and interaction effects. This provides some insight for such a buyer about the benefits judicious auditing, despite the risk of lost revenue due to any non-compliance.

1.2 Essay 2: A Decision-Support System for Detecting and Handling Biased Decision-Makers in Multi-Criteria Group Decision-Making Problems

Detecting and handling biased decision-makers in a group decision-making process is overlooked in the literature. This paper aims to develop an anti-bias statistical approach, including extreme, moderate, and soft versions, as a decision support system for group decision-making (GDM) to detect and handle bias. The extreme version starts with eliminating the biased decision-makers. For this purpose, the decision-makers (DMs) with a Biasedness Index value that is higher than a predefined threshold are removed from the process. Next, it continues with a procedure to mitigate the effect of partially biased DMs by assigning different weights to DMs with respect to their

biasedness level. To do so, two ratios for the remaining DMs are calculated: (i) Overlap Ratio, which shows the relative value of overlap between the confidence interval (CI) of each DM and the maximum possible overlap value; and (ii) Relative confidence interval CI, which reflects the relative value of CI for each DM compared to the confidence interval CI of all DMs. The final step is assigning a weight to each DM, considering the two values Overlap Ratio and Relative Confidence Interval. DMs with opinions closer to the aggregated opinion of all DMs gain more weight. The framework adequately addresses and prescribes possible actions for almost all possible cases in GDM including without any outliers (i.e., acceptable consensus among DMs), cases with partial outliers, and extreme cases with complete disagreement among DMs. The moderate version preassigns a minimum weight to the remaining DMs and then follows the weighting step for the remaining total weight. However, the soft version follows the pre-assignment of weights to all DMs in the initial pool, meaning there is no elimination in this setting. The proposed approach is tested for several scenarios with different sizes. Four performance measures are introduced to evaluate the effectiveness and reliability of the proposed method.

1.3 Essay 3: Expert-Augmented Supervised Feature Selection: Models and Algorithms

Feature-selection techniques have two fundamental conflicting objectives: maximizing classification/regression accuracy while decreasing model complexity (i.e., fewer features) to avoid the curse of dimensionality. Here, we add another layer to the traditional setup by considering academicians' (and potentially practitioners') opinions on the features, a process we refer to as expert-augmented feature selection. We contribute to the literature in three distinct ways with this paper. First, we formulate two simple optimization models (one for classification problems, and one for regression problems) to capture the trade-off between explainability and prediction accuracy. In this formulation, we allow the decision-maker (or domain expert) to specify

a sacrifice percentage in accuracy, to increase explainability by reducing the number of selected features and prioritizing those that are more influential from the perspective of practitioners or academicians. Second, we develop a probabilistic solution generator via an information fusion (PSGIF) algorithm that functions as an ensemble operator, collecting data from several filter techniques to improve the genetic algorithm's exploration and exploitation process. Third, we optimized the parameters of our method using Bayesian optimization, to reduce the algorithm's computational time; this helps the scalability of our developed method. We compare the performance of our proposed methods to some well-known algorithms on 16 publicly available datasets. The analysis revealed that selected versions of our genetic algorithms outperform other benchmark algorithms in terms of average accuracy rate for classification datasets and average MSE for regression datasets. Finally, we develop a posterior ensemble technique for estimating the accuracy-contribution degree of each feature. The result of this algorithm enables us to identify features that are underrated, overrated, or consistently rated by academicians (or practitioners).

Chapter 2

Judicious Audits for Managing Social and Environmental Compliance in the Supplier Base

This work is in preparation for peer-review journal submission and is co-authored with Prof. Nagesh N. Murthy and Dr. Hossein Rikhtehgar Berenji.

2.1 Introduction

Global competition, frequent shifts in client demands, and increased outsourcing have brought more technical challenges in supply chain management. Selecting the most appropriate vendors and allocating orders to them to achieve greater customer loyalty, reduced operating costs, higher profitability, and competitive advantage is a major strategic priority. Supplier selection and order allocation is essential to improving corporate efficiency and build productivity for an enterprise and its supply chain. Thus, decisions on supplier collection and order distribution have become an important part of supply chain management (Azadnia et al., 2015; Kannan et al., 2013; Hamdan & Cheaitou, 2017b; Moheb-Alizadeh & Handfield, 2019).

Elkington (1998) established the three-pillar framework of sustainability: benefit, Earth, and people. Researchers and practitioners occasionally replace sustainability with other concepts such as green corporate social responsibility (CSR). About a quarter of a century earlier, pioneer supply chain researchers began work on sustainability problems (e.g., Klassen and McLaughlin, 1996). CSR practices are not considered a secondary or side objective by most organizations, and organizations are well placed to strategize their missions and make it possible to work together with value-adding efforts in all areas of CSR.

About a decade ago, it appeared that some overseas suppliers of Apple, Dell, and HP had forced their employees to work under hazardous conditions. In another embarrassing event, there were reports of chemicals being dumped in the rivers of China by suppliers of Nike and Adidas. As another example, violating the environmental code of conduct by one of the tier-two suppliers of General Motors resulted in a massive explosion, with 163 people severely injured and 97 killed. Violating CSR values is not limited to environmental issues. An internal audit revealed multiple cases of forced labor and human trafficking in Patagonia's supply chain in Taiwan. This issue was originated by labor brokers due to the labor shortage of many companies that relied on illegal workers; those companies are mostly from Vietnam, Indonesia, Thailand, and the Philippines.

These were just a few examples of bad publicity that hurt a brand's reputation. The consequences of these violations could show up in market share loss or revenue loss. If the source of a violation, in terms of the specific supplier and product, is viable for the company and its customers, the market loss would be marginal and limited to the origin of the violation. If this situation does not occur, the aftereffect will hurt the buyer's revenue or market share as a whole. Violations of environmental and social regulations can have enormous consequences. Research conducted by JB Were Support indicated that "76% will refuse to purchase a company's products or services upon learning it supported an issue contrary to their beliefs". This can be seen as an alarming sign for any company that underestimates the consequences of CSR violations.

Customers are increasingly seeking corporate social responsibility (CSR) and environmental policies from companies. A 2018 poll by JB Were Support found that "63% of Americans, without government regulation, are optimistic corporations, would contribute to social and environmental improvement." There are multiple examples of CSR activities: a reduction of carbon footprint (Coca-Cola's 25% reduction in carbon footprint by 2020); any reform in business capable of mitigating environmental harm (reduction in water usage by 1 billion liters); and terminating the relationship with any agency that breaches CSR values (Lego ending a half-century partnership with Shell Company due to Arctic drilling). By being compliant on sustainability codes of conduct, companies can benefit in their consumer-hunting efforts by focusing more on CSR practices.

Recently, we have seen an increasing trend in the number of companies that have committed to working only with socially and environmentally compliant suppliers. These companies try to use their negotiation power to persuade their supplier base to be compliant with the sustainability code of conduct. However, there is no guarantee of such compliance due to a lack of resources and incentives. Moreover, this commitment turned out to be ineffective, and many of these companies who pledged themselves to follow this have experienced controversies caused by vendors who nonetheless violated the commitment, despite being mindful of sustainability standards.

The simplest solution for this issue is auditing all the suppliers and terminating contracts with non-compliant suppliers if they are not willing to resolve the issues found. However, this approach does not work for most small or medium-sized businesses, since it is a demanding and expensive task to determine all sustainability-related activities for all suppliers participating in a

supply chain. In an effort to find ways to reduce costs, several businesses have used cross-utilization, a concept that involves pooling resources and using supplier-sustainability evaluation methods. It is very cost-effective to have joint auditing in place, but this approach applies only to firms that have a common base of suppliers and may not be applicable to any existing supply chain. That is where the concept of judicious auditing offers a more generalizable solution based on several factors: the likelihood of sustainability compliance, the market type (forgiving vs. sensitive), the degree of market loss, the auditing cost, and examining suppliers' willingness to take corrective action when caught by auditing. Amazon, for example, is implementing a framework for auditing that includes this statement:

> *When violations are identified, suppliers must develop a corrective action plan that details immediate actions to address high-risk issues, and a long-term plan to prevent issues from reoccurring. Where suppliers fail to meet our standards or refuse to make progress on remediating issues, we may choose to terminate the relationship.*

There are a few notable points in their supplier statement above. First, they have not claimed to audit all suppliers. Second, to ease the sustainability compliance burden on suppliers, they partner with independent auditors and confidential worker interviews to verify sustainability compliance. Third, they expect their suppliers to be able to perform the corrective action if any violation is identified. We tried to embed these features in modeling our studied problem.

As anticipated, our results show that on average, having a more compliant supplier base results in higher profit and fewer audited suppliers. Moreover, our results show that on average, as the auditing cost ratio increases, the buyer has less willingness to audit. The same interpretation applies to the corrective action cost ratio and the suppliers. However, there is such a case in which the buyer may prefer to audit even the whole of selected suppliers to prevent the risk of market punishment. We find that if the buyer has a non-compliant supplier base, the buyer will not necessarily be better off if it decides to audit more, since there are some cases in which the market loss is not high enough to consider auditing even if the auditing cost is inexpensive. The general behavior of both developed models is consistent except in a few places. In the marginal-loss model, when the market loss is low, the proportion of suppliers being audited is lower compared to when the market loss is high, regardless of the compliance degree of the supplier base. In contrast to the marginal model, in the total-loss model when the market loss is low, as the sustainability of the

supplier base increases, for a fixed level of auditing cost and corrective action, the buyer chooses to audit more of the suppler base to guarantee the minimum loss.

In this paper, we extended the traditional supplier-selection order-allocation (SSOA) problem by considering all the mentioned features. To the best of our knowledge, this work is the first attempt to design a decision support system (DSS) that captures the essential events to the maximum possible extent. This is a brief overview of our contribution:

Conceptualize and design a DSS for judicious audits to manage social and environmental compliance in the supplier base. The DSS should cover these features:

- auditing cost
- the idea of the likelihood of suppliers being sustainable
- sustainability in the supplier base (probability of a supplier being sustainable)
- supplier's expected corrective action cost
- demand sensitivity to lack of sustainability in the supplier base (i.e., % market loss)
    - *Marginal* market loss that applies separately to non-compliant suppliers
    - *Total* market loss that applies to part of the buyer's total profit

Develop two new ratios that measure the willingness of buyers and suppliers to perform the audit and possible corrective action, respectively, and use it in our sensitivity analysis.

Develop two new performance metrics besides the main variables of interest (i.e., profit):

- The Portion of Suppliers Selected with Auditing (PSSA)
- Sustainability Index (SI)

Develop an innovative framework for extracting insights from a DSS model based on a sensitivity analysis that relies on regression analysis and marginal effects (ME).

The rest of this article is structured as follows. Section 2 reviews relevant SSOA literature and related stylized modeling papers, distinguishing work in this article from previous work. In Section 3, we present and describe the proposed mathematical models for the problem under study.

Section 4 sets out our experiment design and the analysis framework. Finally, in Section 5, we discuss the key results of our study and suggest a range of directions for future studies.

2.2 Literature review

This section contains a comprehensive literature review of the sustainable-supplier selection and order-allocation problem. We start by defining the spectrum and emphasis of our literature review. This study focuses more on the sustainability aspects of this problem and its connections with other elements of the model. As a result, we omit papers on supplier collection and traditional order allocation. For reviews of those topics, we recommend that readers see Wetzstein et al. (2016) and Pasquale et al. (2020).

2.2.1 Methodology

Buyer-supplier partnerships have been mostly modeled using stylized models and mathematical models. Plambeck and Taylor (2016) and Caro et al. (2018) use stylized models to extract insight from abstract models. However, papers on mathematical modeling such as Trapp and Sarkis (2016) and Moheb-Alizadeh and Handfield (2018) provide practical models for use in real-world contexts. The emphasis of this study is on mathematical modeling articles, and we will only briefly discuss (in Section 2.4) some stylized modeling papers in Section 2.4 that proposed and modeled sustainability auditing.

2.2.1.1 Structure (bi-phase versus integrated)

We classify SSSOA papers according to their structure into two categories: bi-phase, and integrated. This refers to the process of selecting/allocating or evaluating suppliers in terms of sustainability/allocation that happens concurrently (integrated) or independently in two phases (bi-phase). In bi-phase papers, two strategies are used. According to some studies, such as Govindan and Sivakumar (2016), MCDM approaches are used to conduct a pre-qualification procedure for sustainability and then simply allow the green supplier to be included in the final pool. Essentially, each supplier must meet a threshold in terms of their aggregated sustainability score or satisfy the minimum requirements of all sustainability criteria. In these streams of articles, there is usually no sustainable component to the allocation phase (e.g., Step 2), and sustainability is included indirectly in the problem. However, other studies such as Ghadimi et al. (2018) and Kellner et al. (2019) have implemented MCDM methods to evaluate the suppliers' indicators, including economic, social, and environmental criteria.

The output of the MCDM method, which is an aggregated score, is frequently used and modeled in the objective functions of proposed mathematical models.

The term “integrated” here implies that there is no idea of sustainability-related supplier assessment or pre-qualification. The sustainability component of these papers is often expressed mathematically in terms of the objective function, constraints, or a combination of the two. For instance, Moheb-Alizadeh and Handfield (2019) integrated CO_2 emissions into the objective function as a dimension of their model's sustainability.

2.2.1.2 MCDM approaches

In nature, the supplier selection problem is an MCDM problem (Kazemi et al., 2014). As a result, there are numerous papers that explain how various MCDM approaches were used and developed, as well as some criteria for supplier selection in general. MCDM methods were mostly used in three phases of the reviewed papers: criteria selection, weight determination, and supplier selection/ evaluation. A significant number of reviewed papers used MCDM approaches for at least two of the applications listed.

a) Criteria selection

Vahidi et al. (2018) examined the problem of SSSOA under disruption and operational risk as an example for use of MCDM methods in the criteria selection. They used the most important sustainable criteria by means of a hybrid SWOT-QFD method in combination with a DEMATEL technique. Babaee Tirkolaee et al. (2020) identified the most critical criteria and sub-criteria in their studied problem using an approach that integrates Fuzzy Analytic Network Process (FANP) and fuzzy DEMATEL.

b) Weight determination

Analytical hierarchy process (AHP) and its fuzzy version (i.e., FAHP) are by far the most widely used MCDM method that is used to aggregate expert’s preferences in calculation procedure of the importance weight of considered criteria and sub-criteria (Kanan et al. 2013; Mohammed et al. 2018; Chowdhury et al. 2020; Mohammed et al. 2021). Rezaei (2016) developed the Best-Worst-Method (BWM), a relatively new MCDM technique for determining the importance weights of criteria based on a systematic pairwise comparison that is more efficient and reliable than AHP in terms of the number of pairwise comparisons needed. This method is

employed as a weight determination method in so many recently published SSSOA studies such as Cheraghalipoura and Farsad (2018); Lo et al. (2018) and Li et al. (2021).

c) Supplier selection/evaluation

Supplier selection/evaluation is by far the widely used application of MCDM methods in reviewed papers. The Technique for Ordering Preferences by Similarity to the Ideal Solution (TOPSIS) and its fuzzy counterpart (i.e., Fuzzy TOPSIS or FTOPSIS) have been widely used to assess suppliers' compliance with sustainability standards (See Mohammed et al. 2018; Babaee Tirkolaee et al. 2020 and Li et al. 2021). Each study used different criteria and different set of weights for supplier evaluation process. Some papers considered just environmental and/or society based criteria in the evaluation process. For example, Keshavarz Ghorabaee et al. (2017) considered seven environmental criteria including Environmental pollution, Resource consumption, Ecological innovation, Environmental management system, Commitment of managers to environmental improvements, Using green technologies in production processes and Using green materials in production processes. The second stream usually consider either of environmental or social-based criteria besides economic related measures. For example, Govindan and Sivakumar (2016) considered cost, quality, delivery, recycle capability and GHG emissions as main criteria to proxy the sustainability score of suppliers. Finally, the last stream incorporates all three aspects of sustainability (i.e., economic (conventional), environment and social). Mohammed et al. (2019) followed this pattern and considered ten sub-criteria and all of these three aspects of sustainability.

As the MCDM approach is not used in this work, we did not mention all the details of these studies and recommend the readers to refer to Batista Schramm et al. (2020) who reviewed 82 papers of sustainable supplier selection, published between 1990 and 2019.

2.2.1.3 Single-objective versus multi-objective approaches

Total cost minimization is by far the most frequent objective function in the literature. Majority of reviewed paper in the SSSOA domain modeled their problems as multi-objective optimization problem. Multi-objective models were found to have a range of two to seven objective functions in the literature. The majority of these papers, however, have either two or three objective functions. Numerous papers addressed the order allocation problem by taking into account the objective function-specific characteristics of the supply chain for which the model was

developed; for example, some papers used green and sustainable criteria such as environmental impact, recycled waste, CO2 emissions, and energy use per product, or criteria to mitigate the impact of disruption risks, or the resilience level of the supply chain.

These papers used different approaches to handle the multi-objective issue. In general, we can categorize these papers to two groups. First stream employs an approach to convert a multi-objective problem to a single objective problem. These approaches are including Simple Additive Weighting method (Govindan and Sivakumar, 2016), weighted comprehensive criterion (Hamdan and Cheaitou, 2017), ε-constraint method (Kellner and Utz, 2019), augmented ε-constraint method (Azadnia et al. 2015; Vahidi et al. 2018), goal programming (Jia et al. 2020), min-max method (Ozgen et al. 2008) and augmented min-max method (Lo et al. 2018).

The second stream has a framework that produce Pareto solutions and select the final solution at the end of optimization process usually using an MCDM method. Govindan et al. (2015) studied a multi-objective sustainable supply chain network design with stochastic demand. They proposed a hybrid metaheuristic algorithm to tackle the problem. Their solution approach provided Pareto solutions and they compared those solutions with previous algorithms in terms of multi-objective performance measures. They concluded that the developed method achieves better solutions compared with the others. Moheb-Alizadeh and Handfield (2019) developed a mixed integer linear model that considers multiple products, multiple periods and multiple transportation modes. Their model takes account for shortage and discount conditions besides carbon emission as the sustainability aspect of the paper. They developed a hybrid algorithm based on Benders decomposition and produced Pareto solutions at first place. Then, they employed a DEA super efficiency model to select the final solution of the model.

There are just a few studies that modeled SSSOA as a single objective problem. Aktin and Gregin (2016) used a questionnaire for measuring the sustainability score of the suppliers and used that information as an input of their models. They developed mixed integer linear programming models that aimed to allocate the demand to the most sustainable suppliers while ensuring the minimum purchasing cost. Trap and Sarkis (2016) developed a mathematical model that considers the sustainability in the forms of suppliers' ratings and the supplies' investment and training on sustainability issues. They provided an algorithmic approach to identify the high quality and diverse solutions. The objective of their model was to maximize the sustainability score of selected suppliers.

Li et al. (2021) proposed a bi-phase framework in which the first phase involves selection of qualified suppliers that meet the risk threshold and the second stage takes care of order allocation. The objective function of their proposed model was to minimize the total cost which are including production cost, transportation cost, delay cost, penalty cost and carbon emission cost.

2.2.2 Sustainability aspect

A noticeable fraction of reviewed papers incorporate sustainability as a score that shows the compliance degree of the supplier with respect to sustainability code of conducts. We refer this approach to "*Score-based*" in this paper. (For example, see Gupta et al., 2016; Hamdan and Cheaitou, 2017). We observed two approaches to have this scores. The first approach assumes that the organizations can evaluate suppliers in terms of sustainability criteria using eco-design criteria or auditing or industry related sustainability standards (Trap and Sarkis, 2016). The second stream, which is dominant, employ MCDM approaches to evaluate suppliers in terms of all sustainability criteria and present the aggregated score of the MCDM approach as the sustainability score of the suppliers. Then, they use this score as an input of the order allocation mathematical model. For example, Mohammed et al. (2019) considered ten criteria in three categories of conventional (cost, quality, delivery reliability and technology capability), green (environmental management system, waste management and pollution production) and social (safety, rights and health of employees, staff development and information disclosure). There are two review papers that focused on the green/sustainable supplier selection problem and we refer the readers to these papers to get more information on the variety of the criteria that considered in sustainability evaluation of the suppliers (Genovese et al. 2013; Zimmer et al. 2015).

The second most frequent sustainability aspect we observe in SSSOA literature is minimization of CO_2 Emission or greenhouse gas emission (GHGE) emission (For example, see Moheb-Alizadeh and Handfield, 2019; Jia et al., 2020). In terms of decision variable, there are just two papers that include sustainability as one of their main decision variables. For example, Govindan and Sivakumar (2016) considered recycle order quantity as the sustainability decision variable of the model besides the order quantity for the raw material. Govindan et al. (2020) studied the SSSOA in closed-loop supply chain which considers multi-product, multi-depot and green vehicle routing problem. The sustainability aspect of their model were the decision to build the disposal and recycling centers as well as the amount shipped from collection center to disposal and

recycling centers. Moreover, they minimized the total fuel consumption of the used vehicles and the total emission of greenhouse gases.

2.2.3 Supply risk and market-based implications

Park et al. (2018) proposed a new framework that uses multi-attribute utility theory to identify the supplier regions based on four regional sustainability indexes regional sustainability indices that is annually reported by authorized agencies. These indexes are including Global Competitiveness Index, Global Enabling Trade Index, Ease of Doing Business Index, and Logistics Performance Index. In second phase of their framework, they proposed a mathematical model with four minimization objectives including the number of defective modules and components, the delivery delay, the supply chain cost and total carbon footprint.

Vahidi et al. (2018) studied the SSOA problem under the operational and disruption risks by considering the resilience and sustainability criteria. They proposed a bi-objective two-stage possibilistic-stochastic programming model to tackle this problem. In first phase, they calculated the sustainability and resilience score of suppliers. Then, they used both of these scores in an aggregated function to maximize the sustainability and resiliency of the supplier base simultaneously. The second objective function of their model is to minimize the total expected cost which is including contract, purchasing and shipping costs with the main and backup suppliers and the expected cost for the undelivered orders based upon the disruption. They set a threshold for GHG emission caused by transportation and a penalty cost is incurred by any supplier that exceeds this limit.

Wong (2020) proposed a model to evaluate suppliers in terms of static and dynamic performances. This study categorize the market customers in terms of sustainability demand to three types including green, inconsistent and red consumers. This market segmentation integrated with supplier's dynamic risk is included in modelling the SSOA problem. Moreover, a threshold is set for order cost, acceptable level of trend value, risk, green consensus among suppliers and desirable greenness level. Market bonus applies to suppliers that have higher green score compared to the desirable level and market penalty applies to suppliers that cannot satisfy the market in terms of product's green level. In total, seven fuzzy objective functions are considered and modeled and solved by fuzzy goal programming approaches. This study demonstrated how different types of green consumers can affect companies' green goals and how the green customers are interested in converting their green beliefs to actual consumption.

2.2.4 Auditing and corrective action

One stream of research has used stylized modeling to investigate how a variety of auditing approaches or the degree of auditing affect a supplier's compliance with a code of conduct. Plambeck and Taylor (2016) consider a buyer-supplier setting to investigate how increased auditing pressure may motivate the supplier to exert higher effort to pass the buyer's audit by hiding information, instead of improving the compliance or fixing the social and environmental violations in the supplier's facility. Caro et al. (2018), Fang and Cho (2020), and Chen et al. (2020) extend the framework of Plambeck and Taylor (2016) to study the efficacy of various auditing approaches (e.g., independent, shared, or joint audit) to push a supplier for exerting a higher level of compliance effort. Caro et al. (2018) first study the independent audit-penalty approach in which the two non-competing buyers conduct their respective audits and impose penalties independently. They then develop two analytical frameworks to evaluate the impact of jointly-auditing a common supplier or sharing results of one's independent audit with the other buyer on the common supplier's compliance. Fang and Cho (2020) develop a model based on a co-operative game in partition function form, enabling them to investigate firms' competitive and co-operative interactions in a market. They consider coalitions of competing buyers sourcing from a common supplier wherein buyers in each coalition either jointly audit the supplier or audit the supplier independently and share the outcome of auditing with buyers in the coalition. Chen et al. (2020) model a scenario with two identical buyers and three suppliers. Each buyer sources two components, with one of them being from the common supplier and the other from a non-common supplier. Considering a budget-constrained setting, they study whether buyers prioritize the common supplier's auditing effort instead of their respective non-common suppliers (i.e., whether to focus on the supplier's centrality). Rikhtehgar Berenji et al. (2020) investigate the role of commitment to contract terms in conjunction with auditing to improve the supplier's compliance with the code of conduct. They find that increasing the degree of commitment to contract terms (i.e., a priori commitment to only price, only quantity, or both) enhances both the supplier's compliance with the code of conduct and overall sustainability compliance in the marketplace. Chen et al. (2020) develop a game-theoretical model to investigate the effect of supplier-auditor collusion on the auditing and contracting strategy. By looking into the cost versus collusion elimination trade-off between a third-party audit and an audit done by the buyer, they also offer explanations for why buyers rely on third-party audits and set higher compliance bar for suppliers

who may commit more social and environmental violations. This research stream only focuses on how different auditing approaches can motivate the supplier to be more compliant with the code of conduct.

To the best of our knowledge, Mazahir and Ardestani-Jaafari (2020) is the most relevant published study to our work. They studied the SSOA problem under the legislation and proposed a two-stage robust optimization model. They defined their problem in a way that a buyer considers some local regulatory requirements associated with environmental/health risks while making sourcing decisions. Two sources of uncertainty, including supplier's compliancy and demand, are considered. However, our proposed model can be differentiated from this work in so many aspects as below:

- Unlike our model, the audit is not a decision variable in their study, and it is treated as a fixed cost. To be more precise, our model allow for a situation in which a buyer can select a supplier and allocate to it even without auditing when beneficial (i.e., judicious auditing)
- Unlike our model, there is no notion of corrective action in their model.
- The supplier compliancy is a binary parameter in their model. i.e., either a supplier is qualified to meet the standard of a specific market or not. However, we define it as a probability ranging from zero to 1 and gives more flexibility to the buyer to make its supplier base more complaint other than meeting the minimum thresholds.
- Unlike our model, there is no notion of market consequences if a supplier fails the audit.

2.3 Problem description

We consider a supply chain wherein a buyer sources a single product from a supplier base with the size I. Each supplier has the unit production cost (c_i), production capacity (m_i), and offers wholesale price (w_i). In the interest of considering a parsimonious framework, we assume that a pre-qualification process has filtered the suppliers who have the desired level of quality, flexibility, lead-time, and other essential criteria from the buyer's perspective. Further, we assume that the demand (D) and market price (p) are exogenous. This model setting is consistent with a classical supplier selection order allocation (SSOA) problem.

We assume that the buyer faces a market that penalizes any suppliers' sustainability misconduct if the buyer sources from non-compliant suppliers. To avoid any penalization, the buyer can audit a given supplier in his selection and order allocation and subsequently incurs the

auditing cost (ac_i). The buyer audits a supplier based on comprehensive instruction and guidelines (i.e., code of conduct) for all aspects of sustainability. We assume a common understanding of what is considered a sustainability violation among the buyer, suppliers, and customers (i.e., market). Note that in contrast with the traditional auditing setting (i.e., "*audit all*" or "*audit none*"), we conceptualize the auditing decision to be judicious, meaning that the buyer can audit even a portion of its supplier base. The buyer always either audits or skips the auditing for the entire order allocated to a selected supplier. The outcome of any audit depends on the compliance degree of the supplier. Hence, there is always a chance of being non-complaint for the suppliers unless they are fully compliant. Similar to Plambeck and Taylor (2016), we define ϕ_i as a degree of compliance for a given supplier that proxies the probability of being compliant with the sustainability code of conduct. We assume that ϕ_i is predictable by the supplier's historical performance and/or getting information from third party agents.

If a supplier passes the audit, the buyer sources from that supplier. However, if a supplier fails the audit, the buyer offers corrective action to resolve the issue. Subsequently, if the supplier can make its reservation profit, the supplier will participate in the corrective action and incur the cost (ca_i). In other words, a supplier will participate in the corrective action if the supplier receives $\gamma\%$ of its anticipated profit. We assume that the corrective action is perfect, which means that the supplier will be fully compliant with the code of conduct if he takes the corrective action.

If the buyer sources from a non-compliant supplier and sells the product to the market, the NGO/media will publicize the non-compliance. Consequently, the buyer's margin will be affected, and the buyer loses $\lambda\%$ of its original profit margin associated with the supplier(s) who was not compliant (marginal loss scenario). This scenario is possible when the buyer has implemented the blockchain technologies.

Figure 2-1 summarizes the sequence of actions for all possible scenarios. When the buyer audits a supplier, two outcomes are possible. With the probability of ϕ_i , the supplier will pass the audit and the buyer sources from that supplier. On the other hand, with the probability of $1-\phi_i$, the supplier fails the audit, and the buyer asks for corrective action. The supplier participates in the corrective action if he makes $\lambda\%$ of its profit.

On the other hand, if the buyer skips the auditing, two outcomes are possible. With the probability of ϕ_i, NGO and media do not catch the non-compliance, and there is no further consequence for the buyer. However, with the probability of $1-\phi_i$, the media and NGOs catch sustainability misconduct and, this affects the buyer's marginal profit, as discussed above.

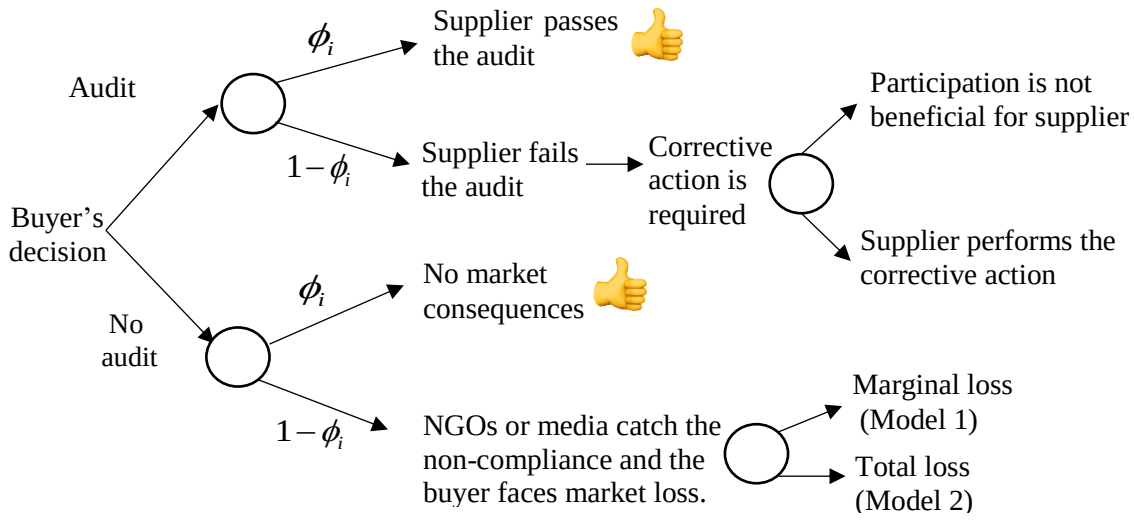

Figure 2-1. Sequence of actions

2.3.1 Model 1 (marginal loss)

Notations used in our developed mathematical formulation are as follows.

Sets

I Set of pre-qualified suppliers $I=\{1,...,I\}$

K Set of auditing decisions $K=\{a,\bar{a}\}$ where a means auditing happens and $\bar{a}$ means skip auditing

D Deterministic demand

c_i The production cost of i_{th} supplier per unit

m_i The production capacity of i_{th} supplier

ac_i The auditing cost of i_{th} supplier

p	The selling price for the buyer	ca_i	The corrective action cost of i_{th} supplier
w_i	The wholesale price of i_{th} supplier	ϕ_i	*A priori* likelihood of i_{th} supplier's compliance
λ	The loss in the margin for the buyer when the supplier's non-compliance is publicized		
γ	The minimum percentage of profit for a supplier to participate in the corrective action		

Indices

i	Supplier index $i \in I$	k	Audit index $k \in K$

Parameters

Decision variables

x_{ia}	The allocated amount of demand to i_{th} supplier with an audit (Integer variable)
$x_{i\bar{a}}$	The allocated amount of demand to i_{th} supplier without audit (Integer variable)

$$y_i = \begin{cases} 1 & \text{if buyer audit } i^{th} \text{ supplier} \\ 0 & \text{otherwise} \end{cases}$$

Auxiliary Decision Variable

$$b_i = \begin{cases} 1 & \text{if any portion of demand is allocated to } i^{th} \text{ supplier} \\ 0 & \text{otherwise} \end{cases}$$

x_{ia}, $x_{i\bar{a}}$ and y_i are the main decision variables and b_i is used to facilitate the modeling. The formulation of our problem is as follows:

$$Max\ Z = \underbrace{\sum_{i=1}^{I} x_{ia}(p - w_i)}_{\text{Part 1}} + \underbrace{\sum_{i=1}^{I} \phi_i\, x_{i\bar{a}}(p - w_i)}_{\text{Part 2}} + \underbrace{(1-\lambda)\sum_{i=1}^{I}(1-\phi_i)x_{i\bar{a}}(p - w_i)}_{\text{Part 3}} - \underbrace{\sum_{i=1}^{I} y_i\ ac_i}_{\text{Part 4}} \quad (1)$$

$$x_{ik} \le y_i\, m_i \qquad \forall i = 1,...,I; k = a \quad (2)$$

$$x_{ik} \le (1 - y_i)\, m_i \qquad \forall i = 1,...,I; k = \bar{a} \quad (3)$$

$$\sum_{k \in K} \sum_{i=1}^{I} x_{ik} = D \quad (4)$$

$$\sum_{k \in K} x_{ik} \leq b_i \, m_i \qquad \forall i = 1, ..., I \qquad (5)$$

$$\sum_{k \in K} x_{ik} > b_i - 1 \qquad \forall i = 1, ..., I \qquad (6)$$

$$y_i \leq b_i \qquad \forall i = 1, ..., I \qquad (7)$$

$$\phi_i (w_i - c_i) x_{ik} + (1 - \phi_i)\left[(w_i - c_i) x_{ik} - y_i \, ca_i\right] \geq \gamma \, x_{ik} (w_i - c_i) \qquad \forall i = 1, ..., I; k = a \qquad (8)$$

$$x_{ik} \geq 0 \qquad \forall i = 1, ..., I; k = a, \bar{a} \qquad (9)$$

y_i and b_i are binary variables $\qquad \forall i = 1, ..., I \qquad (10)$

The objective function of our model (Eq. 1) maximizes the buyer's expected profit. The objective function reflects the expected buyer's profit function wherein the potential market loss due to publicized non-compliance exclusively applies to the portion of demand sourced from the non-compliant supplier(s). As shown in Figure 2-1, the buyer has two auditing choices, either to audit a selected supplier or skip the auditing. Part 1 and part 4 of the objective function represent the buyer's profit and auditing cost, respectively when the buyer chooses to audit any supplier. Parts 2 and 3 represent the buyer's profit and marginal loss due to the supplier's non-compliance, respectively, when the buyer chooses to skip auditing. If the buyer decides to audit, a supplier (i.e. $y_i = 1$), the buyer incurs the auditing cost (part 4). Subsequently, the buyer guarantees collecting its profit (part 1) without facing any loss due to supplier's non-compliance. Part 2 represents the buyer's expected profit when the buyer skips auditing, and the supplier turns out to be compliant; thus, there is no loss. Mathematically, it is the summation of the portion of maximum achievable revenue per supplier. Note that part 2 depends on each supplier's compliance degree (i.e., ϕ_i), which proxies the probability of being compliant with the code of conduct. Finally, part 3 represents the buyer's expected profit when the buyer skips the auditing, and the supplier is not compliant. In this case, the buyer faces market disruption and loses λ% of the portion of demand that was sourced without auditing from the non-compliant supplier.

Eq. 2 and 3 ensure two issues simultaneously. First, the auditing decision is mutually exclusive, meaning that any single supplier is either being audited or not, given its financial advantage on the buyer's side. In other words, when the audit occurs, the buyer always audits the entire order allocated to a supplier, not a portion of that. Second, they guarantee that the allocated

order to any given supplier is less than or equal to its production capacity. Eq. 4 ensures that the allocated amount to all suppliers with or without auditing equals the total demand. Constraints 5 and 6 are added to force b_i to take zero if the allocation is not made and take one if the allocation with or without auditing to any single supplier is a value between zero to its production capacity. In other words, we track the suppliers who received an order with b_i to enforce another restriction in the next constraint. Eq. 7 ensures that the buyer does not audit a supplier unless allocation is made to that specific supplier. Eq. 8 is the supplier's participation constraint. It shows that a supplier who fails an audit participates in the corrective action if the supplier's expected profit is more than the supplier's reservation profit (i.e., above γ% of his expected total profit). Finally, Eq. 9 and 10 determines the variable bounds.

2.3.2 Model 2 (total loss)

The parameters and sets used to formulate model 2 are as same as Model 1. Model 2 has two more auxiliary variables, two extra constraints, and its objective function is different from Model 1. To save space, we just mention the differences and do not repeat the similar components in Model 2. As mentioned earlier, in this model, the market reaction to sustainability failure applies to the total profit due to the invisibility of the source of failure. This is carried out by multiplying a ratio that is defined based on the sustainability compliance of selected suppliers and auditing decisions.

Decision variables

x_{ik} and y_i are the main decision variables here, the same as in Model 1. Please refer to the definitions provided for these two variables in Model 1.

Auxiliary Variables

b_i used as a positive allocation indicator, the same as in Model 1.

Ψ_i Sustainability assurance probability of i^{th} supplier

Ω Sustainability assurance probability of the entire supply chain

x_{ia}, $x_{i\bar{a}}$ and y_i are the main decision variables, and b_i, Ψ_i and Ω are used to facilitate the modeling. The formulation of Model 2 is as follows:

$$Max \ \underbrace{\Omega \sum_{i=1}^{I} (x_{ia} + x_{i\bar{a}})(p - w_i)}_{\text{Part 1}} - \underbrace{\sum_{i=1}^{I} y_i \ ac_i}_{\text{Part 2}} + \underbrace{(1-\lambda)\left[\sum_{i=1}^{I} (x_{ia} + x_{i\bar{a}})(p - w_i)\right](1-\Omega)}_{\text{Part 3}} \quad (11)$$

(Eq 2-10)

$$\Psi_i = \phi_i b_i + (1-\phi_i) y_i + (1-b_i) \quad (12)$$

$$\Omega = \prod_{i=1}^{I} \Psi_i \quad (13)$$

As same as Model 1, the buyer wants to maximize the overall expected profit. Here, we formulate the buyer's profit in such a scenario that the market loss for a probable failure in the sustainability issue applies to the total profit, which means the market does not have visibility over the source of that product. We modeled the objective function of the buyer using three parts, as can be seen in Eq. 11. Part 1 of Eq. 11 calculates the expected revenue for the portion of demand that is risk-free in terms of market violations of sustainability issues. Part 2 shows the total auditing cost needed to pay to make any supplier risk-free. Part 3 demonstrates the expected revenue for the portion of demand that was sourced by a supplier that is not fully compliant (i.e. $\phi_i \neq 1$) or was not audited by the buyer (i.e., $b_i = 1 \text{and } y_i = 0$).

We elaborate on this objective function more by two hypothetical cases. In case 1, the buyer decides to audit all selected suppliers, and in case 2, the buyer chooses to skip auditing for at least one of the selected suppliers. If case 1 happens, the sustainability risk-freeness probability of selected suppliers takes the value of 1 (i.e., $\Omega = 1 \text{which in turn leads to} (1-\Omega) = 0$), which means he guarantees zero future market loss for himself by paying the auditing costs for all selected suppliers. In other words, Part 3 of the objective function becomes inactive. To be more exact, the supplier base is guaranteed to be compliant; therefore, there is no risk. Now suppose the buyer decides to skip auditing for at least one of the selected suppliers (i.e., case 2). In this hypothetical case, all parts of the objective function are active, and the model proposes the buyer to audit the suppliers if it makes the buyer better off after paying the auditing cost.

We call Eq. 12 an expression that calculates the *Risk-freeness probability of any given supplier*. In other words, it is the chance of a supplier being complaint based on buyers' decisions. To be more specific, there are three possibilities that we modeled by this constraint:

Case 1: When a supplier is not selected for sourcing

In this case, first, b_i will be forced to take zero because no allocation with or without auditing is made (i.e., Eq.5 and 6). Then, y_i is forced to take zero by Eq. 7. Now, we have all information (main decision variables) needed to calculate Ψ_i. By plugging $y_i = b_i = 0$ into constraint 10, we have $\Psi_i = 1$, which indicates that this supplier is risk-free because it does not come to the supplier base to have any share in the total risk.

Case 2: When a supplier is selected for sourcing with auditing

In this case, first, b_i will be forced to take one because allocation with auditing is made (i.e., Eq.5 and 6). Then, y_i takes one since the buyer decides to audit this specific supplier. If we plug $y_i = b_i = 1$ into constraint 10, we have $\Psi_i = 1$, which indicates that this supplier is risk-free due to auditing, which is as same as case 1. However, the buyer should incur the auditing cost to achieve the full risk-freeness for this supplier.

Case 3: When a supplier is selected for sourcing without auditing

In this case, first, b_i will be forced to take one because allocation without auditing is made (i.e., Eq.5 and 6). Then, y_i takes zero since the buyer decides to skip auditing this supplier. By plugging $b_i = 1$ and $y_i = 0$ into constraint 10, we have $\Psi_i = \phi_i$, which indicates that this supplier is not necessarily risk-free anymore. However, there is a special case in which ϕ_i is equal to one itself, meaning that this supplier showed 100% sustainability compliance before.

Recall there was a multiplier in Part 1 and 3 of the objective function for model 2.We call that multiplier, as an expression that calculates *Risk-freeness probability for the entire supply* based on buyers' decisions in Eq. 13. In other words, it is the chance of entire supply being complaint based on buyers' decisions. It is worth noting that Ω can take its maximum value (i.e., one) when the buyer decides to audit all of its selected suppliers.

2.4 Scenario Design, Analysis, and Results

We present the details of our research methodology in this section. First, we elaborate on the numerical analysis by introducing the parameters, variable factors, and levels. Then, we discuss the scenario development and problem generation procedure. Next, we discuss the scenario

analysis for both models and finally we elaborate on the steps we took to implement the average marginal analysis.

2.4.1 Scenario development

Note that this study is motivated by examining the effect of specific parameters on the buyer's auditing decision (i.e., judicious audit). Hence, we focus on capturing the impact of changing these parameters and their interaction on the performance metrics. We identify the first group of parameters as factors of interest that include the loss in the margin (i.e., λ), *a priori* likelihood of sustainability compliance (i.e., ϕ_i), auditing cost (i.e., ac_i), and corrective action (i.e., ca_i). The second group of parameters is demand (i.e., D), production capacity (i.e., m_i), price (i.e., p), wholesale price (i.e., w_i), production cost per unit (i.e., c_i), and the minimum percentage of profit for any supplier to participate in corrective action (i.e., γ). To be able to provide the managerial insights, we further generalize our model by developing two new ratios which are substituted with ac_i and ca_i, respectively: (1) Auditing Cost Ratio (i.e., ACR_i) and (2) Corrective Action Ratio (i.e., CAR_i). We present these two ratios as follows:

$$ACR_i = ac_i/(p - w_i)m_i \qquad \forall i = 1,...,I \tag{14}$$

$$CAR_i = ca_i/(w_i - c_i)m_i \qquad \forall i = 1,...,I \tag{15}$$

ACR_i is the ratio of the buyer's auditing cost over its profit. This metric proxies the buyer's profitability degree when the buyer chooses to audit a supplier. CAR_i represents the ratio of cost of the supplier's corrective action over its profit. This metric proxies the supplier's profitability degree when the supplier fails the audit and is asked to undertake the corrective action.

To develop scenarios for our analysis, we consider different values for the above-mentioned parameters. First of all, we set $I = 10$ (i.e., size of supplier base), $p = 100$ (market price), $D = 2400$ (demand in the market) and $\gamma = 0.8$. We use the uniform distribution (shown in Eq. 16) to generate random values given an average for any parameter that could be potentially different across suppliers in the supplier base.

$$\text{U(Average value} \times (1 - 0.15)\text{, Average value} \times (1 + 0.15)\text{)} \tag{16}$$

For example, for the suppliers' production capacity, we set the average capacity of the supplier base at 400 and randomly generate ten values by U($400\times(1-0.15), 400\times(1+0.15)$). We consider the average value for other parameters as follows: $\overline{m}=400$, $\overline{w}=50$ and $\bar{c}=25$. Similar to the above-mentioned process, we generate random values for the factors of interest (ϕ_i, ACR_i and CAR_i). Note that we cannot use Eq.16 to generate random ϕ_i because it creates values higher than one, which is meaningless in our model. Hence, we generate the random values for ϕ_i by a uniform distribution between 0.91 and 0.99. We present the formulas for the average of factors of interests ($\bar{\phi}$, $\overline{ACR}$ and $\overline{CAR}$) as follows:

$$\bar{\phi}=\sum_{i=1}^{I}\phi_i \Big/ I \qquad (17) \qquad \overline{ACR}=\sum_{i=1}^{I}ACR_i \Big/ I \qquad (18) \qquad \overline{CAR}=\sum_{i=1}^{I}CAR_i \Big/ I \qquad (19)$$

To have a parsimonious model and provide managerial insights, we consider two values for factors of interest, which create sixteen (i.e. 2^4) scenarios. Table 2-1 provides all details about the factors and their values.

Table 2-1. Scenario development for factors of interest

	λ	$\bar{\phi}$	$\overline{ACR}$	$\overline{CAR}$
Low value	0.1	0.05	0.025	0.025
High value	0.8	0.95	0.75	0.75

Table 2-2 demonstrates the values used for $\overline{ac}$ and $\overline{ca}$ to get the low and high values of the $\overline{ACR}$ and $\overline{CAR}$ provided in Table 2-1. Recall that the supplier base has ten suppliers, and by generating ten random problems for each combination in Table 2-1, we have 160 problems (i.e., 10×16).

Table 2-2. Auditing and corrective action costs values

	$\overline{ac}$	$\overline{ca}$
Low value	500	15000
High value	250	7500

2.4.2 Performance metrics

In this work, the buyer's profit is considered as main objective. We normalized the profit value for each model separately by dividing all optimal values (i.e., optimal profit) to the maximum profit value among 160 observations to make the analysis unitless and be able to generalize our insights to any case with different parameters. It's worth noting that obtaining the global optimal solution in model 2 is not guaranteed due to the model's highly nonlinear nature.

To evaluate the buyer's judicious auditing, we introduce a couple of performance measurements. The first one is the buyer's profit. It represents the buyer's financial performance under different scenarios. The next measurement is the proportion of suppliers selected with auditing (PSSA), and it is shown in Eq. 20. $PSSA$ can get any values between zero and 1. $PSSA$ equal to zero means that the buyer's optimal decision is to skip auditing the entire supplier base. $PSSA$ equal to one means that the buyer should audit all of its selected suppliers, and any value for $PSSA$ which is between zero and one implies partial (judicious) auditing.

$$PSSA = \sum_{i=1}^{I} y_i \Big/ \sum_{i=1}^{I} b_i \tag{20}$$

Eq. 21 shows the next performance measurement, which is the sustainability index (SI). The sustainability index is the weighted proportion of demand sourced from the complaint suppliers. This measurement is a proxy to capture the sustainability score of the selected suppliers in the supplier base, including the chance of market loss. The maximum value that SI takes is one, and it happens when the buyer audits all of its selected suppliers. However, to calculate the lower bound for the SI, suppliers should be sorted in descending order based on ϕ_i . Then, the maximum allocation should be made from the supplier that has the lowest ϕ_i and move on to the next supplier until the demand is fulfilled.

$$SI = \sum_{i=1}^{I} x_{ia} + \sum_{i=1}^{I} \phi_i \, x_{i\bar{a}} \Big/ D \tag{21}$$

We use a commercial solver to solve mixed-integer programming models (presented in the previous section) and find the optimal solutions. All above-mentioned scenarios are solved on a machine with a 4th Gen Intel® Core™ i7 processor and 16GB RAM. Then, we recorded both objective function value (i.e., profit) and the optimal solutions. Further, we calculate *PSSA*, and SI

based on the optimal solutions. In the next subsection, we present our methodology to derive insights from scenarios discussed in this section and their associated optimal solutions.

2.4.3 Benefit of judicious audits over naïve actions

In this part, we numerically show the benefit of judicious audits vis-à-vis naïve strategies (auditing none (AN) or auditing all (AA) selected suppliers– case by case for 160 randomly generated problems when possible. Table 2-3 demonstrates optimal PSSA value and % Average Improvement over Naïve policies (AION) for both models. To calculate AION, we find the optimal profit value for naïve policy of each case. Then, compare the base model's optimal profit with the optimal value naïve policy. As we see in Table 2-3, the naïve policy for 7 case of all 16 cases are almost clear in buyer's point of view. These cases are including 1, 5, 7, 9, 10, 13 and 14.

In Model 1, as indicated in Table 2-3, improvement happens in 3 cases out of 16 possible combinations (Case#1, 5 and 7) and the judicious auditing has no benefit over blanket policies for four cases out of these seven cases. The results demonstrate an improvement for 22 out of 160 problems, which represents 13.75% of all cases. This shows the capability of our model compared with blanket policies. It should be mentioned that the best profit value we get for the naïve policy of case 5 is negative (i.e., loss). It turned out that this improvement has the highest value (i.e., about 92.5%) when the supplier base is non-complaint, corrective action is in favor of suppliers (i.e., very cheap), the market is unforgivable, and the buyer has to pay a high auditing cost. Now, suppose the buyer has unidimensional thinking and decide to perform or skip auditing based on a single parameter. When the supplier base is non-complaint, the buyer is prone to audit all of the selected suppliers. However, the results say the opposite thing (i.e., "Audit None") for 50 out of 80 problems. Now, suppose the buyer decides solely based on the auditing cost. So, the buyer decides to skip the audit for all of the selected suppliers when the auditing cost is high. The results confirm this decision for 62 problems (%.77.5). Nevertheless, in 18 problems (22.5%), we see full or partial audit as the buyer's optimal decision.

In Model 2, we can see this improvement for 20 out of 160 cases, about 12.5 % of all possible cases. Moreover, we see the optimal decisions made by our model dominates unidimensional strategies similar to Model 1. The other difference we observe between Model 1 and 2 is that the partial auditing cases happened in different cases for Model 1 (Case# 1,3,7,11,12) and 2 (Case# 10 and 11). Table 2-4 shows detailed information about the number of Audit Al (AA), Audit None (AN), and Audit Partial (AP) of selected suppliers for the cases that, on average

does are not zero or one (partial auditing). The difference we see between Model 1 and Model 2 in this table is that the results for Model 2 are quite robust, and there is no variation for all partial cases (both 100%). This shows that the buyer's decision for any single supplier depends on his decision regarding the rest of the selected suppliers. However, this is not the case in forgivable markets, and the buyer can decide about any single supplier independently.

Table 2-3. Average PSSA and the average improvement of our model compared to naïve actions

Case #	Factors of Interest				Model 1			Model 2		
	$\bar{\phi}$	λ	$\overline{ACR}$	$\overline{CAR}$	PSSA		%AION	PSSA		%AION
					Optimal	Naïve		Optimal	Naïve	
1	L	L	L	L	0.932	**AA**	**31.004**	1.000	**AA**	0.000
2	L	L	L	H	0.000	-	-	0.000	-	-
3	L	H	L	L	0.986	-	-	1.000	-	-
4	L	H	L	H	0.000	-	-	0.000	-	-
5	L	L	H	L	0.000	**AA**	*	0.000	**AA**	*
6	L	L	H	H	0.000	-	-	0.000	-	-
7	L	H	H	L	0.767	**AA**	**92.495**	1.000	**AA**	0.000
8	L	H	H	H	0.000	-	-	0.000	-	-
9	H	L	L	L	0.000	**AN**	0.000	0.652	**AN**	**1.707**
10	H	L	L	H	0.000	**AN**	0.000	0.652	**AN**	**1.707**
11	H	H	L	L	0.607	-	-	1.000	-	-
12	H	H	L	H	0.579	-	-	1.000	-	-
13	H	L	H	L	0.000	**AN**	0.000	0.000	**AN**	0.000
14	H	L	H	H	0.000	**AN**	0.000	0.000	**AN**	0.000
15	H	H	H	L	0.000	-	-	0.000	-	-
16	H	H	H	H	0.000	-	-	0.000	-	-

Table 2-4. The detailed information about auditing decision for partial auditing cases

	Case #	$\bar{\phi}$	λ	$\overline{ACR}$	$\overline{CAR}$	AA	AN	AP
Model 1	1	L	L	L	L	60%	0	40%
	3	L	H	L	L	90%	0	10%
	7	L	H	H	L	20%	0	80%
	11	H	H	L	L	0	10%	90%
	12	H	H	L	H	0	10%	90%
Model 2	9	H	L	L	L	0	0	100%
	10	H	L	L	H	0	0	100%

2.4.4 Scenario analysis for Model 1

In this section, we summarize the main effects of interest factors on the objective function. Then, we examine all possible interaction effects of these parameters on all three performance measures (Profit, PSSA, and SI). We simplify the Eq.1 and call it Eq. 22 to explain the main effects of interest factors.

$$Max\ \ Z = \underbrace{\sum_{i=1}^{I}(x_{ia}+x_{i\bar{a}})(p-w_i)}_{\text{Part 1}} \ \underbrace{-\lambda\sum_{i=1}^{I}(1-\phi_i)x_{i\bar{a}}(p-w_i)}_{\text{Part 2}} \underbrace{-\sum_{i=1}^{I}y_i\ ac_i}_{\text{Part 3}} \tag{22}$$

Note that Eq. 22 is increasingly linear in ϕ_i and decreasingly linear in λ and ac_i. If we follow the same approach for Eq. 8 and rearrange it, we can observe that an increase in corrective action cost may decrease the solution space for the allocation with auditing (x_{ia}), which means the objective function does deteriorate or does not get affected.

As mentioned earlier, moving from low to the high value of interest factors negatively affects the profit function except ϕ_i. For that reason, we select two contradictory factors as the main variables of our plots to better visualize the interaction effects. We have sixteen unique combinations for the factors of interest and three metrics. We explain each of these combinations for all metrics separately and skip explaining those which have the same pattern.

In Figure 2-2, we analyze the interaction effects of $\bar{\phi}$ and $\overline{ACR}$ on metrics while keeping λ and $\overline{CAR}$ Low. When $\bar{\phi}$ is high, normalized profit is higher no matter what the value of auditing cost is. This implies that the market loss value is not high enough to persuade the buyer to audit suppliers, and the buyer skips auditing in this scenario. Figure 2-2B confirms that $PSSA$ is zero, which means the buyer is worse off by skipping the audit. Figure 2-2B shows that the audit does make more sense for the buyer when the supplier base is less complaint, and auditing cost is low. The other observation we see from Figure 2-2B is that the high auditing cost does not cause any predictability power for $PSSA$ when we move from low to high $\bar{\phi}$. One might argue that when λ , $\overline{ACR}$ and $\overline{CAR}$ are low, no matter what $\bar{\phi}$ is the buyer would be worse off if he audits all of the supplier base. However, this is not always true, and there are some cases in which buyers prefer to skip auditing because the probable market loss is less than total auditing costs. In Figure 2-2C, as we move from low to high $\bar{\phi}$, we expect to have higher SI . However, this is not the case when auditing cost is low because the buyer prefers to skip auditing, as we explained before.

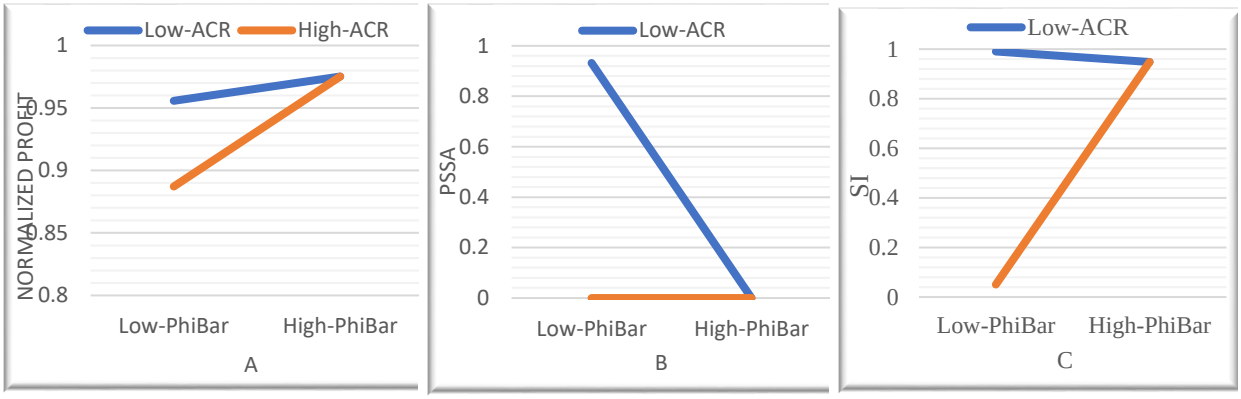

Figure 2-2. Model 1: Interaction effects of $\bar{\phi}$ and $\overline{ACR}$ on metrics while keeping λ low and $\overline{CAR}$ low

Figure 2-3 deals with the interaction effects of $\bar{\phi}$ and $\overline{ACR}$ on metrics while keeping λ low and $\overline{CAR}$ high. One might expect that when λ and $\overline{ACR}$ are in favor of the buyer (i.e., have low values), the buyer will audit all supplier base to guarantee sustainability compliance and avoid future market loss. Although the latter argument makes sense, it ignores the suppliers' willingness to participate in corrective action. As we see in Figure 2-3, there are cases in which suppliers do not prefer to participate in corrective action, and the buyer does not have the auditing option anymore. That's why we do not observe any difference between low and high $\overline{ACR}$.

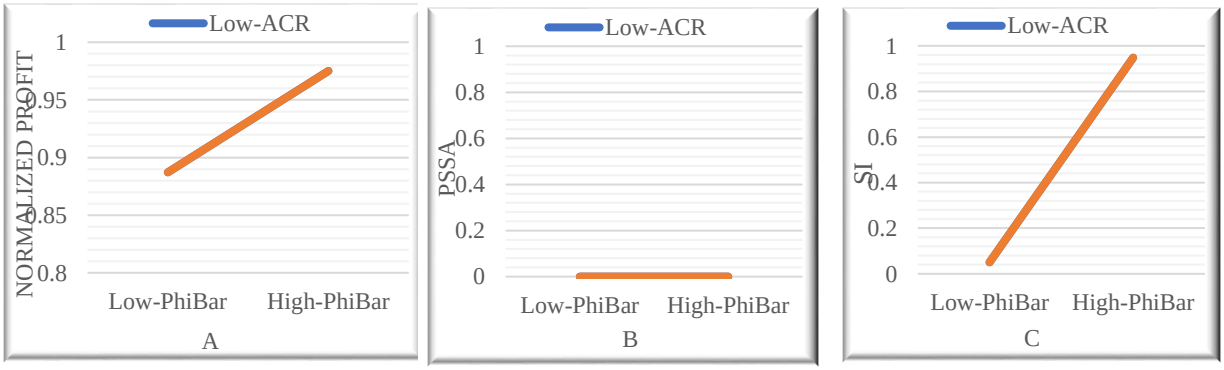

Figure 2-3. Model 1: Interaction effects of $\bar{\phi}$ and $\overline{ACR}$ on metrics while keeping λ low and $\overline{CAR}$ high

Figure 2-4 deals with the interaction effects of $\bar{\phi}$ and $\overline{ACR}$ on metrics while keeping λ high and $\overline{CAR}$ low. In Figure 2-4A, as we move from a non-compliant supplier base to an almost compliant supplier base, we do not see a meaningful difference in profit when auditing cost is low.

One might expect that when λ is high (i.e., in favor of buyer) and $\overline{CAR}$ is low (i.e., in favor of suppliers), it would be a win-win strategy for buyer and suppliers if buyer audit all selected suppliers. However, Figure 2-4B contradicts this argument and verifies it only for one scenario when both the supplier base is non-complaint and auditing costs are low.

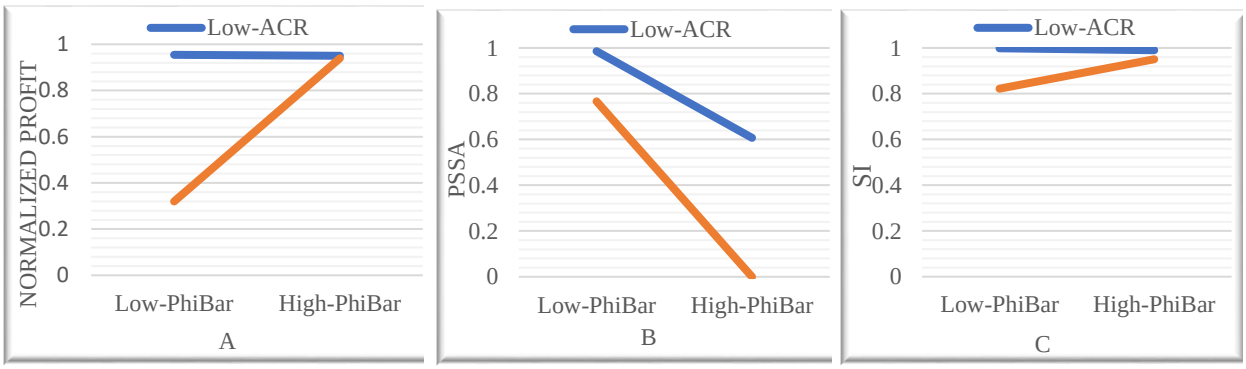

Figure 2-4. Model 1: Interaction effects of $\bar{\phi}$ and $\overline{ACR}$ on metrics while keeping λ high and $\overline{CAR}$ low

Figure 2-5 demonstrates the interaction effects of $\bar{\phi}$ and $\overline{ACR}$ on metrics while keeping λ and $\overline{CAR}$ high. As we move from low to high $\bar{\phi}$, different values of $\overline{ACR}$ do not have a meaningful impact on normalized profit and SI. When λ and $\overline{CAR}$ are high, one might expect that whatever is optimal with high auditing cost holds for low auditing cost, too. In this case, we might expect the buyer skips auditing for low auditing cost to be as same as high auditing cost. However, on average, the buyer prefers to audit a portion of the supplier base, which might happen because the unwillingness of suppliers to participate in corrective action limits the buyers' hand for auditing.

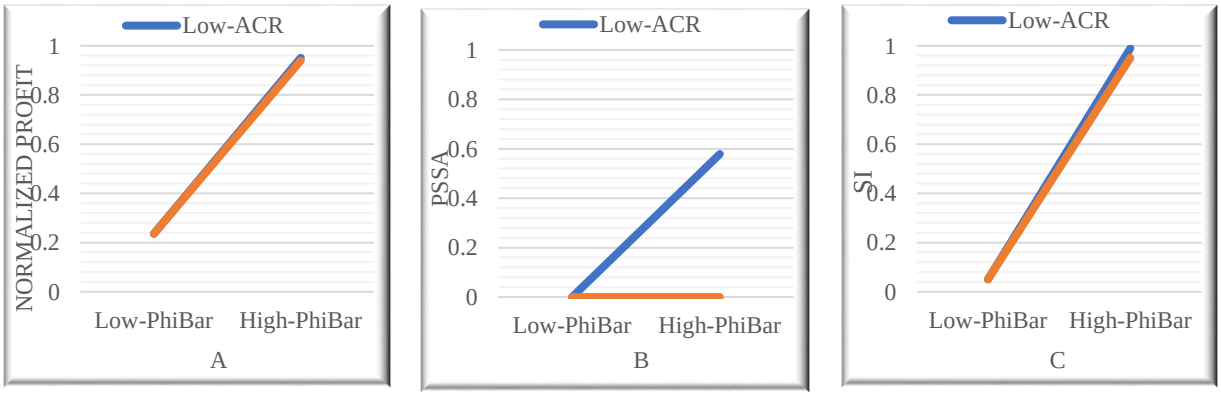

Figure 2-5. Model 1: Interaction effects of $\bar{\phi}$ and $\overline{ACR}$ on metrics while keeping λ high and $\overline{CAR}$ high

2.4.5 Scenario analysis for Model 2

In this section, we summarize the main effects of interest factors on the objective function of Model 2. Then, we focus on the interaction effects of these parameters that do not have the same behavior and pattern similar to Model 1. We simplify Eq. 11 and call it Eq. 23 to explain the main effects of interest factors.

$$Max\ \ Z = \underbrace{(1-\lambda+\lambda\Omega)\sum_{i=1}^{I}(x_{ia}+x_{i\bar{a}})(p-w_i)}_{\text{Part 1}} - \underbrace{\sum_{i=1}^{I} y_i\ ac_i}_{\text{Part 2}} \tag{23}$$

Note that Eq. 23 is increasingly linear in ϕ_i (See Eq. 12 and 13) and decreasingly linear in λ (as $0 \le \Omega \le 1$) and ac_i. Similar to the approach we used for Model 1, we can show that an increase in corrective action cost may decrease the solution space for the allocation with auditing (x_{ia}), which means the objective function does deteriorate or does not get affected.

The interaction effects of $\bar{\phi}$ and $\overline{ACR}$ on normalized profit and SI are pretty similar to Model 1 for cases in which we keep λ and $\overline{CAR}$ in the following combinations respectively: (Low-Low), (Low, High) and (High, High). However, we observe a difference in terms of $PSSA$ for these three combinations when $\overline{ACR}$ is low. As we see in Figure 2-6A, for the (Low-Low) case, we observe a noticeable difference compared to Model 1. As we move from non-compliant supplier base to almost-complaint supplier base, when $\overline{ACR}$ is low, the Model 1's optimal decision is skipping the audit for almost all selected supplier. However, in Model 2, the buyer's optimal decision is auditing about 65% of the selected suppliers. This implies that skipping the audit for unforgivable markets (i.e., Model 2) will have huge consequences for the buyer's profit, and the buyer would be worse off if he audits more in unforgivable markets. In Figure 2-6B, we observe an increase in $PSSA$, when we keep λ low and high $\overline{CAR}$, the Model 1's optimal decision is skipping the audit for almost all selected suppliers. However, in Model 2, when $\overline{ACR}$ is low, the buyer would be worse off if he audits 65% of the selected suppliers. The rationale behind this event is similar to the previous case. In Figure 2-6C, when we keep λ and $\overline{CAR}$ high, and the $\overline{ACR}$ is low , we observe that the buyer's optimal decision is to audit all selected suppliers. However, in Model 1 and on average, about half of selected suppliers would be audited.

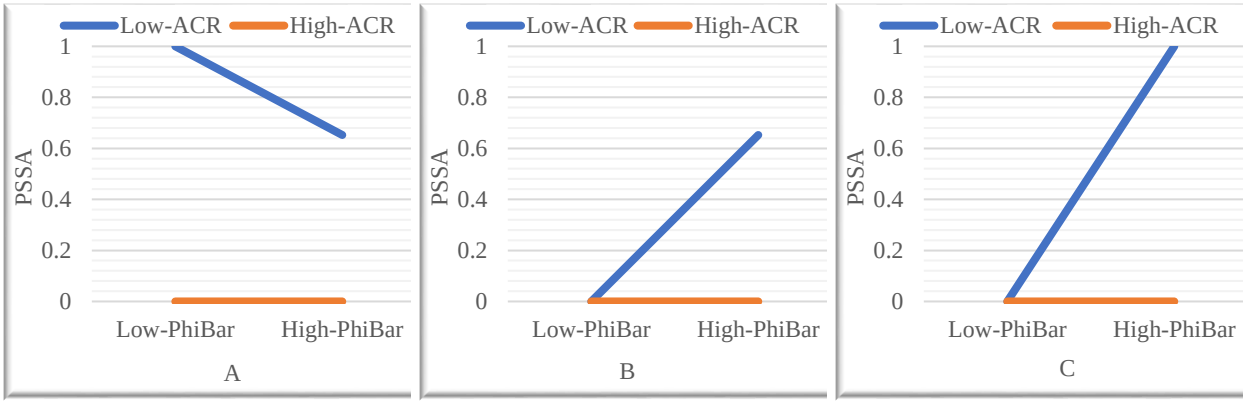

Figure 2-6. Model 2: Interaction effects of $\bar{\phi}$ and $\overline{ACR}$ on metrics while keeping λ low and $\overline{CAR}$ low

There is one case that the interaction effects of $\bar{\phi}$ and $\overline{ACR}$ are different across all three metrics, and this case happens when we keep both of λ and $\overline{CAR}$ high. As we see in Figure 2-7A, when $\overline{ACR}$ is high, Model 2's results are pretty similar to Model 1. However, in low $\overline{ACR}$, change in $\bar{\phi}$ does not affect the buyer's profit because in Model 2, the buyer is more inclined to audit the selected suppliers to prevent future market loss, and the difference in profit in these two cases is just the auditing cost, which is very low and does not drastically change the profit function. The difference between these two models for $PSSA$ and SI is explainable by the same rationale we had for profit.

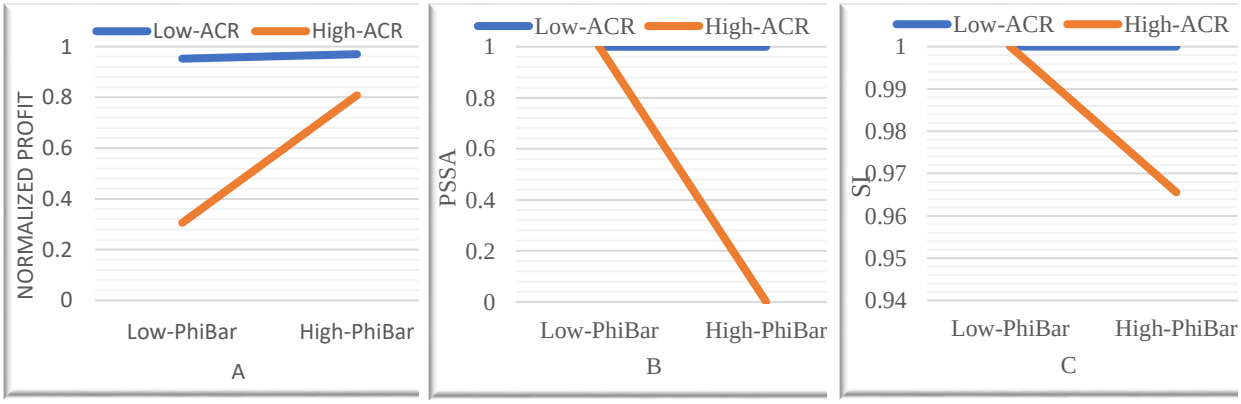

Figure 2-7. Model 2: Interaction effects of $\bar{\phi}$ and $\overline{ACR}$ on metrics while keeping λ high and $\overline{CAR}$ high

2.4.6 Average marginal effect

Note that we have created 160 problems and solved them in the previous section. Now we use the data (optimal solutions and given parameters) from these problems, and we fit a model to find a good representative. Recall that the process of picking the correct line for this model is called "fitting". There are different ways to do this – least squares is possibly the most used one. To do so, we use a multiple linear regression (MLR) with all second order, third order, and fourth-order terms. We have three metrics of interest, i.e., normalized profit, *PSSA*, and *SI*. Therefore, we utilize the MLR model three times for each of these metrics. Note that the process is the same for Model 1 and 2. Eq. 24 shows the functional form used for all of these three models. Z in this equation (dependent variable) can be substituted by profit, $PSSA$ or SI. Note that to run the regression and compare different problem sets together, we normalize the optimal profit to a value from zero to one, using the highest profit in the solution sets.

$$Z = \beta_0 + \underbrace{\beta_1\bar{\phi} + \beta_2\lambda + \beta_3\overline{ACR} + \beta_4\overline{CAR}}_{\text{First order effects}} + \underbrace{\beta_5\bar{\phi}\lambda + \beta_6\bar{\phi}\,\overline{ACR} + \beta_7\bar{\phi}\,\overline{CAR} + \beta_8\lambda\,\overline{ACR} + \beta_9\lambda\,\overline{CAR}}_{\text{Second order effects}} +$$
$$\beta_{10}\,\overline{ACR}\;\overline{CAR} + \underbrace{\beta_{11}\bar{\phi}\lambda\,\overline{ACR} + \beta_{12}\bar{\phi}\,\overline{ACR}\;\overline{CAR} + \beta_{13}\bar{\phi}\,\lambda\overline{CAR} + \beta_{14}\lambda\,\overline{ACR}\;\overline{CAR}}_{\text{Third order effects}} + \underbrace{\beta_{15}\bar{\phi}\,\lambda\,\overline{ACR}\;\overline{CAR}}_{\text{Fourth order effects}} \quad (24)$$

Next, we introduce average marginal effects (AME) to derive insights from our models. Generally speaking, taking the derivative is one of the most informative ways of explaining the fitted results. Basically, it calculates the change in the dependent variable (i.e., profit, $PSSA$, or SI) for a small change in the covariate (i.e., $\bar{\phi}$, λ, $\overline{ACR}$ and $\overline{CAR}$). For each model, we compute the AME of a metric (i.e., profit, $PSSA$, or SI) with respect to a factor of interest (i.e., $\bar{\phi}$, λ, $\overline{ACR}$ and $\overline{CAR}$) in the low or high values of other factors. Note that average marginal effect calculates the expected change of a metric for a change in a factor, given other factors of interest is given at a certain point. For example, suppose we are interested in capturing the impact of a small change in sustainability compliance on normalized profit (Z) while loss in the margin (λ), average auditing cost ratio $(\overline{ACR})$ and average corrective action cost ratio $(\overline{CAR})$ all get the lowest defined value. This translates to $(\partial(Z)/\partial(\bar{\phi})\,|\;\lambda = 0.1, \overline{ACR} = 0.025, \overline{CAR} = 0.025)$. After calculating the MLR function of Z for any observation in our model, we form Eq. 24 and take the derivative of this function with respect to $\bar{\phi}$. Eq. 25 illustrates the outcome of the analysis. Note

that we plug in 0.1, 0.025 and 0.025 for the terms λ, $\overline{ACR}$ and $\overline{CAR}$, respectively (i.e., in multipliers of β_5, β_6, β_7, β_{11}, β_{12}, β_{13} and β_{15}).

$$E(\partial(Z)/\partial(\bar{\phi}) \mid \lambda = 0.1, \overline{ACR} = 0.025, \overline{CAR} = 0.025) = \beta_1 + 0.1\beta_5 + 0.025\beta_6 + 0.025\beta_7 + 0.0025\beta_{11} + 0.000625\beta_{12} + 0.0025\beta_{13} + 0.0000625\beta_{15}$$

Table 2-5 demonstrates the AME in all performance metrics for a tiny change in $\bar{\phi}$, λ, $\overline{ACR}$. Here, we try to highlight some interesting observations from this table. As we move from forgivable to sensitive market (i.e., from Model 1 to 2), we observe no deterioration in AME of all performance measures for a tiny change in $\bar{\phi}$. In the next three paragraphs, we focus on the important observations for each performance measures.

As per normalized profit, the highest positive difference between model 2 and 1 is 0.03 which occurs in cases 4, 11 and 12. The common feature of these three cases is the low value of $\overline{ACR}$. The highest negative difference between model 2 and 1 in terms of profit is 0.03 which occurs in cases 23 and 24. This happens when supplier base is almost complaint and market loss is high. In model 1 and among all possible scenarios of a tiny change in factors of interest, we see that a tiny change in $\bar{\phi}$ has the most profitable change while all other factors are in their high values (i.e., Case 8). On the other hand, the worst possible consequence for profit by a tiny change in corrective action is -0.99 loss and this happens while $\bar{\phi}$ and $\overline{ACR}$ are in their low values and market is unforgivable (case 13).

2.5 Conclusion and Future Research

Supplier selection order allocation is a developing field that has captivated the interest of researchers over the last few decades. Customers are increasingly demanding social and environmental considerations, and handling the buyer-supplier relationship has become much more complicated. As a result, the supplier base's sustainability compliance, consumer expectations about sustainability concerns, and its penalty mechanism to punish non-compliant buyers/suppliers have added a new layer to the conventional supplier selection order allocation problem. To address the new challenge, new models and insights are needed to assist supply chain managers in taking some of these complicated issues into account.

Table 2-5. AME of all performance measures with respect to parameters of interest

				$\partial(Z)/\partial(\bar{\phi})$		$\partial(PSSA)/\partial(\bar{\phi})$		$\partial(SI)/\partial(\bar{\phi})$	
Case#	λ	$\overline{ACR}$	$\overline{CAR}$	M1	M2	M1	M2	M1	M2
1	L	L	L	0.02	0.02	-1.04	-0.39	-0.05	-0.01
2	L	L	H	0.03	0.03	-0.96	-0.31	0.02	0.07
3	H	L	L	0.00	0.02	-0.42	0.00	-0.01	0.00
4	H	L	H	0.05	0.08	-0.35	0.08	0.06	0.07
5	L	H	L	0.03	0.03	-0.96	-0.36	0.02	0.06
6	L	H	H	0.03	0.03	-0.90	-0.29	0.09	0.13
7	H	H	L	0.04	0.06	-0.45	-0.08	0.00	0.00
8	H	H	H	0.09	0.11	-0.38	0.00	0.07	0.07
				$\partial(Z)/\partial(\lambda)$		$\partial(PSSA)/\partial(\lambda)$		$\partial(SI)/\partial(\lambda)$	
	$\bar{\phi}$	$\overline{ACR}$	$\overline{CAR}$	M1	M2	M1	M2	M1	M2
9	L	L	L	0.00	0.00	0.08	0.00	0.01	0.00
10	L	L	H	-0.07	-0.07	0.07	0.00	0.01	0.00
11	H	L	L	-0.04	-0.01	0.87	0.50	0.06	0.01
12	H	L	H	-0.04	-0.01	0.86	0.50	0.06	0.01
13	L	H	L	-0.06	-0.06	0.15	0.10	0.09	0.09
14	L	H	H	-0.12	-0.12	0.14	0.09	0.08	0.09
15	H	H	L	-0.04	-0.02	0.81	0.46	0.06	0.01
16	H	H	H	-0.04	-0.02	0.80	0.46	0.06	0.01
				$\partial(Z)/\partial(\overline{ACR})$		$\partial(PSSA)/\partial(\overline{ACR})$		$\partial(SI)/\partial(\overline{ACR})$	
	$\bar{\phi}$	λ	$\overline{CAR}$	M1	M2	M1	M2	M1	M2
17	L	L	L	-0.09	-0.08	-1.29	-1.38	-1.30	-1.31
18	L	L	H	-0.09	-0.07	-1.20	-1.28	-1.21	-1.22
19	H	L	L	0.00	-0.01	0.00	-0.90	0.00	-0.06
20	H	L	H	0.00	-0.01	0.00	-0.90	0.00	-0.06
21	L	H	L	-0.88	-0.89	-0.30	0.00	-0.24	0.00
22	L	H	H	-0.82	-0.83	-0.28	0.00	-0.23	0.00
23	H	H	L	-0.01	-0.22	-0.84	-1.38	-0.05	-0.05
24	H	H	H	-0.01	-0.22	-0.83	-1.38	-0.05	-0.05

				$\partial(Z)/\partial(\overline{CAR})$		$\partial(PSSA)/\partial(\overline{CAR})$		$\partial(SI)/\partial(\overline{CAR})$	
	$\bar{\phi}$	λ	$\overline{CAR}$	M1	M2	M1	M2	M1	M2
25	L	L	L	-0.09	-0.08	-1.29	-1.38	-1.30	-1.31
26	L	L	H	-0.09	-0.07	-1.20	-1.28	-1.21	-1.22
27	H	L	L	0.00	0.00	0.00	0.00	0.00	0.00
28	H	L	H	0.00	0.00	0.00	0.00	0.00	0.00
29	L	H	L	-0.99	-1.04	-1.36	-1.38	-1.31	-1.31
30	L	H	H	-0.93	-0.98	-1.34	-1.38	-1.29	-1.31
31	H	H	L	0.00	0.00	-0.04	0.00	0.00	0.00
32	H	H	H	0.00	0.00	-0.04	0.00	0.00	0.00

This paper further enhances and examines the problem of sustainable supplier selection and allocation of orders in order to get closer to real world issues. In all facets of the supply chain, including consistency, flexibility and lead times, the suppliers available to meet the order are

supposed to have crossed the buyer's threshold. In conventional order allocation literature, the audit is either overlooked or regarded as a fixed expense, which means that suppliers are audited when chosen. In this study, however, we implement a judicious audit that means, if it is found useful, that the buyer can also save the audit. In other words, the auditing decision is distinct from the selection-allocation decision. We simulate situations in which the supplier is apprehended by the media/NGOs. Furthermore, our model takes into account situations in which the supplier fails the audit and is pressured by the buyer to conduct the corrective action that she deems beneficial.

We model the sustainable supplier selection order allocation problem depending on the market customer's attitude regarding the sustainability compliance of in two versions. These two versions are including total market loss and marginal market loss which are modeled as MILP and MINLP, respectively. These models aimed to maximize the buyer's expected profit. This paper has the following contributions, which are 1) Conceptualizes and designs the model-based DSS for the allocation of order to sustainable suppliers to manage the social and environmental compliance 2) The notion of marginal loss and total loss in supply chain literature is proposed and implemented for the first time 3) Develop two new ratios that are highly anticipated to measure the buyer's ability to conduct audits and suppliers to take corrective measures 4) Develop two new ratios which can calculate the willingness of buyers to perform audits and suppliers to carry out corrective action 5) Developing a practical and innovative framework to extract insights from any data-driven based sensitivity analysis based on regression analysis, marginal effects, and interaction effects.

Here, we summarize briefly some of the insights from the models we have developed. We show that, as predicted, a more compliant supplier base provides the buyer with benefit. The findings also indicate that the ability of the buyer to audit decreases as the audit cost ratio rises. It can nevertheless be shown that buyers tend to audit even their entire supplier base to avoid the possibility of likely failure in sustainability. Another interesting insight from our model occurs when the buyer has a supplier base that is almost entirely non-compliant. One might argue that the buyer would be better off auditing the entire supplier base in this situation. However, our findings indicate that there are some instances where the market loss is insufficient to justify auditing, even though the auditing cost is low. We discover that the corrective action ratio has little predictive power, mostly in cases where the supplier base is almost compliant due to the buyer's decision to avoid auditing, in which case corrective action is irrelevant. Additionally, there are instances where

buyers choose to audit the majority of the supplier base, recognizing that the majority of the supplier base would be more likely to engage in corrective action. These situations typically arise when the supplier base is nearly non-compliant and the cost of corrective action is minimal. If the supplier base becomes more complaint, the percentage of audited suppliers declines regardless of market loss value. It's worth noting that the pace at which this decrease occurs is significantly higher when the market loss is lower. Our findings suggest that when both auditing and corrective action costs are minimal, the sustainability index takes the maximum possible value (nearly one). This occurs as a result of the buyer and supplier's activities being in sync. The buyer often opts out of auditing the supplier base when the market loss is small and the auditing expense is large.

Despite these novel features and contributions, this study has modeling and methodological weaknesses that can be discussed in future studies. For instance, in this study, we assume that the market loss is exogenous and that the buyer is fully aware of the type and value of the market loss. However, this presumption can be violated in practice. As a result, a new study will consider market loss as a function of the buyer's audit decisions, the sustainability degree of selected suppliers, and selection-allocation decisions. Another potential area of future study is to incorporate suppliers' market power into the model. This may also be considered a customer's perception of the supplier's brand. Competition and collaboration amongst major suppliers might also be studied for future research in this field. This model is limited to a single buyer. However, it would be worthwhile to analyze the effect of multiple buyers on suppliers' decision-making processes. Another intriguing extension for future research is to explore scenarios in which the buyer will contribute to corrective action process in instances where the supplier is unable to engage in the contract. Finally, another possibility is to loosen the assumption of deterministic values for the model and instead use stochastic programming methods to deal with the ambiguity associated with a few of the model's parameters.

References Cited

Aktin, T., & Gergin, Z. (2016). Mathematical modelling of sustainable procurement strategies: three case studies. *Journal of cleaner production*, 113, 767-780.

Azadnia, A. H., Saman, M. Z. M., & Wong, K. Y. (2015). Sustainable supplier selection and order lot-sizing: an integrated multi-objective decision-making process. *International Journal of Production Research*, 53(2), 383-408.

Caro F, Chintapalli P, Rajaram K, Tang CS (2018) Improving supplier compliance through joint and shared audits with a collective penalty. *Manufacturing & Service Operations Management* 20(2):363–380.

Chen, L., Yao, S., & Zhu, K. (2020). Responsible sourcing under supplier-auditor collusion. Manufacturing & Service Operations Management, 22(6), 1234-1250.

Cheraghalipour, A., & Farsad, S. (2018). A bi-objective sustainable supplier selection and order allocation considering quantity discounts under disruption risks: A case study in plastic industry. *Computers & Industrial Engineering*, 118, 237-250.

Chowdhury, M. M. H., Paul, S. K., Sianaki, O. A., & Quaddus, M. A. (2020). Dynamic sustainability requirements of stakeholders and the supply portfolio. *Journal of Cleaner Production*, 255, 120148.

Di Pasquale, V., Nenni, M. E., & Riemma, S. (2020). Order allocation in purchasing management: a review of state-of-the-art studies from a supply chain perspective. International Journal of Production Research, 58(15), 4741-4766.

Elkington, J. 1998. Cannibals with Forks: The Triple Bottom Line of 21st-Century Business. New Society, Stony Creek, CT.

Fang, X., & Cho, S. H. (2020). Cooperative approaches to managing social responsibility in a market with externalities. *Manufacturing & Service Operations Management*, 22(6), 1215-1233.

Genovese, A., Lenny Koh, S. C., Bruno, G., & Esposito, E. (2013). Greener supplier selection: state of the art and some empirical evidence. *International Journal of Production Research*, 51(10), 2868-2886.

Ghadimi, P., Toosi, F. G., & Heavey, C. (2018). A multi-agent systems approach for sustainable supplier selection and order allocation in a partnership supply chain. *European Journal of Operational Research*, 269(1), 286-301.

Ghorabaee, M. K., Amiri, M., Zavadskas, E. K., Turskis, Z., & Antucheviciene, J. (2017). A new multi-criteria model based on interval type-2 fuzzy sets and EDAS method for supplier evaluation and order allocation with environmental considerations. *Computers & Industrial Engineering*, 112, 156-174.

Govindan, K., & Sivakumar, R. (2016). Green supplier selection and order allocation in a low-carbon paper industry: integrated multi-criteria heterogeneous decision-making and multi-objective linear programming approaches. *Annals of Operations Research*, 238(1-2), 243-276.

Govindan, K., Jafarian, A., & Nourbakhsh, V. (2015). Bi-objective integrating sustainable order allocation and sustainable supply chain network strategic design with stochastic demand using a novel robust hybrid multi-objective metaheuristic. *Computers & Operations Research*, 62, 112-130.

Govindan, K., Mina, H., Esmaeili, A., & Gholami-Zanjani, S. M. (2020). An integrated hybrid approach for circular supplier selection and closed loop supply chain network design under uncertainty. *Journal of Cleaner Production*, 242, 118317.

Gupta, P., Govindan, K., Mehlawat, M. K., & Kumar, S. (2016). A weighted possibilistic programming approach for sustainable vendor selection and order allocation in fuzzy environment. The *International Journal of Advanced Manufacturing Technology*, 86(5), 1785-1804.

Hamdan, S., & Cheaitou, A. (2017a). Dynamic green supplier selection and order allocation with quantity discounts and varying supplier availability. *Computers & Industrial Engineering*, 110, 573-589.

Hamdan, S., & Cheaitou, A. (2017b). Supplier selection and order allocation with green criteria: An MCDM and multi-objective optimization approach. *Computers & Operations Research, 81*, 282-304.

Ho, W., Xu, X., & Dey, P. K. (2010). Multi-criteria decision making approaches for supplier evaluation and selection: A literature review. *European Journal of operational research*, 202(1), 16-24.

Jia, R., Liu, Y., & Bai, X. (2020). Sustainable supplier selection and order allocation: Distributionally robust goal programming model and tractable approximation. *Computers & Industrial Engineering*, 140, 106267.

Kannan, D., Khodaverdi, R., Olfat, L., Jafarian, A., & Diabat, A. (2013). Integrated fuzzy multi criteria decision making method and multi-objective programming approach for supplier selection and order allocation in a green supply chain. *Journal of Cleaner production*, 47, 355-367.

Kazemi, N., Ehsani, E., & Glock, C. H. (2014). Multi-objective supplier selection and order allocation under quantity discounts with fuzzy goals and fuzzy constraints. *International Journal of Applied Decision Sciences, 7*(1), 66-96.

Kellner, F., & Utz, S. (2019). Sustainability in supplier selection and order allocation: Combining integer variables with Markowitz portfolio theory. *Journal of Cleaner Production*, 214, 462-474.

Kellner, F., Lienland, B., & Utz, S. (2019). An a posteriori decision support methodology for solving the multi-criteria supplier selection problem. *European Journal of Operational Research*, 272(2), 505-522.

Li, F., Wu, C. H., Zhou, L., Xu, G., Liu, Y., & Tsai, S. B. (2021). A model integrating environmental concerns and supply risks for dynamic sustainable supplier selection and order allocation. *Soft Computing*, 25(1), 535-549.

Lo, H. W., Liou, J. J., Wang, H. S., & Tsai, Y. S. (2018). An integrated model for solving problems in green supplier selection and order allocation. *Journal of cleaner production*, 190, 339-352.

Mazahir, S., & Ardestani-Jaafari, A. (2020). Robust global sourcing under compliance legislation. *European Journal of Operational Research*, 284(1), 152-163.

Mohammed, A., Harris, I., Soroka, A., Naim, M., Ramjaun, T., & Yazdani, M. (2021). Gresilient supplier assessment and order allocation planning. *Annals of Operations Research*, 296(1), 335-362.

Mohammed, A., Setchi, R., Filip, M., Harris, I., & Li, X. (2018). An integrated methodology for a sustainable two-stage supplier selection and order allocation problem. *Journal of Cleaner Production*, 192, 99-114.

Moheb-Alizadeh, H., & Handfield, R. (2018). An integrated chance-constrained stochastic model for efficient and sustainable supplier selection and order allocation. *International Journal of Production Research*, 56(21), 6890-6916.

Moheb-Alizadeh, H., & Handfield, R. (2019). Sustainable supplier selection and order allocation: A novel multi-objective programming model with a hybrid solution approach. *Computers & industrial engineering*, 129, 192-209.

Park, K., Kremer, G. E. O., & Ma, J. (2018). A regional information-based multi-attribute and multi-objective decision-making approach for sustainable supplier selection and order allocation. *Journal of cleaner production*, 187, 590-604.

Plambeck EL, Taylor TA (2016) Supplier evasion of a buyer's audit: Implications for motivating supplier social and environmental responsibility. *Manufacturing & Service Operations Management*18(2):184–197

Rezaei, J. (2016). Best-worst multi-criteria decision-making method: Some properties and a linear model. *Omega*, 64, 126-130.

Rikhtehgar Berenji, H. Murthy, N., Yang, Z. (2021) Committing to Contract for a Supplier's Social and Environmental Compliance. *Working Paper*.

Schramm, V. B., Cabral, L. P. B., & Schramm, F. (2020). Approaches for supporting sustainable supplier selection-A literature review. *Journal of Cleaner Production*, 123089.

Tirkolaee, E. B., Mardani, A., Dashtian, Z., Soltani, M., & Weber, G. W. (2020). A novel hybrid method using fuzzy decision making and multi-objective programming for sustainable-reliable supplier selection in two-echelon supply chain design. *Journal of Cleaner Production*, 250, 119517.

Trapp, A. C., & Sarkis, J. (2016). Identifying Robust portfolios of suppliers: a sustainability selection and development perspective. *Journal of Cleaner Production, 112*, 2088-2100.

Vahidi, F., Torabi, S. A., & Ramezankhani, M. J. (2018). Sustainable supplier selection and order allocation under operational and disruption risks. *Journal of Cleaner Production*, 174, 1351-1365.

Wetzstein, A., Hartmann, E., Benton Jr, W. C., & Hohenstein, N. O. (2016). A systematic assessment of supplier selection literature–state-of-the-art and future scope. *International Journal of Production Economics*, 182, 304-323.

Wong, J. T. (2020). Dynamic procurement risk management with supplier portfolio selection and order allocation under green market segmentation. *Journal of Cleaner Production*, 253, 119835.

Zimmer, K., Fröhling, M., & Schultmann, F. (2016). Sustainable supplier management–a review of models supporting sustainable supplier selection, monitoring and development. *International Journal of Production Research*, 54(5), 1412-1442.

Chapter 3

A Decision-Support System for Detecting and Handling Biased Decision Makers in Multi-Criteria Group Decision-Making Problems

This is a joint work with Babak Aslani and Dr. Jafar Rezaei. This work is published in *Expert Systems with Applications* (a peer-reviewed journal)

3.1 Introduction

Decision-making, defined as processing the information to determine the best alternative from among a set of possible alternatives, plays a crucial role in personal life and professional interactions in business, industry, and social contexts (Kabak & Ervural, 2017). As the nature of problems is becoming more complex, considering several criteria in decision-making processes is a necessary action. Therefore, Multi-Criteria Decision-Making (MCDM) is making its way to different disciplines as a useful tool for assessing alternatives based on various norms. MCDM is a branch of Operations Research dealing with problems that typically relate to evaluate, select, sort, or rank multiple alternatives that typically involve various (conflicting) criteria. MCDM approach can deal with a range of decision problems to help decision-makers achieve (relatively) consistent and robust solutions (Cinelli et al., 2020).

MCDM problems have been classified in different ways, one of which is based on the number of decision-makers (DM) contributing to the process, which leads to two classes, single decision-maker and group decision-maker. Single DM class is a process in which only a single DM is responsible for defining the problem and assessing the alternatives based on a set of criteria. In the group decision-making class, on the other hand, the opinion of several DMs are collected and aggregated to make the final decision. Nowadays, group decision-making (GDM) plays a crucial role in decision support systems (DSS) and information systems in the presence of ever-increasing complexity of real-life problems. The intensified complexity of the decision-making process in a range of contexts such as health, business, management, and politics, has made stakeholders and decision-makers rely on group wisdom instead of a single individual. The outcome of decision-making problems usually affects a range of stakeholders at different levels with different preferences. So, in a GDM process related to a real problem, several DMs with expertise in various fields are present. These DMs are usually looking for a single final decision

in the process, though (Bouzarour-Amokrane et al., 2015; Q. Dong et al., 2019; F. Jin et al., 2017; Y. Liu et al., 2019).

GDM as a response to the ever-growing complexity of modern environments related to decision-making problems has found its way in industry section, where technical groups make the decisions on designing products, advancing plans and strategies, and in service sector such as healthcare, in which the critical decisions are made by a group of advisors and experts (Ambrus et al., 2009). The GDM process generally starts with identifying the characteristics of the problem, including alternatives, criteria related to the alternatives, and their importance (weights), as well as DMs and stakeholders. During this stage, compromising among DMs to have an agreement over the mentioned features is highly essential. Following that, a score is assigned to each alternative on each criterion, either by the DMs or other sources. Aggregating the opinions of DMs to rank or sort the alternatives or to select the best option is the final stage of the process. In other words, in GDM problems, separate preferences of DMs should be aggregated in a collective and well-structured way to make the final decision (Liao et al., 2018; Xia & Chen, 2015; W. Xu et al., 2019). GDM process usually includes the opinions of DMs who have different backgrounds and knowledge bases. As a result, interpreting and analyzing the preferences of these DMs is a complex task compared to single DM processes (Wittenbaum et al., 2002; Yu et al., 2018).

Particular challenges have been recognized in the literature for the GDM process. One possible shortcoming of GDM is creating a diffusion of responsibility resulting in a lack of accountability for the outcomes of the process. Moreover, group decisions can make it easier for members to deny personal responsibility and blame others for the undesired final decision (Palomares et al., 2014). GDM involves many complex and conflicting aspects intrinsic to human individuality and human nature. For instance, when a team of experts with various (usually conflicting) personal goals and intentions takes part in the process, having disagreement is a natural expectation (Parreiras et al., 2010). As another challenge, because of different individual concerns about alternatives, DMs usually adopt heterogeneous preference representation structures to express their preferences (X.-h. Xu et al., 2019). Experiments related to GDM in different fields show clear contrasts between the results of the GDM process compared to decisions reached by DMs making choices in isolation. DMs in a real decision-making problem may have various fields of expertise, meaning that each DM can have a dissimilar effect on the final decision (Hafezalkotob & Hafezalkotob, 2017).

GDM can be categorized based on its dynamics to cooperative or non-cooperative. Cooperative GDM is significant in engineering, medical, and scientific fields, while non-cooperative GDM is common in economic and political areas. Even in cooperative GDM, reaching a complete agreement among all members of the group on the final solution is nearly impossible since DMs, who are supposed to have similar goals, may have some opinion conflicts, in practice (Y. Dong et al., 2018). Having consistency and consensus are other noteworthy challenges in the GDM process. Consistency is directly related to the credibility of the GDM process. Consensus, on the other hand, means that the majority of DMs accept the final result of the process (Song \& Li, 2019). In the case of political context, the presence of the correlated opinions can pose another challenge to the aggregation of individual opinions. For instance, the opinions of the members associated with a certain political party are expected to be highly correlated. These correlated clusters can be identified and labeled as outliers in a GDM aggregation phase since they do not reflect the true opinions of DMs (parliament members or congressmen in this context), but only represent their political attachments. This specific problem is not in the scope of the current study, but has been tackled in the literature (see Akram et al., 2019; Akram et al., 2020(a); Akram et al., 2020(b); Gonzalez-Arteaga et al., 2016; Liu et al., 2020).

One of the main challenges in the GDM process is having biased DM(s), which can significantly challenge the result of the process in such a way that the stakeholders might not accept the final decision. The biased DM(s) distort the consensus among other DMs. The unbiased DM(s) will not reach an agreement over the results in the presence of biased DM(s). These mentioned challenges show the significance of detecting and managing biased DMs in the GDM process (Kabak & Ervural, 2017; Liao et al., 2019; Mohammadi & Rezaei, 2019; Song & Li, 2019; Yu et al., 2018).

The studies related to group decision-making have handled bias in the process from different points of view. A stream in the literature is mainly focused on identifying sources of bias (see Jones and Roelofsma (2000), Xiao and Benbasat (2018), Ceschi et al. (2019)). These works are mainly concerned with the reason of having bias in the GDM and do not present any solution to resolve or eliminate bias. Adopting consensus reaching process (CRP) is another rich body of knowledge in handling biasedness in GDM (see Bouzarour-Amokrane et al. (2015), W. Liu et al. (2018), W. Xu et al. (2019)). These papers generally present frameworks to improve the consensus among the DMs by closing their preferences (scores) assigned to the alternatives in each criterion.

DMs are even allowed to change their opinions based on receiving feedback in some configurations such as W. Liu et al. (2018). Another research avenue related to bias is identifying the biased DMs (outliers) in the GDM process (see McShane et al. (2013)). Even though this problem has been addressed in a range of contexts such as business in Lee (2014), NBA draft season in Ichniowski and Preston (2017), there is a shortage of efforts to present a general framework for identifying and managing outliers in GDM.

In specific situations, the opinions of biased DMs can be excluded from the final pool of decision-making. For instance, if the aggregated opinion of many DMs is desired (like voting systems), eliminating DMs who represent the minority voices is justified. Likewise, when the collective agreement is acceptable by everyone even those whose opinions have been eliminated, we can rationalize the elimination of DMs in the GDM process. One example is the GDM procedures using agents to collect and aggregate anonymous votes of DMs to find the results with the highest degree of consensus. For these cases, we propose the Extreme Anti-Biased Method (EABM) and Moderate Anti-Biased Method (MABM) version of our proposed method, both of which include elimination and weighting based on the consensus degree of each DM with the remaining pool. However, in the moderated configuration, a threshold of weight is assigned to each DM to mitigate the biasedness in assigning the weights only based on consensus. On the other hand, if all DMs are somehow related to the final results in terms of interest, like stakeholders or board of directors, the opinions of outliers cannot be eliminated easily (or at all). As the DMs observe the consensus reaching process and they do not consent to be eliminated from the GDM process, a different version of the presented framework is designed for these environments. To handle these cases, we propose the Soft Anti-Biased Method (SABM), in which there is no elimination, but the weights are assigned to all DMs based on a combination of a predefined threshold and consensus-based weights.

The main contribution of this study is to present a general statistical-based framework to detect and handle biased DMs in a GDM process. The diverse versions of the proposed framework are capable of providing viable solutions to all the possible scenarios in the aggregation phase of a GDM process. If there is no outlier in the pool, the framework will assign a relative weight to each DM based on the agreement level to the total pool. In the case that some of the DMs are biased, the framework can either detect and exclude them from the decision-making process or incorporate the biasedness in assigning the weights to the DMs. Finally, if there is no common

ground for all DMs, the agent responsible for the aggregation can warn the DMs to adjust their opinions or even incorporate a new set of DMs to the pool to find some common opinions on which she/he can establish the aggregated opinion. Even though detecting and handing outliers in GDM has high importance in various contexts, such as voting systems, supplier section and order allocation, and portfolio optimization, there are not enough works in the literature to tackle the problem.

The remaining of the paper is structured as follows. In Section 2, the literature review, as well as the gap, is presented. The proposed method is explained in detail in section 3. Section 4 includes the result of applying the proposed approach to a numerical example. The result of the method on a large sample of problems and a detailed analysis is provided. Finally, the conclusion and suggestions for possible future research are provided in Section 5.

3.2 Literature Review

As a main challenge in the GDM process, when multiple DMs are involved, it is necessary to aggregate individual judgments into a single representative judgment for the entire group. The studies in this research stream have tried to use different aggregation methods to achieve a collective decision in a well-structured way.

Ordered Weighted Aggregation (OWA) and its extensions as a popular aggregation operator for GDM is present in several works such as Xu and Jian (2007), P. Liu (2011), and L. Jin et al. (2018). Other aggregation procedures has been intensively proposed in various contexts, such as individual judgments and priorities in Forman & Peniwati (1998), group calibration process of slightly different preferences in Rokou & Kirytopoulos (2014), Introducing a representative value function as the aggregated opinion in Kadziski et al., (2013), Stochastic Group Decision Support System (GDSS) for the GDM with uncertainty in Mokhtari (2013), and incomplete individual ratings in group resource allocation decisions in Stengel (2013).The main shortcoming of these operators is that it cannot appropriately operate in many real cases in which the arguments being aggregated have different significance. This issue massively challenges the credibility of the combined aggregated decision in GDM.

The Delphi method is a framework based on the results of multiple rounds of questionnaires sent to a group of experts. The anonymous responses are aggregated and shared with the group after each round. The experts can adjust their opinions in subsequent rounds, based on their perception of the group's opinion. As the Delphi method seeks to reach a desirable

response through consensus, it has practicality in the aggregation stage of GDM. For instance, a novel Delphi method known as agile Delphi was used as an aggregation approach in Xie et al. (2012). They aimed to present a GDM process to rapidly develop and execute the response approach to the unconventional emergency events. Their agile Delphi decision-making platform based on ASP.NET. allowed experts to conduct appropriate alternative portfolio and enables researchers to find a new combination of experts to make the decision-making more efficient. Lack of physical interactions, long response time, and the usefulness of received information from experts are the main drawbacks of this method.

A richer body of knowledge in the literature of GDM contains studies that tried to increase the consistency and consensus in the aggregation phase of the GDM process. Wibowo and Deng (2013) presented an interactive algorithm for consensus building in the group decision-making process. They touched the subjectiveness and the imprecision of the decision-making process by using intuitionistic fuzzy numbers. A decision support system framework was also presented for improving the effectiveness of the consensus building process. Q. Dong and Cooper (2016) proposed a novel consensus reaching model in a dynamic decision environment within the context of the group Analytic Hierarchy Process (AHP). They adopted a Markov chain method to determine the DMs' weights of importance for the aggregation process with respect to the group members' opinion transition probabilities. Their developed model also provided feedback suggestions to adaptively adjust the credibility of each DM in each round so that the dynamic feature of decision-making process is also addressed. Wu and Xu (2016) developed separate consistency and consensus processes to deal with hesitant fuzzy linguistic preference relation (HFLPR) individual rationality and group rationality. They included an easy understandable local revision strategy in their framework. The feedback system was based directly on the consensus degrees to reduce the proximity measure calculations.

F. Jin et al. (2017) proposed two new GDM methods to improve the multiplicative consistency of linguistic preference relation (LPRs) until they are acceptable, and the priority weight vector of the alternatives is derived from adjusted LPRs. They introduced novel concepts of order consistency, multiplicative consistency, and a consistency index in their proposed approach. Other components of their methodology were two linear optimization models to generate the normalized crisp individual and group weight vectors and two GDM methods to help DMs in obtaining reasonable and reliable results. W. Liu et al. (2018) proposed an iteration-based

consensus building framework for GDM problems with self-confident multiplicative preference relations. They developed an extended logarithmic least squares method to derive the individual and collective priority vectors from the self-confident multiplicative preference relations. They used a two-step feedback adjustment mechanism to assist the decision-makers to improve the consensus level. Q. Dong et al. (2019) introduced a novel framework to achieve a potential consensus considering too polarized opinions. By considering the existing dynamic relationships among the experts, they suggested using Social Network Analysis to reach agreement in this specific class of GDM problems. They claimed that the framework is particularly useful when achieving the agreement through CRPs is impossible. Morente-Molinera et al. (2019) presented a novel model for experts to carry out GDM processes using free text and alternative pairwise comparisons and two ways of applying consensus measures over the GDM process. In the context of social networks, they adopted Sentiment analysis procedures to analyze free texts and extract the preferences that the experts provide about the alternatives. They also presented two ways of applying consensus measures over the GDM process.

Song and Hu (2019) developed an iterative algorithm to improve the consistency to obtain the probabilistic linguistic preference relation (PLPR) with satisfactory consistency in large-scale GDM problems. They used PLPR in the presence of partial preference information coming from stakeholders in the GDM procedure. A probability computation model was defined by mathematical programming to derive the missing probabilities of PLPR. Song and Li (2019) developed a GDM model based on obtained multiplicative consistency linguistic preference relations (LPRs) in consideration of the group consensus reaching process. Their proposed mathematical programming method was able to find the incomplete LPR with the highest consistency and increasing inconsistent LPR to complete consistent LPR using multiplicative consistency based on given incomplete HFLPR. Tang et al. (2019) developed a programming model to judge the consistency of multiplicative linguistic intuitionistic fuzzy preference relations (MLIFPRs), and an approach to derive consistent MLIFPRs. They developed a direct consensus framework based on cumulative prospect theory to solve GDM problems with eight kinds of preference representation structures. X.-h. Xu et al. (2019) developed a confidence consensus-based model for large-scale group decision making that provides a novel approach to addressing non-cooperative behaviors. They defined new concepts of collective adjustment suggestion and rationality degree. Then, they combined the rationality and non-cooperation of the adjustment

information to construct the concept of a confidence level to measures the impartiality and objectivity of the adjustment information and for managing non-cooperative behaviors. Y. Liu et al. (2019) constructed a two-step optimization model to solve for the group consensus ideal scheme and its measure value matrix. They adopted Nash's bargaining idea and maximizing group negotiation satisfaction and minimizing system coordination deviation in their developed framework. They measured the closeness degree of decision-maker information and attribute information by using the distance between the group measure matrices of scheme and consensus ideal scheme.

The main challenge of consensus reaching process frameworks is that since they ignore external factors, such as weather conditions, price fluctuations, alternatives and/or experts' availability, they lose their practicality in many real-life settings. Higher time consuming, more time on constant preferences supervision and the higher complexity with respect to dealing with experts' non-cooperative behaviors are notable patterns of CRPs when they are applied on large-scale GDM problems, too.

Handling bias in GDM process is another section of the literature, in which either the source of bias is identified, or the bias is managed by different approaches. One major source of bias is in the data collection phase of GDM, in which DMs assert their preferences (or scores) of alternatives in the defined criteria. Cognitive bias, which is a systematic pattern of deviation from norm or rationality in judgment, and is very common in the GDM process, is addressed in Stewart (2005). Anchoring bias happens when we (DM) think too little our judgments can be skewed by irrelevant information that we happened to see, hear, or think about a moment ago. This source is analyzed comprehensively in the work of Lieder et al. (2018). Jacobi and Hobbs (2007) investigated the splitting bias, which refers to a situation where presenting an attribute in more detail, may increase the weight it receives. For more details on the studies of this stream of literature, we refer to a comprehensive study of Montibeller and Von Winterfeldt (2015), which provides an overview of different cognitive biases at different stages of the decision-making process. Even though these various types of bias are important in the GDM process, we are not going to address them in our framework.

In the current study, we assume that there is bias in the form of biased DM(s) in the GDM process, and we need to identify and handle it in an organized manner. In this section, we focus on papers trying to detect and eliminate the present bias in the GDM problems. There is a limited

number of studies in the literature, in which handling biased DMs in a specific case and context is addressed. For instance, McShane et al. (2013) presented the OntoAgent model for automatic detection of decision-making biases, using clinical medicine as a sample application area. In particular, the intelligent agent was able to follow doctor-patient interaction and warn the doctor if the patient's decisions were being affected by biased reasoning. Bouzarour-Amokrane et al. (2015) proposed a resolution model based on individual bipolar assessment. Each decision-maker evaluates alternatives through selectability and rejectability measures, representing the positive and negative aspects of alternatives. They also included the impact of human behavior (including influence, individualism, fear, caution) on decisional capacity. They adopted consensus building models based on game theory for collective decision problems to achieve a common solution in the GDM process. Y. Dong et al. (2016) proposed a novel consensus framework for managing non-cooperative behaviors in GDM. A self-management mechanism was responsible for generating experts' weights dynamically based on multi-attribute mutual evaluation matrices (MMEMs). During the process, the experts could provide and update their MMEMs regarding the experts' performances (e.g., professional skill, cooperation, and fairness) so that their weights were updated. He et al. (2018) introduced a quantum framework based on quantum probability theory to model the subjectivity in multi-attribute group decision making (MAGDM) when subjective beliefs towards DMs' independence or relations are present. The opinions of DMs are viewed as various wave functions that are occurring at the same time. Then the beliefs will interfere with each other and influence the aggregated result. Classical MAGDM techniques were used for the preparation stage. Then, they constructed a Bayesian network and extended it to a quantum framework. When all the DMs are deemed independent, the quantum framework will degenerate into a classical Bayesian network.

Based on the papers reviewed, a significant gap in the scope of our study, detection, and handling bias in the GDM environment, is identified. The studies addressed bias in the GDM process in the literature are more focused on using various models and techniques to increase the agreement level among DMs, which is an essential aspect of GDM. Although there are some previous works approaching bias detection in individual cases of GDM problems, presenting a general framework to this purpose is a significant gap in the literature. The main contribution of our work is to develop a systematic approach to tackle one crucial challenge in GDM, identifying and handling the outliers (biased) DMs to improve the credibility of the aggregated opinion. To

bridge this gap, we proposed a two-stage statistical approach to identify outliers and modify the final weights of DMs in the decision-making process. The proposed framework in this study can be applied to any GDM problem regardless of size and context. Also, this framework can handle extreme situations without any agreement as well as the partial and total agreement among DMs. The details of the suggested framework will be provided in Section 3.

3.3 Methodology

In this section, we first explain EABM version of our proposed method in detail. To save the space, we only then focus on the differences among the moderate and soft versions and the extreme one.

The EABM for handling bias in GDM is composed of two phases. First, by forming the normalized integrated matrix of scores, the biased DMs are detected and eliminated from the decision-making process. Then, by following a well-defined procedure based on the relative amount of overlap for each DM by other DMs and the total data, a weight value is assigned to each DM. This section includes the detailed process of these steps of the proposed framework. Table 3-1 presents the notations used in the following sections to help the readability of the equations.

Table 3-1. Nomenclature of the proposed framework

Notation	Description
d_i	Decision-maker (DM) i, $i=1,...,\|I\|$
a_j	Alternative j, $j=1,...,\|J\|$
c_k	Criterion k, $k=1,...,\|K\|$
s_{ijk}	Score assigned by d_i to a_j with respect to c_k
ψ_{ijk}	The normalized value of s_{ijk}
$S_k^{\min}$	Minimum value of c_k across all DMs and alternatives or $\min\left[s_{ijk}\right]$ for all i and j
$S_k^{\max}$	Maximum value of c_k across all DMs and alternatives or $\max\left[s_{ijk}\right]$ for all i and j
$\bar{x}_i$	Average of normalized scores of d_i
$\bar{\bar{X}}$	Average of total normalized scores for all DMs
σ_i	The standard deviation of normalized scores of d_i

Table 3-1. (continued)

Notation	Description
σ	The standard deviation of total normalized scores for all DMs
CI_i	Confidence Interval for d_i
UB_i	Upper bound of Confidence Interval for d_i
LB_i	Lower bound of Confidence Interval for d_i
l_i	Length of Confidence Interval for d_i
CI	Confidence Interval for total normalized scores for all DMs
O_{mn}	Overlap among CI_m and CI_n, $m,n \in i$
o_i	Overall overlap for d_i
M_i	Maximum possible value of overlap of d_i with other DMs
α	Significance level for calculating CI's (The range is usually between 0.9 and 0.99)
$\bar{o}_i$	Overlap Ratio for d_i
CI_i	Relative CI for d_i
B_i	Biasedness Index for d_i
B	Threshold for detecting biased DM (The range can be different from 0 to $\lvert I-1 \rvert$)
R	Threshold for the number of eliminated DMs in the first step
N	The number of normalized scores of each DM in the beginning of the GDM process
I	The number of DMs in the initial pool of the GDM process
I'	The number of DMs remained in the pool in the second phase of the framework
λ	The total number scores for all DMs in the second phase of the framework
ω_i	Weight assigned to d_i
$\hat{\omega}_i$	Initial Weight of d_i
$\omega_{\min}$	The minimum predefined weight for MABM and SABM versions
γ	Total predefined weight

Please note that since several DMs from the initial pool might be excluded as outliers, we defined two sets for DMs. The first set is the initial pool and starts from $i = 1,...,|I|$. The pool in the second phase starts from $i = 1,...,|I'|$. Regardless of this change, the initial label of DMs is kept for evaluating the alternatives based on the assigned weights and the associated performance matrix. Please also note that for parameter B, the range can change from 0 to $|I-1|$ and for α, this variation is usually between 0.9 and 0.99 from a statistical point of view.

3.3.1 Outlier elimination phase

3.3.1.1 Gathering data for the decision-making process

Group decision-making process includes a number of decision-makers, $d_1,...,d_I$, who evaluate several alternatives, $a_1,...,a_j$, with respect to some criteria, $c_1,...,c_K$. It is worth mentioning that the score table, which is also called performance matrix, sometimes is available with objective scores, while it is sometimes formed by the evaluations done by the experts in a subjective manner or the combination of both. In this study, we consider the last two cases. The following matrices show the raw data available at the beginning of the process.

$$\begin{array}{r} \\ a_1 \\ d_1 \Rightarrow a_2 \\ \dots \\ a_J \end{array}\begin{array}{c} \begin{array}{cccc} c_1 & c_2 & \dots & c_K \end{array} \\ \begin{pmatrix} s_{111} & s_{112} & \dots & s_{11K} \\ s_{121} & s_{122} & \dots & s_{12K} \\ \dots & \dots & \dots & \dots \\ s_{1J1} & s_{1J2} & \dots & s_{1JK} \end{pmatrix} \end{array} ,...,d_I \Rightarrow \begin{array}{r} \\ a_1 \\ a_2 \\ \dots \\ a_J \end{array}\begin{array}{c} \begin{array}{cccc} c_1 & c_2 & \dots & c_K \end{array} \\ \begin{pmatrix} s_{I11} & s_{I12} & \dots & s_{111} \\ s_{I21} & s_{I22} & \dots & s_{I2K} \\ \dots & \dots & \dots & \dots \\ s_{IJ1} & s_{IJ2} & \dots & s_{IJK} \end{pmatrix} \end{array}$$

3.3.1.2 Creating the integrated decision matrix

As the first step, the opinions of decision-makers should be integrated as a comprehensive decision matrix. Matrix P^1 is comprised of the score of each decision-maker to all alternatives in every criterion. Matrix P^1 is a sample decision matrix for a situation with I decision-makers, J alternatives, and K criteria.

$$P^1 = \begin{array}{r} d_1 \\ \dots \\ d_I \end{array}\begin{pmatrix} s_{111} & s_{112} & \dots & s_{11k} \\ \dots & \dots & \dots & \dots \\ s_{1J1} & s_{1J2} & \dots & s_{1JK} \\ s_{I11} & s_{I12} & \dots & s_{I1K} \\ \dots & \dots & \dots & \dots \\ s_{IJ1} & s_{IJ2} & \dots & s_{IJK} \end{pmatrix}$$

3.3.1.3 Normalizing the integrated decision matrix

In real-world problems, criteria usually have various natures and scales. So, in order to compare these criteria regardless of their units and magnitude, they should be normalized to obtain dimensionless scores. To normalize the scores several methods are developed in the literature like Max, Max-Min, and logarithmic approaches (Vafaei et al., 2016). To convert the scores given to various alternatives, the decision matrix should be normalized by using equation 1 to 4. Equation 1 shows the normalization procedure for criterion with positive nature (like profit, quality, and so on). For negative criterion such as cost and pollution, we can use Equation 2. Equations 3 and 4 show that the minimum value and the maximum value are calculated in each column related to the criterion by $\min\left[s_{ijk} \right]$ and $\max\left[s_{ijk} \right]$, which are the minimum and maximum value among each column of the original decision matrix (Anojkumar et al., 2014). The final output of this normalization process is the matrix labeled as P^1_{norm}.

$$\psi_{ijk} = \frac{S_{ijk} - S_k^{\min}}{S_k^{\max} - S_k^{\min}}, \forall i = 1,..,|I|, j = 1,...,|J|, k = 1,...,|K| \Rightarrow \text{For positive criterion} \quad (1)$$

$$\psi_{ijk} = \frac{S_k^{\max} - S_{ijk}}{S_k^{\max} - S_k^{\min}}, \forall i = 1,..,|I|, j = 1,...,|J|, k = 1,...,|K| \Rightarrow \text{For negative criterion} \quad (2)$$

$$S_k^{\min} = \min\left[s_{ijk} \right], \forall i = 1,..,|I|, j = 1,...,|J| \quad (3)$$

$$S_k^{\max} = \max\left[s_{ijk} \right], \forall i = 1,..,|I|, j = 1,...,|J| \quad (4)$$

$$P^1_{norm} = \begin{matrix} \\ d_1 \\ \\ \\ d_I \\ \\ \end{matrix} \begin{pmatrix} \psi_{111} & \psi_{112} & \cdots & \psi_{11k} \\ \cdots & \cdots & \cdots & \cdots \\ \psi_{1J1} & \psi_{1J2} & \cdots & \psi_{1JK} \\ \psi_{I11} & \psi_{I12} & \cdots & \psi_{111} \\ \cdots & \cdots & \cdots & \cdots \\ \psi_{IJ1} & \psi_{IJ2} & \cdots & \psi_{IJK} \end{pmatrix}$$

3.3.1.4 Reshaping the normalized matrix to a column matrix

To compare the decision-makers in a structured way, the normalized decision matrix should be converted to several columns separately mirroring the scores of each decision-maker to all alternatives in each criterion. This reshaped matrix is shown as matrix P^2.

$$P^2 = \begin{pmatrix} \psi_{111} & \psi_{112} & \cdots & \psi_{11k} \\ \cdots & \cdots & \cdots & \cdots \\ \psi_{11K} & \psi_{21K} & \cdots & \psi_{I1K} \\ \\ \cdots & \cdots & \cdots & \cdots \\ \psi_{1JK} & \psi_{2JK} & \cdots & \psi_{IJK} \end{pmatrix}$$

Having this column matrix, a confidence interval (CI) for each decision-maker should be calculated. Equations 5 and 6 show the required formulas to find the lower bound (LB) and the upper bound (UB) of these intervals for each column (DM). The $\overline{x}_i$ reflects the average scores of d_i assigned to all the alternatives in all criteria, and σ_i shows the standard deviation of the scores for this DM. N is the total number of scores available in the integrated matrix (such as P^2) for each DM, which is the product of the number of alternatives in the number of criteria ($J \times K$). Since we do not know the exact distribution of the scores, t-student distribution is used for calculating the CIs. Three parameters (Thresholds) should be set before further calculation. The first threshold is the acceptable Biasedness level (B). This threshold is a base for comparing the biasedness level of all DMs.

As the first step to calculate the B_i, the CI of all DMs will be compared together. Then, if that DM has overlap with all other DMs, its B_i will be equal to $|I-1|$ and if that DM has no overlap with any DM, its B_i will be equal to zero. Basically, B_i is defined as the number of DMs with having an overlap in their CI's with a specific DM. This threshold (B) can get any values between 0 and $|I-1|$. Choosing a value close to zero means more open to hear extreme voices and closer to $|I-1|$ means more restrictive to hear extreme voices from DMs. The B_i value is calculated for one DM in the numerical example in Section 3 to clarify the procedure. The second threshold is α which is the confidence level of statistical tests. Finally, a threshold R should be defined for the ratio of DMs eliminated in the first phase of the framework. This parameter gives the aggregation agent the freedom to send the DMs warning to review and resubmit their scores or (if possible) to consider new DMs in the process to reach an agreement.

$$\bar{x}_i = \frac{\sum_{j=1}^{J}\sum_{k=1}^{K}\psi_{ijk}}{N},\ \sigma_i = \sqrt{\frac{1}{N-1}\sum_{j=1}^{J}\sum_{k=1}^{K}(\psi_{ijk}-\bar{x}_i)^2}\ ,\forall i=1,...,I \tag{5}$$

$$CI_i = (\bar{x}_i - t(N-1,\alpha)\frac{\sigma_i}{\sqrt{N}}, \bar{x}_i + t(N-1,\alpha)\frac{\sigma_i}{\sqrt{N}}) \tag{6}$$

To find the decision-makers who are biased or outliers in comparison with their counterparts, these CI's should be compared in terms of their B. DMs with lower B_i values than the predefined threshold B are excluded from the process as the output of this phase.

3.3.2 Weight assignment phase

After eliminating the biased DMs in the previous step, an overall comparison should be conducted for the remaining scores in the decision matrix. In this step, a pairwise comparison of CIs for all DMs is conducted to obtain an overlap ratio of DMs. Having these values, the relative CI_i to CI is also calculated to assign a proportionate weight to the remaining DMs. While there is no elimination in this phase, DMs who have more agreement with other DMs and the total opinion are considered more influential by having higher weight values. A similar concept is presented in the work of Bonner et al., (2002), in which the experts with higher expertise receive more weight in the aggregation stage.

3.3.2.1 Overlap ratio calculation

As the first step of this phase, the CIs of the remaining DMs in the pool of GDM are compared to one another. For each DM, the overlaps with CIs of other DMs are accumulated to a value known as overlap ratio. Matrix P^3 shows the remaining scores after eliminating the outliers in the first step. Please note that although the new set of unbiased DMs starts from 1 to $I^{'}$, the initial label of DMs is preserved for the final evaluations.

$$P^3 = \begin{array}{c} \\ d_1 \\ d_2 \\ \dots \\ d_{I^{'}} \end{array} \begin{array}{c} \begin{array}{cccc} c_1 & c_2 & \dots & c_K \end{array} \\ \begin{pmatrix} \psi_{111} & \psi_{122} & \dots & \psi_{1JK} \\ \psi_{211} & \psi_{222} & \dots & \psi_{2JK} \\ \dots & \dots & \dots & \dots \\ \psi_{I^{'}11} & \psi_{I^{'}22} & \dots & \psi_{I^{'}JK} \end{pmatrix} \end{array}, \forall i = 1,...,I^{'} \in \{Unbiased\ DMs\}$$

$$\forall m,n \in \{unbiased\ DMs\} \Rightarrow \begin{cases} if\ CI_m\ has\ overlap\ with\ CI_n \Rightarrow \begin{cases} O_m \rightarrow O_m + O_{mn} \\ O_n \rightarrow O_n + O_{mn} \end{cases} \\ else \Rightarrow \begin{cases} O_m \rightarrow O_m \\ O_n \rightarrow O_n \end{cases} \end{cases}$$

Figure 3-1 shows the overlap among CIs in three possible situations: a) Total overlap, in which one of the CIs is completely overlapped with the other one, b) Partial overlap, in which only a portion of CIs are overlapped, and c) No overlap, in which the CIs do not have any overlap whatsoever. It is worth mentioning that the reverse cases are not included in the figure to save the brevity of the paper.

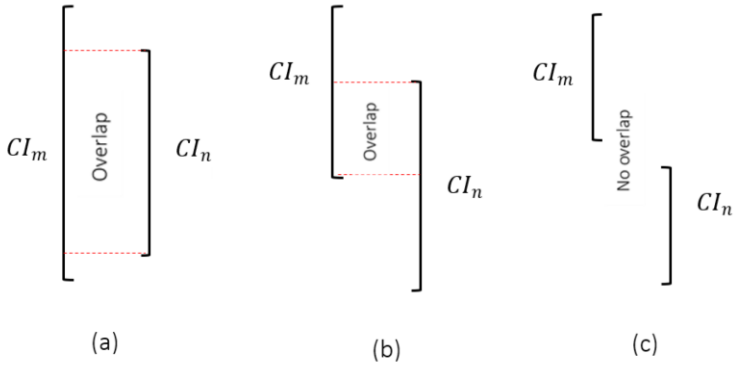

Figure 3-1. The graphical representation of overlap among CIs

The result of this process is a matrix of overlaps such as P^4. The sum of each row will yield the total overlap of each DM with other DMs. It is worth mentioning that the diagonal values of the overlap matrix are the length of CI for the associated DM. However, since we are interested in having the overlap among CIs of different DMs, we reduced the diagonal values from the sum of overlaps for all DMs.

$$P^4 = \begin{array}{c} \\ d_1 \\ d_2 \\ \dots \\ d_{I'} \end{array} \begin{array}{c} \begin{array}{cccc} d_1 & d_2 & \dots & d_{I'} \end{array} \\ \begin{pmatrix} l_1 & O_{12} & \dots & O_{1I'} \\ O_{21} & l_2 & \dots & O_{2I'} \\ \dots & \dots & \dots & \dots \\ O_{I'1} & O_{I'2} & \dots & l_{I'} \end{pmatrix} \end{array}, o_i = \sum_{q=1}^{I'} O_{iq} - l_i, l_i = UB_i - LB_i$$

For each DM, there is an upper bound of CI overlap with others M_i, which is the product of the length of its CI and the number of other DMs (Equation 7). Dividing the actual ratio and the maximum possible value in Equation 8 will give us the overlap ratio for each DM.

$$M_i = (I^{'} - 1) \times l_i \tag{7}$$

$$\tilde{o}_i = \frac{o_i}{M_i} \tag{8}$$

3.3.2.2 Relative CI calculation

To have the relative CI of each DM to the total CI, we use equations 5 and 6 to find the CI_i. Then, Equations 9 and 10 are used to calculate the total CI based on matrix C input. Note that in equations 11 and 12, the total number of scores in the decision matrix after eliminating the biased ones ($\lambda = N \times I^{'}$) in the first step (matrix P^3) is used to find the CI.

$$\bar{x} = \frac{\sum_{i=1}^{I^{'}} \sum_{j=1}^{J} \sum_{k=1}^{K} \psi_{ijk}}{\lambda},\ \sigma = \sqrt{\frac{1}{\lambda - 1} \sum_{i=1}^{I^{'}} \sum_{j=1}^{J} \sum_{k=1}^{K} (\psi_{ijk} - \bar{x})^2} \tag{9}$$

$$CI = (\bar{x} - t(\lambda - 1, \alpha) \frac{\sigma}{\sqrt{\lambda}}, \bar{x} + t(\lambda - 1, \alpha) \frac{\sigma}{\sqrt{\lambda}}) \tag{10}$$

The calculated CI for each DM is then compared with the CI for total data in Equation 11 to find the relative confidence interval value.

$$CI_i = \frac{CI_i}{CI} \tag{11}$$

3.3.2.3 Weight calculation

Having the overlap ratio and the relative CI, we can calculate a baseline value for assigning weights to DMs as it is shown in Equation 12.

$$\omega_i = \frac{\tilde{o}_i \times CI_i}{\sum_{i=1}^{I^{'}} \tilde{o}_i \times CI_i} \tag{12}$$

Equation 12 shows that a decision-maker has a higher weight (compared to other individual decision-makers) if that decision-maker has more relative agreement with the other decision-makers.

3.3.3 MABM version of the proposed approach

As we mentioned, the MABM version is different from the EABM one in the weighting phase. After eliminating the biased DMs in the elimination phase, a combinatory approach is used to assign weights. First, a minimum weight is assigned to all DMs. Equation 13 shows this predefined minimum value, in which I' is the number of remaining DMs in the pool of GDM and γ parameter controls the total share of this part of weight in the total weight. Then the other half of weight is distributed among DMs based on the weighting scheme described in section 3.3.2.3. Equation 14 shows the method to compute the final weight for DMs.

$$\omega_{\min} = \frac{\gamma}{I'} \tag{13}$$

$$\omega_i^{MABM} = \omega_{\min} + (1-\gamma)\omega_i \tag{14}$$

3.3.4 SABM version of the proposed approach

The SABM version skips the elimination phase, and only follows the logic of MABM version for all DMs in the initial pool. The rationale of this version is the fact that sometimes excluding DMs from the GDM pool is impractical, so that we need to tolerate biasedness in a logical way. Therefore, the part of the final weight that comes from the overlap of each DM with other DMs reflect the biasedness in these situations. Again, the minimum predefined weight is the same for this version as the SABM.

3.3.5 The pseudo-code of the proposed approach

As the final stage of this section, we present the pseudo-code for EABM version of the proposed framework in Figure 3-2. Since the SABM and MADBM are slightly different from this version, we did not include the pseudo-codes for these variants.

3.3.6 Application and possible scenario

Based on the nature of GDM problems, different versions of the proposed framework in this study can be applied on certain categories of problems. Table 3-2 shows the applicability of our framework based on the specific context of a GDM process. Table 3-3 also summarizes the proposed solution of the framework in all possible scenarios in terms of having outliers in the pool of DMs.

Table 3-2. Applications of the proposed framework

Version	GDM context
EABM and MABM	The aggregated opinion of many DMs is desired (like voting systems). The collective agreement is acceptable by everyone. An agent collects and aggregates anonymous votes of DMs.
SABM	DMs have interest, like stakeholders or board of directors. DMs observe the consensus-reaching process. DMs do not consent to be eliminated from the GDM process.

Table 3-3. Proposed solution for the possible scenarios in a GDM process

Scenario	Proposed solution
(i) The pool of DMs does not contain any outlier	The weights of DMs are calculated based on the level of accordance with the total opinion.
(ii) The pool of DMs contains outliers	Detect the outliers and handle the biasedness in a systematic approach.
(iii) The agreement is unreachable in the GDM	Tune the parameters to reach minimum desirable consensus or give equal weights.

Begin

/* GDM initialization and parameter setting/

Gather the opinions of DMs about all alternatives in each criterion;

Set the B and α value for CI; Combine the data in an integrated decision matrix;

Begin normalization Determine the nature of the criteria (positive or negative);

Find the Min and Max values for each criterion; Calculate the normalized values;

End

/* End of initialization/

/* Column elimination of biased DMs/

Transform the data to a column matrix; Calculate the CIs and B_i for each DM;

If B_i is lower than the B

Eliminate d_i as a biased one;

Else

Keep d_i in the decision-making process;

End if

/* End of column elimination /

If No consensus is reachable, or the number of the eliminated DMs is higher than R

Tune the B and α parameters to have a minimum consensus or give the equal weights to all DMs

End if

/* Weight assignment phase/

Calculate the CIs for the remaining DMs and the total data;

for $DM\ m = 1\ to\ I'$ **do**

  for $DM\ n = 1\ to\ I'$ **do**

    If CI_m has overlap with CI_n

      Calculate the overlap O_{mn} and add to the total overlap of O_m and O_n;

    Else

      The total overlap values will remain unchanged;

    End if

    End

    End

  end

end

for $DM\ i = 1\ to\ I'$ **do**

  Calculate the overlap ratio $\widetilde{O}_i$;

  Calculate the proportional value of CI_i to the total CI;

  Calculate the w_i by normalizing the product of $\widetilde{O}_i$ and $\widetilde{CI}_i$;

end

/* End Weight assignment phase /

/*Calculating performance measures/

Calculate MAD, AWD, NEDM, and REDM measures;

/*End of calculating the performance measures/

End

Figure 3-2. Pseudo-code of EABM

3.4 Computational Results

This part includes an explanatory example to clarify the designed steps in the process, along with the analysis of the proposed method on a categorized series of test problems.

3.4.1 Numerical example

In this section, one example GDM problem with 5 DMs, 6 alternatives, and 3 criteria is considered to explain the steps of the proposed approach in detail. Table 3-4 shows the normalized scored assigned to the alternatives in each criterion sorted based on 5 present DMs in the process.

Table 3-4. The normalized scores for the numerical example

		DM1	DM2	DM3	DM4	DM5
Alternative 1	Criterion 1	0.48692	0.57043	0.67077	0.59369	0.36061
	Criterion 2	0.41586	0.75376	0.57771	0.51611	0.34695
	Criterion 3	0.54536	0.30971	0.49205	0.60229	0.21558
Alternative 2	Criterion 1	0.45033	0.67194	0.72409	0.60409	0.35013
	Criterion 2	0.52772	0.59235	0.69956	0.58426	0.28778
	Criterion 3	0.34813	0.83423	0.54178	0.62333	0.17443
Alternative 3	Criterion 1	0.48894	0.61275	0.75015	0.68182	0.16171
	Criterion 2	0.32891	0.72741	0.55521	0.56144	0.42836
	Criterion 3	0.48867	0.71386	0.71333	0.4974	0.30785
Alternative 4	Criterion 1	0.37971	0.43229	0.62814	0.49796	0.39818
	Criterion 2	0.37648	0.79708	0.63033	0.6325	0.26244
	Criterion 3	0.42778	0.53919	0.76235	0.68966	0.34949
Alternative 5	Criterion 1	0.33189	0.27685	0.70551	0.51662	0.20792
	Criterion 2	0.43112	0.56941	0.56176	0.57655	0.30157
	Criterion 3	0.33699	0.77226	0.72425	0.503	0.46047
Alternative 6	Criterion 1	0.54532	0.42602	0.64256	0.59018	0.24999
	Criterion 2	0.52755	0.77619	0.50469	0.63534	0.40827
	Criterion 3	0.40366	0.85662	0.41516	0.5835	0.46604

3.4.2 EABM version

As the first step, we need to calculate the CIs for all DMs, which is presented in Figure 3-3.

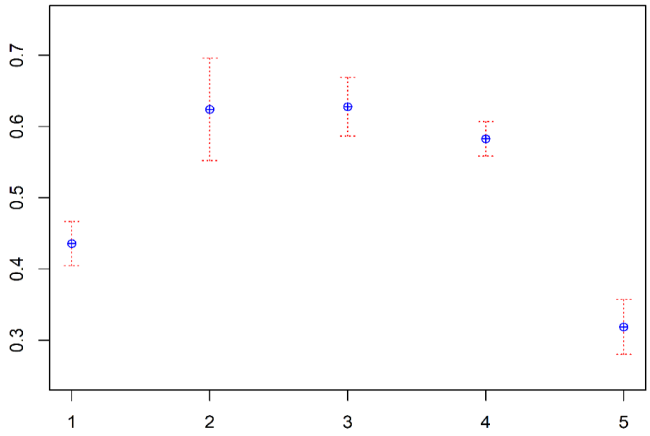

Figure 3-3. CIs of DMs

Since d_1 and d_5 do not have any overlap with other DMs, they will be excluded from the pool of decision-making as outliers. This elimination is done by comparing the CIs of the DMs. For example, the CI of opinions for d_1 does not have any overlap with other DMs. Likewise, the CI of d_5 is outside the range of CIs for other DMs. On the other hand, d_2, d_3, and d_4 have overlap with two other DMs. By considering a threshold $B = 2$, d_1 and d_5 are eliminated since they do not have any overlap with other DMs. Then, Table 3-5 shows the overlap ratio for the remaining DMs. Please note that diagonal values are the length of CI for each DM (Refer to section 3.3.2.1).

Table 3-5. The overlap ratio table

	d_2	d_3	d_4	$\sum O_{mn}$	o_i	$\tilde{o}_i$
d_2	0.0935	0.08216	0.04844	0.224101	0.130601	0.454877
d_3	0.08216	0.10595	0.2032	0.20843	0.102479	0.623649
d_4	0.04844	0.02032	0.00365	0.07241	0.068758	0.709726

For instance, the overlap between d_2 and d_3 shows that the opinions of these two DMs are closer compared to other combinations. d_3 and d_4 have the least overlap in this specific case, though. Column $\sum O_{mn}$ is simply calculated by summation over each row, and column o_i is obtained by reducing the diagonal values form the summations. Finally, column $\tilde{o}_i$ is derived by dividing the previous column over the maximum possible overlap value for each DM. This column shows that d_4 has the highest ratio, meaning that this DM has 70 percent of potential overlap with other DMs. In contrast, d_2 has the lowest ratio, which means its CI has an average overlap with others. The next step is to calculate the relative CI for each DM. The result of this step is provided in Table 3-6.

Table 3-6. The relative CI

	d_2	d_3	d_4
CI_i	0.14356	0.08216	0.04844
CI_i	2.06408	1.18133	0.69648

The row CI_i reflects the length of CI for each DM. By dividing this row on the CI total, the next row of relative CI will be obtained. For example, the relative value for d_2 mirrors the fact that its CI is much wider than the CI of total opinions. Among the DMs in the pool, d_4 has the tightest CI compared to the CI total.

As the final step, the multiplication of overlap ratio and relative CI should be divided on the sum of these productions calculated to have the final weights of DMs. Table 3-7 shows the results of this step. Having the weights, we can aggregate the opinions of the DMs in the pool. The weighted sum of scores for each alternative provides the final ranking of the GDM problem.

Table 3-7. The weight calculations

	d_2	d_3	d_4	Sum
ω_i	0.43269	0.33952	0.2278	1

3.4.3 MABM version

We need to first find the minimum predefined weight for the remaining DMs to obtain the results for this numerical example for the MABM version. In this example, we consider the γ parameter equal to 0.5 so that the total predefined comprises 50 percent of the total weight. We use the following formula to find this predefined weight for each DM:

$$\omega_{\min} = \frac{0.5}{3} = 0.1667 \tag{15}$$

Then, we multiply the weight calculated in Table 3.7 in 0.5 to normalize this segment of the final weight. As the final step we add the normalized weights to the minimum predefined weight to find the weights in this version. For example, for d_4 the weight is:

$$\omega_4 = 0.1667 + 0.5 \times 0.2278 = 0.2806 \tag{16}$$

The final weights are shown in Table 3-8.

Table 3-8. The weight calculations for moderate version

	d_2	d_3	d_4	Sum
ω_i	0.383	0.3364	0.2806	1

3.4.4 SABM version

Finally, since we do not have any elimination in the SABM version, the threshold weight will be acquired from the following equation:

$$\omega_{\min} = \frac{0.5}{5} = 0.1 \tag{17}$$

Then, the other part of weight is based on the overlaps between the DMs like Table 3-5. Having these inputs, we can calculate the weights for all DMs. For instance, for DM2 the weight is calculated as:

$$\omega_2 = 0.1 + 0.5 \times 0.43269 = 0.3163 \tag{18}$$

Table 3-9 shows the final weights for all DMs in the poll in the soft version.

Table 3-9. The weight calculations for soft version

	d_1	d_2	d_3	d_4	d_5	Sum
ω_i	0.1	0.3163	0.3364	0.2806		1

3.4.5 Test Scenarios

To test the performance of the proposed approach, we first identified the factors affecting the output of the process, including the number of DMs, the significance level for the statistical calculation of CI, and the B threshold. Table 3-10 shows the considered levels for the mentioned variables.

Table 3-10. The levels for scenarios

	Level 1	Level 2	Level 3
I	3	5	10
B	$\frac{I}{3}$	$\frac{I}{2}$	$I-1$
α	0.9	0.95	0.99

A total number of 27 combinations is considered as the benchmark problems for the proposed approach. In each category, 1,000 random instances are generated. The codes of the algorithm are written in MATLAB© software package and are executed on a PC with Intel® Xenon® Bronze 3160 1.70 GHz CPU with 16 GB RAM. On average, the computational time for each problem was about 22 minutes, resulting in 10-hour time to obtain the results for all the cases.

3.4.6 Performance measures

To compare and analyze the performance of our approach, we defined four performance measures as follows:

3.4.6.1 Mean absolute deviation (MAD)

This indicator mirrors the average difference among the (normalized) weights of the DMs present in the decision-making process and their initial weight (which is considered equal among the DMs) before applying the method. Note that for the biased DMs, the final weights are

considered zero. Equation 19 captures this concept, in which ω_i and $\hat{\omega}_i$ are the final normalized and the initial weights for the d_i, respectively.

$$MAD = \sum_{i=1}^{I} \left| \omega_i - \hat{\omega}_i \right| \tag{19}$$

3.4.6.2 Average weight deviation (AWD)

This measure captures the deviation of weights assigned to the DMs remained in the GDM pool compared to the equally distributed weights scheme. Equation 20 shows the formula for this performance indicator:

$$AWD = \sum_{i=1}^{I'} \left| \omega_i - \frac{1}{I'} \right|, \forall i = 1, ..., I' \in \{\text{Unbiased DMs}\} \tag{20}$$

3.4.6.3 Number of eliminated decision makers (NEDM)

As the number of DMs plays a significant role in the decision-making problems, we record the number of eliminated (biased) DMs as a result of applying our method on the GDM procedure. This measure can be calculated by subtracting the number of DMs in the second phased (I') and the beginning of the process (I) reflected in equation 21. Please note that this measure is not applicable for the soft version as there is no elimination in this setting.

$$NEDM = I - I' \tag{21}$$

3.4.6.4 Ratio of eliminated decision makers (REDM)

This performance measure captures the ratio of the eliminated DMs in the proposed framework to the total number of DMs in the GDM process. Equation 22 shows the formulation for this indicator. Again, please note that this measure is not applicable for the soft version.

$$REDM = \frac{NEDM}{I} \tag{22}$$

3.4.7 Output analysis

This section includes the detailed results of the devised scenarios, the trend of the defined performance measures, and the noteworthy patterns. The analysis is performed in three parts. First, the trends of various elimination situations in the generated problems are presented to

discuss the patterns and the prescription of the framework to handle the process. Then, to assess the relationship among variables, their interaction, and performance measures, a Pearson correlation analysis is executed. Finally, to evaluate the effect of the variables and their interactions on the performance measures, a multivariate linear regression analysis is designed. The rationale for these separated analyses is to assess two different aspects of the problem. By doing the correlation analysis, we want to investigate the whole possible combinations among the parameters and performance measures in a holistic approach. On the other hand, in the regression analysis, we want to evaluate the impact of variables and their interactions on each performance measure separately.

3.4.7.1 Elimination analysis

Figure 3-4 demonstrates three trends in the test scenarios; 1) The share of problems in each case that none of the DMs are eliminated in the proposed framework (the first step), 2) The share of problems in which a part of DMs (from 1 DM to $I-1$) are excluded from the final pool, and 3) The share of problems without any DM at the second stage as all of them are eliminated in the first stage. This analysis is only plausible for the EABM and MABM versions of the proposed method.

The numbers of the scenarios are based on the levels of the parameters in Table 3-6. The share of problems with no elimination confronted a descending trend in general, meaning that as the number of DMs, B, and α rises, the possibility for the elimination of completely biased DMs goes up. In contrast, for partial elimination, the share grows as the parameters increase in the testbed. This trend will be explained in the regression analysis section. Finally, the share of problems in which there is no reachable consensus among the DMs in the first step is relatively constant with certain spikes. This share increases for test cases with higher levels (2 and 3) for B and α parameters. However, we included these problems in the class of problems where no DM is eliminated. In these cases, as reaching an agreement is impossible, there are some options. For one thing, the agent can tune the parameters (mainly B and α) to the point that a minimum desirable consensus is achievable. This minimum consensus is different in various context, like more than 50 percent in some political systems and even the biggest minority in other cases. So, this parameter is entirely in the control of the aggregator agent. Another option is considering an equally distributed weight for all DMs. As there is not any overlaps between the CIs of any two DMs, this strategy assures to keep all the DMs in the pool of GDM. Since we generated random

cases in this study and there is no aggregator agent to tune the parameters, we considered the equal weight scenario for these specific situation. However, in real-life applications, both of the proposed solutions can be implemented when there is no consensus in the first phase.

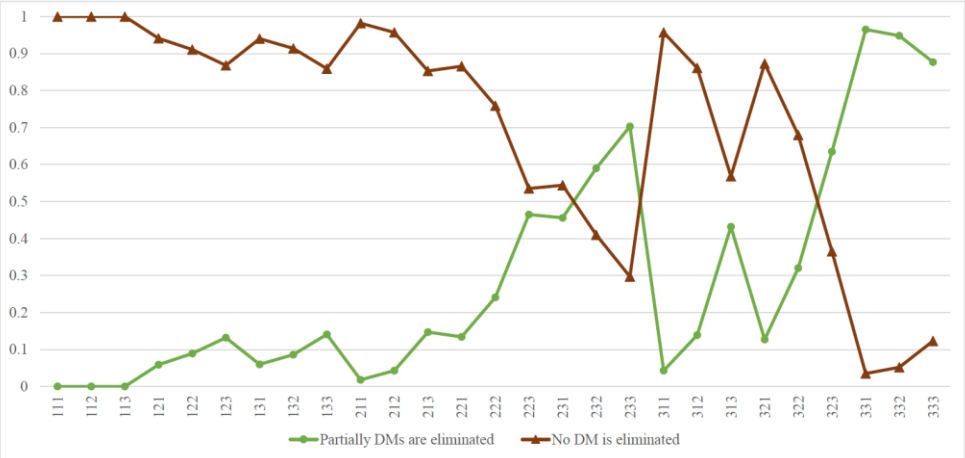

Figure 3-4. Trend of elimination in the scenarios

3.4.7.2 Correlation analysis

Figure 3-5 shows the results of the correlation analysis for the output of generated problems for the EABM (a), MABM (b), and the SABM (c) versions of the proposed algorithm. The insignificant correlations at level 0.001 are marked by asterisks in this Figure.

There are some interesting correlations between the parameters and measures. For example, there is a strong correlation among NEDM measure and the interactions of NDM and B and NDM, B and α . On the other hand, a relatively meaningful negative correlation exists among MAD and AWD measures and NEDM and REDM measures, meaning they display opposite behaviors in the generated examples. Parameter α has weak correlations with all other ones, though. The patterns are almost similar in all versions, and the only difference is that the correlation between α and the MAD performance measure is insignificant in this configuration and the correlation between AWD and MAD is lower compared to other cases.

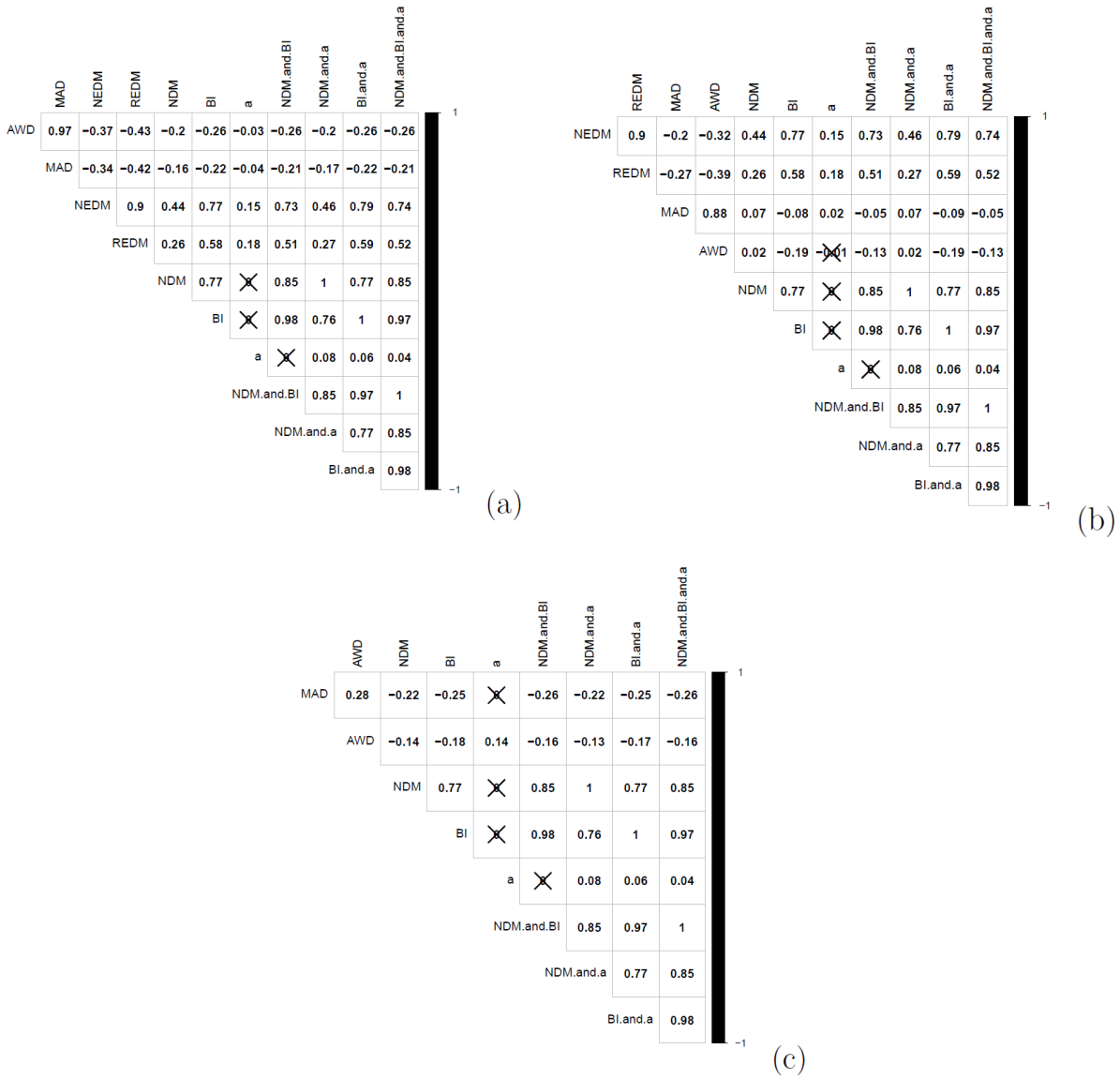

Figure 3-5. Correlation analysis result

3.4.7.3 Regression analysis

Table 3-11 presents the results of a multivariate linear regression conducted on the output of the generated scenarios for the extreme version. To save the space, we present the tables for other versions in the appendix (a) for MABM and (b) for SABM.

Table 3-11. The results of the regression analysis

Measure	MAD		AWD		NEDM		REDM	
	Coeff	P-Value	Coeff	P-Value	Coeff	P-Value	Coeff	P-Value
Constant		0.00		0.000		0.000		0.000
NDM	-0.531***	0.000	-0.558***	0.000	1.215***	0.000	-0.001***	0.000
B	1.192***	0.000	1.392***	0.000	-1.817***	0.000	-0.766***	0.000
α	0.185***	0.000	0.221***	0.000	-0.079***	0.000	-0.089***	0.000
NDM and B	-0.542**	0.000	0.341*	0.000	2.693***	0.000	1.304***	0.000
NDM and α	-0.590***	0.000	-0.679***	0.000	0.462***	0.000	1.001***	0.000
B and α	-1.627***	0.000	-1.917***	0.000	2.329***	0.000	0.770***	0.000
NDM, B, and α	1.460	0.000	1.453***	0.000	-3.765***	0.000	-1.303***	0.000

Coeff= Coefficients, Significance Codes: "***"=0.001, "**"=0.01, "*"=0.05

Analyzing the patterns emerged in Table 3-11 will lead to the following insights:

(i) B has the highest impact on the weight-related measures, as higher value (level) forces the method to exclude more DMs. Therefore, the changes in the weights are more noticeable. Conversely, this parameter also has negative coefficients for NEDM and REDM measures.

(ii) In the performance related to weights, MAD and AWD, parameter α has a high coefficient in the estimated models. This can be justified by the fact that by increasing α, the CIs become wider. As a result, the possible overlap among DMs increases, resulting in more fluctuations in their weights compared to the beginning of the process. However, the coefficient for α for two other performance measures, NEDM and REDM, is negative. In other words, with wider CIs, the number of eliminated DMs and the ratio to the total number of DMs decreases.

(iii) The NDM parameter has a considerable negative impact on weight related measures and a strong positive one on the NEDM measure. The effect of NDM on REDM measure is almost negligible, though.

(iv) Interestingly, the interaction of B and α has the highest negative effect on all the performance measures related to change of weights and the second highest positive one on the elimination-related measures. This intensified effect can be associated with the synergy of these parameters to convert a problem to an extreme one.

(v) The interaction of NDM and α and NDM and B has similar behavior to that of B and α, while other interaction (NDM and B and α) shows conflicting behavior in terms of the direction of coefficients, yet all of the relations are statistically significant.

3.5 Conclusion and Future Research

This paper presents a structured framework to deal with biased decision-makers in the group decision-making process. Our focus was on aggregating the evaluations of a set of alternatives with respect to a set of criteria provided by a set of independent DMs. It starts with determining the entirely biased DMs and removes them from the decision-making pool. Then, it continues with remained DMs and assigns a weight to each DM based on its consensus compared with other DMs. The proposed algorithm is developed using two ratios, which are called Overlap Ratio and Relative Confidence Interval. The Overlap Ratio indicates the relative value of overlap between CI of each DM and other decision-makers by making the pairwise comparison of their confidence intervals. The Relative Confidence Interval shows the relative joint range of CI for each DM compared to the total CI of the remained decision-makers.

The proposed method is one of the first attempts in the context of group decision making where one is concerned with biased decision-makers, tries to detect the sources of biases systematically, and finally to handle the partially biased DMs. One of the main applications of our proposed methods is where there is an agent (main DM) who is responsible for aggregating the DMs opinion, and DMs do not have information either about aggregation procedure or the opinion of other DMs. In this setting, the biased DMs can be excluded from the final decision-making pool. However, there are two possible scenarios in case DMs do have information about the aggregation procedure. Either collective agreement is acceptable by all DMs, or it is not. The proposed framework is not applicable in the latter case. Finally, our developed method can be used in the initial phase of any group decision-making procedure, falling into the mentioned categories. The proposed framework also provides satisfactory flexibility in dealing with all possible situations in a GDM problem. If the pool of DMs does not contain any outlier, the weights of DMs are calculated based on the level of accordance with the total opinion. However, the framework can detect the outliers and handle the biasedness in a systematic approach in other cases. In the presence of biasedness, the framework provides different prescription for the cases. The biased DMs can be excluded (if possible) from the GDM problem (EABM version) or they can be kept

but assign their partial weights based on a combination of a minimum predefined weight and their disagreement level from other DMs (SABM version). As a middle way, the MABM version excludes the biased DMs, but assign weights to the remaining DMs based on a combination of minimum weight and relative weight based on the level of agreement.) Finally, if an agreement is unreachable in the GDM, the framework suggests revising the opinions or having new sources for the evaluation of alternatives based on the criteria. As a result, we can claim that the proposed framework and its variants cover adequately all possible scenarios in a GDM setting from strict exclusion of biased DMs to only reflecting biasedness in the weighting phase. As previously stated, the moderate and soft versions of our framework do not exclude any decision maker from disagreeing with the norm, and any bias will be reflected in the weighting. One may argue that this is in conflict with the concepts of Diversity, Equity, and Inclusion. We acknowledge that this might be considered as a potential drawback of this framework. However, if the DM selects for the soft version, this disadvantage is negligible because she has the freedom to distribute the majority of the weight equally among all DMS.

Several paths can be followed as fruitful research directions for expanding the current study. This research tried to detect the outlier DMs in total. However, we aim to develop another comprehensive method to detect biased DMs towards particular alternatives and mirror this biasedness by either ignoring their opinions for those alternatives or assigning relatively less weight to them. Another direction could be applying our proposed method is real-world group decision-making problems. This area of research requires more attention, particularly developing new algorithms for individual cases. In specific, in the political context, the decisions are highly correlated (similar opinions of voters belong to a political party for example). Even though this situation is different from a routine GDM problem, the proposed methodology in this study can be extended to incorporate the correlated decisions and handle these inherently different outliers as they adversely affect the output of the aggregation phase. Another prospective research direction is to examine this problem from the perspective of the DEI and to develop a framework that prioritizes underrepresented individuals in decision-making processes.

References Cited

Akram, M., Adeel, A., & Alcantud, J. C. R. (2019). Group decision-making methods based on hesitant N-soft sets. Expert Systems with Applications, 115, 95-105.

Akram, M., Bashir, A., & Garg, H. (2020)(a). Decision-making model under complex picture fuzzy Hamacher aggregation operators. Computational & Applied Mathematics, 39(3), 1-38.

Akram, M., Yaqoob, N., Ali, G., & Chammam, W. (2020)(b). Extensions of Dombi Aggregation Operators for Decision Making under-Polar Fuzzy Information. Journal of Mathematics, 2020.

Ambrus, A., Greiner, B., & Pathak, P. (2009). Group versus individual decision-making: Is there a shift. Institute for Advanced Study, School of Social Science Economics Working Paper, 91.

Anojkumar, L., Ilangkumaran, M., & Sasirekha, V. (2014). Comparative analysis of MCDM methods for pipe material selection in sugar industry. Expert Systems with Applications, 41(6), 2964-2980.

Bonner, B. L., Baumann, M. R., & Dalal, R. S. (2002). The effects of member expertise on group decision-making & performance. Organizational Behavior & Human Decision Processes, 88(2), 719-736.

Bouzarour-Amokrane, Y., Tchangani, A., & Peres, F. (2015). A bipolar consensus approach for group decision making problems. Expert Systems with Applications, 42(3), 1759-1772.

Ceschi, A., Costantini, A., Sartori, R., Weller, J., & Di Fabio, A. (2019). Dimensions of decision-making: an evidence-based classification of heuristics & biases. Personality & Individual Differences, 146, 188-200.

Cinelli, M., Kadzinski, M., Gonzalez, M., & Slowinski, R. (2020). How to Support the Application of Multiple Criteria Decision Analysis? Let Us Start with a Comprehensive Taxonomy. Omega, 102261.

Dong, Q. & Cooper, O. (2016). A peer-to-peer dynamic adaptive consensus reaching model for the group AHP decision making. European Journal of Operational Research, 250(2), 521-530.

Dong, Q., Zhou, X., & Martínez, L. (2019). A hybrid group decision making framework for achieving agreed solutions based on stable opinions. Information Sciences, 490, 227-243.

Dong, Y., Zha, Q., Zhang, H., Kou, G., Fujita, H., Chiclana, F., & Herrera-Viedma, E. (2018). Consensus reaching in social network group decision making: Research paradigms & challenges. Knowledge-Based Systems, 162, 3-13.

Dong, Y., Zhang, H., & Herrera-Viedma, E. (2016). Integrating experts' weights generated dynamically into the consensus reaching process and its applications in managing non-cooperative behaviors. Decision Support Systems, 84, 1-15.

Forman, E., & Peniwati, K. (1998). Aggregating individual judgments and priorities with the analytic hierarchy process. European journal of operational research, 108(1), 165-169.

Gonzalez-Arteaga, T., Alcantud, J. C. R., & de &res Calle, R. (2016). A new consensus ranking approach for correlated ordinal information based on Mahalanobis distance. Information Sciences, 372, 546-564.

Hafezalkotob, A. & Hafezalkotob, A. (2017). A novel approach for combination of individual and group decisions based on fuzzy best-worst method. Applied Soft Computing, 59, 316-325.

He, Z., Chan, F. T., & Jiang, W. (2018). A quantum framework for modelling subjectivity in multi-attribute group decision making. Computers & Industrial Engineering, 124, 560-572.

Ichniowski, C. & Preston, A. (2017). Does March Madness lead to irrational exuberance in the NBA draft? High-value employee selection decisions & decision-making bias. Journal of Economic Behavior & Organization, 142, 105-119.

Jacobi, S. K. & Hobbs, B. F. (2007). Quantifying & mitigating the splitting bias and other value tree-induced weighting biases. Decision Analysis, 4(4), 194-210.

Jin, F., Ni, Z., Pei, L., Chen, H., Tao, Z., Zhu, X., & Ni, L. (2017). Approaches to group decision making with linguistic preference relations based on multiplicative consistency. Computers & Industrial Engineering, 114, 69-79.

Jin, L., Mesiar, R., & Yager, R. R. (2018). The paradigm of induced ordered weighted averaging aggregation process with application in uncertain linguistic evaluation. Granular Computing, 1-7.

Jones, P. E. & Roelofsma, P. H. (2000). The potential for social contextual and group biases in team decision-making: Biases, conditions & psychological mechanisms. Ergonomics, 43(8), 1129-1152.

Kabak, Ö. & Ervural, B. (2017). Multiple attribute group decision making: A generic conceptual framework & a classification scheme. Knowledge-Based Systems, 123, 13-30.

Kadzinski, M., Greco, S., & Slowinski, R. (2013). Selection of a representative value function for robust ordinal regression in group decision making. Group Decision & Negotiation, 22(3), 429-462.

Lee, Y.-C. (2014). Impacts of decision-making biases on eWOM retrust and risk-reducing strategies. Computers in Human Behavior, 40, 101-110.

Liao, H., Wu, X., Mi, X., & Herrera, F. (2019). An integrated method for cognitive complex multiple experts multiple criteria decision making based on ELECTRE III with weighted Borda rule. Omega.

Liao, H., Zhang, C., & Luo, L. (2018). A multiple attribute group decision making method based on two novel intuitionistic multiplicative distance measures. Information Sciences, 467, 766-783.

Lieder, F., Griffiths, T. L., Huys, Q. J., & Goodman, N. D. (2018). The anchoring bias reflects rational use of cognitive resources. Psychonomic bulletin \& review, 25(1), 322-349.

Liu, P. (2011). A weighted aggregation operators multi-attribute group decision-making method based on interval-valued trapezoidal fuzzy numbers. Expert Systems with Applications, 38(1), 1053-1060.

Liu, P., Shahzadi, G., & Akram, M. (2020). Specific Types of q-rung Picture Fuzzy Yager Aggregation Operators for Decision-Making. International Journal of Computational Intelligence Systems, 13(1), 1072-1091.

Liu, W., Zhang, H., Chen, X., & Yu, S. (2018). Managing consensus and self-confidence in multiplicative preference relations in group decision making. Knowledge-Based Systems, 162, 62-73.

Liu, Y., Du, J.-l., & Wang, Y.-h. (2019). An improved grey group decision-making approach. Applied Soft Computing, 76, 78-88.

McShane, M., Nirenburg, S., & Jarrell, B. (2013). Modeling decision-making biases. Biologically Inspired Cognitive Architectures, 3, 39-50.

Mohammadi, M. & Rezaei, J. (2019). Bayesian Best-Worst Method: A Probabilistic Group Decision Making Model. Omega.

Mokhtari, S. (2013). Developing a Group Decision Support System (GDSS) For Decision Making Under Uncertainty.

Montibeller, G. & Von Winterfeldt, D. (2015). Cognitive and motivational biases in decision and risk analysis. Risk analysis, 35(7), 1230-1251.

Morente-Molinera, J. A., Kou, G., Samuylov, K., Ureña, R., & Herrera-Viedma, E. (2019). Carrying out consensual Group Decision Making processes under social networks using sentiment analysis over comparative expressions. Knowledge-Based Systems, 165, 335-345.

Palomares, I., Estrella, F. J., Martínez, L., & Herrera, F. (2014). Consensus under a fuzzy context: Taxonomy, analysis framework AFRYCA & experimental case of study. Information Fusion, 20, 252-271.

Parreiras, R. O., Ekel, P. Y., Martini, J. S. C., & Palhares, R. M. (2010). A flexible consensus scheme for multicriteria group decision making under linguistic assessments. Information Sciences, 180(7), 1075-1089.

Rokou, E., & Kirytopoulos, K. (2014). A calibrated group decision process. Group Decision & Negotiation, 23(6), 1369-1384.

Song, Y. & Hu, J. (2019). Large-scale group decision making with multiple stakeholders based on probabilistic linguistic preference relation. Applied Soft Computing, 80, 712-722.

Song, Y. & Li, G. (2019). A mathematical programming approach to manage group decision making with incomplete hesitant fuzzy linguistic preference relations. Computers & Industrial Engineering.

Stengel, D. N. (2013). Aggregating incomplete individual ratings in group resource allocation decisions. Group Decision & Negotiation, 22(2), 235-258.

Stewart, T. (2005). Goal programming and cognitive biases in decision-making. Journal of the Operational Research Society, 56(10), 1166-1175.

Tang, J., Meng, F., & Zhang, Y. (2019). Programming model-based group decision making with multiplicative linguistic intuitionistic fuzzy preference relations. Computers &Industrial Engineering.

Vafaei, N., Ribeiro, R. A., & Camarinha-Matos, L. M. (2016). Normalization techniques for multi-criteria decision making: analytical hierarchy process case study. Paper presented at the doctoral conference on computing, electrical & industrial systems.

Wibowo, S. & Deng, H. (2013). Consensus-based decision support for multicriteria group decision making. Computers \& Industrial Engineering, 66(4), 625-633.

Wittenbaum, G. M., Vaughan, S. I., & Strasser, G. (2002). Coordination in task-performing groups. In Theory & research on small groups (pp. 177-204). Springer, Boston, MA.

Wu, Z. & Xu, J. (2016). Managing consistency & consensus in group decision making with hesitant fuzzy linguistic preference relations. Omega, 65, 28-40.

Xia, M. & Chen, J. (2015). Multi-criteria group decision making based on bilateral agreements. European Journal of Operational Research, 240(3), 756-764.

Xiao, B., & Benbasat, I. (2018). An empirical examination of the influence of biased personalized product recommendations on consumers' decision making outcomes. Decision Support Systems, 110, 46-57.

Xie, K., Liu, J., Chen, G., Wang, P., & Chaudhry, S. S. (2012). Group decision-making in an unconventional emergency situation using agile Delphi approach. Information Technology & Management, 13(4), 351-361.

Xu, W., Huang, S., & Li, J. (2019). A novel consensus reaching framework for heterogeneous group decision making based on cumulative prospect theory. Computers & Industrial Engineering, 128, 325-335.

Xu, X.-h., Du, Z.-j., Chen, X.-h., & Cai, C.-g. (2019). Confidence consensus-based model for large-scale group decision making: A novel approach to managing non-cooperative behaviors. Information Sciences, 477, 410-427.

Xu, Z. S., & Jian, C. H. E. N. (2007). Approach to group decision making based on interval-valued intuitionistic judgment matrices. Systems Engineering-Theory & Practice, 27(4), 126-133.

Yu, G.-F., Li, D.-F., & Fei, W. (2018). A novel method for heterogeneous multi-attribute group decision making with preference deviation. Computers & Industrial Engineering, 124, 58-64.

Chapter 4

Expert-Augmented Supervised Feature Selection: Models and Algorithms

This work is co-authored with Prof. Michael Pangburn, Prof. Dursun Delen, Dr. Saeed Piri, and Mohsen Mirhashemi. It is in preparation for submission for publication.

4.1 Introduction

Data mining and knowledge discovery contribute significantly to today's competitive economy. Recent technological advancements and the benefits of machine-learning projects have persuaded businesses to gather and store vast volumes of organized data or unstructured data. This data may include those of consumers, patients, workers, students, and suppliers, among others. The scale of these data sources keeps increasing, with data lakes storing terabytes of data. Not only is the number and dimension of our data rising, but the understanding we seek from this data is becoming increasingly complicated. Determining which subset of this data is truly relevant is a difficult challenge. The term “feature selection” (or “variable selection”) refers to the process of selecting the optimal subset of features based on a predefined criterion (or multiple criteria). Computer scientists, data analysts, mathematicians, and even biologists have all paid close attention to feature selection. The increased availability of data and the progress of technology intensify the need for new approaches and models to address the challenges of feature selection (Li et al., 2019; Bertolazzi et al., 2016).

Including all possible features in a model might naively be thought to yield both more information and more discrimination power, but this is not generally the case. There are several advantages to implementing limited feature selection. First, it may enhance the performance of the models by reducing their loss function or increasing their accuracy rate. This happens as a result of reducing the risk of model overfitting. Second, it improves the interpretability of the results by constructing

simpler models. Third, it removes noise/irrelevant data, making the data cleaner and increasing its degree of understandability (Byeon & Rasheed, 2008). Generally, feature selection is effective for reducing the time and memory requirements that often plague large-scale analyses. Furthermore, greater complexity not only increases costs in time and memory, it also has a detrimental effect on the performance of algorithms due to the outliers effect and the noise effect (Elhamifar and Vidal, 2013).

Recent research indicates the limitations of human decision-making in comparison to decision support systems based on machine learning. Verboven et al. (2020) demonstrated in a field experiment that experts are frequently inconsistent in their decision-making, noting the importance of strategic decision support systems. The authors revealed many fundamental flaws in human decision-making using two large industry-provided human-resource datasets Additionally, the authors showed that unsupervised learning with an autoencoder is an advantageous tool for strategic decision-making. Even though this argument is consistent with other papers, this does not imply that the experts' opinions are not helpful. For instance, Coussement et al. (2015) concluded that an algorithmic DSS outperforms human professionals at recognizing customer satisfaction from product reviews. Still, the authors indicate that the performance of the predictions is improved by combining the human expert opinions with the data-mining model to create a justifiable DSS. Additionally, Oosterlinck et al. (2020) gathered expert comments on recognizing fraudulent behavior in relation to a telecom family product. They demonstrated how these expert opinions assist in transforming a fraud prediction problem from a one-class problem to a two-class problem, thereby improving fraud detection and bringing domain-expert forecasts in line with model predictions.

This paper proposes a novel and robust framework for expert-augmented supervised feature selection and demonstrates its efficiency and stability on numerous datasets. Our work has made the following contributions:

- To reflect the trade-off between explainability and prediction accuracy, we developed two mathematical models. In this formulation, we allow the decision-maker (or domain expert) to specify a maximum acceptable sacrifice percentage, to increase explainability by reducing the number of selected features and prioritizing the features that are more relevant from the perspective of practitioners or academicians.
- As a contribution to the solution methodology, we primarily developed an algorithm called Probabilistic Solution Generator through Information Fusion (PSGIF) to ensure efficient exploration and exploitation for the genetic algorithm. Additionally, we tuned our algorithm parameters using Bayesian optimization to reduce the algorithm's computational time, which helps the scalability of the DSS we created.
- We developed a posterior ensemble algorithm to measure the true prediction power of features. The output of this algorithm enables us to classify features as underrated, overrated, or consistently rated by domain experts.

The paper is structured as follows. In Section 2, we discuss the various techniques for feature selection and the relevant literature. Section 3 defines the problem, provides notations, and elaborates on the proposed mathematical models. In Section 4, we discuss our developed methods, ensemble models, and approach to parameter tuning. Then, we provide the results of our numerical analysis on 16 publicly available datasets, to demonstrate the effectiveness of our methods. In Section 5, to capture the trade-off between explainability and accuracy, we conduct a sensitivity analysis by varying the mathematical model parameters. Additionally, we show how to estimate

the actual explainability power of each feature and how this output can be used to classify features into three categories: underrated, overrated, and appropriately rated by practitioners/academicians. Finally, Section 6 summarizes the paper's contributions and suggests some potential study directions for further investigation.

4.2 Feature-Selection Methods

Data scientists desire models that are highly predictive yet easily interpretable. Regrettably, however, there tends to be a trade-off between predictive performance and interpretability, and it is rarely feasible to achieve both simultaneously (Shmueli, 2010). Generally, it is preferable to have fewer variables (or features) in a model. As a result, the focus of the research on feature selection has been on minimizing the number of features while maintaining predictive performance. Numerous issues can be facilitated by feature-selection methods (FSM). Data with a large number of dimensions and a small sample size are attracting attention in a variety of disciplines. In terms of feature-selection applications, a highly practical application of FSM is in the analysis of medical and bioinformatics data, where datasets often contain a significant number of features.

Additionally, FSM can be used to alleviate a particular issue in the interplay between features and a model or to minimize the complexity of the model. Models such as neural networks and support vector machines are highly sensitive to irrelevant features (or predictors). In some instances, these kinds of features might result in worse prediction performance. Models such as logistic or linear regression are susceptible to features that are correlated. Eliminating correlated features reduces multicollinearity and thereby enables the fitting of these sorts of models. Even if a predictive model is resistant to more features, the smallest number of features that give acceptable results makes perfect scientific logic. Removal of features can lower the cost of gathering data or enhance the

prediction performance of the algorithm. FSM can be classified in many ways. From one point of view, there are two forms of FSM: supervised, and unsupervised. FSM can be used in both classification problems (Tang et al., 2014) and regression problems (Andersen & Bro, 2010). Our study focuses on supervised feature selection. Nevertheless, our developed model and algorithms can simply be adapted to unsupervised feature selection with a few modifications.

FSM can be classified according to how they are chosen. Someone might simply rank the features according to their individual merits and then choose the required number of them. Ranker methods take a similar strategy (Kira & Rendell, 1992). On the other hand, one may choose to assess a subset of the features based on their cumulative contribution, posing a more complicated subset-selection challenge that is intrinsically combinatorial in nature and widely acknowledged as a hard problem. For the latter case, certain approaches are created to generate a solution set iteratively by adding features and comparing them to those already in the set; this forward selection approach is paired with a backward selection approach in which features are iteratively removed from the existing set (Bertolazzi et al., 2016). Both of these two approaches are used by our developed algorithms.

Another perspective to assess FSM is to categorize the methods according to their search techniques, the number of objectives, and evaluation measures. There are three distinct types of search techniques: exhaustive search; heuristic search (which includes forward and backward selection procedures (Pudil et al. 1994; Mao et al., 2012), and evolutionary computing (Oreski et al., 2012; Emary et al., 2016). With regard to this classification, our technique can be described as intelligent evolutionary computation. FSM can have one or more objective functions. A single-objective technique aggregates feature counts and classification (or regression) accuracy into a single utility function. In contrast, the multi-objective approach is a way to determine the Pareto

frontier of a trade-off solution. We use an aggregation approach to transform three objective functions into a single value, which is referred to as a single objective function in the literature.

Four types of evaluation methods are used in feature selection: filter; wrapper; intrinsic (or embedded); and hybrid approaches. Methods for filter feature selection use statistical approaches to analyze the interaction between each feature and the target output; these scores are used to prioritize and choose the features that will be used in the model. These approaches evaluate the subset of features without using a classifier. Wrapper methods generate several models from various subsets of features and then choose the features that result in the highest-performance model according to a criterion. In their evaluation step, wrapper techniques make use of classifiers. Intrinsic methods, the third type, make feature selection a component of the model's learning process. This class of methods (such as Lasso) imposes a penalty on model complexity (i.e., the number of features) in order to minimize a model's degree of overfitting or variance by increasing the bias. Finally, hybrid methods combine at least two different classes of algorithms to provide more robust results. Our developed framework can be termed hybrid, since it incorporates both filter and wrapper techniques and uses a penalty function for the number of features aligned with intrinsic methods. Chandrashekar and Sahin (2014) provide a detailed overview of several feature-selection approaches, including some examples and a brief discussion of their stability. The reader is referred to Li et al. (2017) for the most recent surveys on this subject.

4.3 Problem Definition and Models

The literature includes numerous definitions for feature selection (Dash & Liu, 1997). Fundamentally, feature selection can be defined as the process of choosing a subset of features from a larger set. Its objective is to generate an effective feature subset capable of representing the data in an informative manner. In essence, feature selection eliminates redundant and superfluous

features while maintaining (and maybe improving) classification performance. It is well established that feature selection prevents overfitting, decreases computation time and space requirements, and improves performance.

Any feature selection algorithm's primary objective is prediction accuracy. In this paper, we develop two mathematical models (one for classification, and one for regression) as a mechanism for selecting a subset of features that maximizes this primary objective while boosting explainability. Even though maximizing the explainability by incorporating domain experts' opinions is an important goal, in reality, a prediction accuracy threshold is typically established to ensure that the final solution meets the decision maker's expectations. To guarantee high prediction accuracy, we propose a hard threshold for filtering the solution space's least favorite solutions. This threshold should be neither very high nor excessively low. If we take the restrictive approach (setting a very high accuracy threshold for classification and a very low MSE threshold for regression problems), the algorithm may run out of feasible options. On the other hand, if we take a more relaxed approach (setting a very low accuracy threshold for classification and a very high MSE threshold for regression problems), the solution space will include solutions with an unsatisfactory rate of prediction accuracy.

We establish our notation in Table 4-1.We next present our two mathematical models.

Mathematically, it is impossible to propose a single model that is applicable for all types of supervised learning problems, since prediction accuracy in classification refers to a higher accuracy rate and in regression refers to a lower MSE value. Hence, we develop one model for classification datasets (Model 1) and another model for regression datasets (Model 2). These two models are nearly identical except for the objective function's Parts 1 and 4 and any constraint that includes an accuracy rate or MSE.

Table 4-1. Notations and descriptions

Notation	Description
	Set
t	Feature index $t = 1,...,T$
s	Selected Features' set index $s = 1,...,S$
	Parameter
o_f	The best obtained F-score by information fusion for classification problems
o_m	The best obtained MSE value by information fusion for regression problems
sa	The sacrifice percentage in accuracy performance to gain more explainability
f_t	This is set to 1 if the t^{th} feature should be among selected features; otherwise, set to zero.
e_t	Subjective explainable score of t^{th} feature provided by a domain expert
w_c	The weight of a model's complexity in terms of the number of features
w_o	The subjective opinion's weight
w_f	The importance weight of getting F-score over a threshold in classification problems
w_p	The penalty weight of getting F-score less than a threshold (or MSE higher than a threshold)
w_m	The MSE value's importance in regression problems
	Decision Variables
X_t	This is set to 1 if the t^{th} feature is selected. We denote s as the set of selected features.
y_s	This is set to 1 if the F-score (or accuracy score) of the s^{th} set is higher than the threshold.
N	This represents the number of selected features for each feasible solution.
FS^s	F-score (or accuracy score) of classifier by training model on set s
MSE^s	MSE of regression model by training model on set s

Model 1: Expert-Augmented Feature Selection in Classification Problems

Objective function

$$Max\ Z = \underbrace{w_f\left(\frac{FS^s}{o_f}\right)}_{Part1} + \underbrace{w_o \sum_{t=1}^{T} X_t e_t}_{Part2} + \underbrace{w_c\left(1-\frac{N}{T}\right)}_{Part3} - \underbrace{w_p(1-y_s)\left(\frac{1-FS^s}{1-AT}\right)}_{Part4} \quad (1)$$

Part 1 of Eq. 1 will be activated if the F-score of the s^{th} set is greater than or equal to the predefined threshold (i.e., $y_s = 1$). If activated, it receives the bonus of w_f for any contribution over the threshold. In this scenario, Part 4 of Eq. 1 will be inactive, since $1 - y_s$ will be equal to zero. Now, suppose the F-score of the s^{th} set is lower than the predefined threshold (i.e., $y_s = 0$). In such a scenario, Part 1 of Eq. 1 is inactive, and Part 4 of the objective function will be activated. In this case, that solution will get a penalty of w_p multiplied by $\frac{1-FS^s}{1-AT}$. Part 2 tries to embed the subjective opinion of a domain expert and maximizes it by considering the most important features (from academicians' or practitioners' point of view) among the selected features. Finally, Part 3 aims to minimize the required features to increase the explainability degree of the model.

Constraints

$$X_t \geq f_t \qquad \forall t \in T \quad (2)$$

$$N = \sum_{t=1}^{T} X_t \quad (3)$$

$$o_f - sa = AT \quad (4)$$

$$FS^s - AT \geq M\,(y_s - 1) \qquad \forall s \in S \quad (5)$$

$$FS^s - AT < M\,y_s \qquad \forall s \in S \quad (6)$$

$$X_t \in \{0,1\} \qquad \forall t \in T \quad (7)$$

$$y_s \in \{0,1\} \qquad \forall s \in S \quad (8)$$

Eq. 2 ensures that the t^{th} feature is selected if the user has a strong preference over this feature. Eq. 3 calculates the number of selected features. Eq. 4 defines the desirable accuracy threshold. Eq. 5 and Eq. 6 ensure that y_s takes value one if the F-score (or accuracy score) of the s^{th} set is higher than the accuracy threshold. Otherwise, y_s takes value zero. To be more precise, any of the s^{th} set with an accuracy score below the threshold will be penalized by Part 4 of the objective function. Eq. 7 shows the decision variable X_t can take either zero or one. Eq. 8 shows the decision variable y_s can take either zero or one.

Model 2 is the proposed model that we have developed for regression datasets. Because MSE is one of the most well-known performance measures for regression problems and is different from prediction accuracy of classification problems, we make all of the necessary adjustments to Model 1 in order to have a model that is applicable for regression problems. In Model 2, we modify some of the parameters used in Model 1. For example, in the objective function and Eq. 10, 11, and 12, w_f, o_f, and FS^s are substituted by w_m, o_m, and MSE^s, respectively.

Model 2: Expert-Augmented Feature Selection in Regression-Based Problems

Objective function

$$Max\ Z = \underbrace{w_m\left(\frac{o_m}{MSE^s}\right)}_{Part\,1} + \underbrace{w_o \sum_{t=1}^{T} X_t e_t}_{Part\,2} + \underbrace{w_c\left(1-\frac{N}{T}\right)}_{Part\,3} - \underbrace{w_p(1-y_s)\left(\frac{MSE^s}{AT}\right)}_{Part\,4} \quad (9)$$

The objective function of Model 2 (Eq. 9) is very similar to Model 1's except in Part 1 and Part 4. In datasets with continuous output (regression-based), we use MSE as a performance measure of the trained model. Part 2 and Part 3 are exactly the same as in Model 1, and the explanations provided for them in Model 1 apply here, too.

Constraints

(2), (3), (7), (8)

$$o_m(1+sa) = AT \tag{10}$$

$$AT - MSE^s \geq M\,(y_s - 1) \quad \forall s \in S \tag{11}$$

$$AT - MSE^s < M\,y_s \quad \forall s \in S \tag{12}$$

In Model 2, Eq. 2 , 3, 7, and 8 are repeated from Model 1. Eq. 10 defines the accuracy threshold for regression problems. Eq. 11 and Eq. 12 ensure y_s takes value one if the MSE value of the s^{th} set is lower than a threshold. Otherwise, y_s takes zero.

4.4 Methodology

4.4.1 Proposed algorithms for feature selection

As previously stated, the feature-selection problem is an NP-Hard combinatorial problem (Garey and Johnson, 1979), making it intractable to solve even for problems of medium size. The time required to optimally solve the proposed models grows exponentially with the magnitude of the problem, making it a computationally unfeasible alternative. We would need to go through 2^t combinations of variables to discover the optimal combination for any feature-selection problem with *t* features. For instance, consider a problem with twenty features that requires the training of about one million models and performance evaluation of each. Such an undertaking would be prohibitively costly in terms of both computation and time. To address this issue, evolutionary algorithms are an excellent solution to overcome the challenge of feature selection (Gunavathi and Premalatha, 2015; Serrano-Silva et al., 2018).

A genetic algorithm (GA) is an evolutionary search algorithm that mimics natural selection, to find answers that are close to optimal (Holland, 1975). GAs are well-established meta-heuristic algorithms that have been effectively used in a wide variety of optimization problems, including

portfolio decisions of many large-scale projects (Sampath et al., 2021), location and capacity optimization (Ramirez-Nafarrate et al., 2021), Enterprise Architecture Modelling (Pérez-Castillo et al., 2020), optimization of cloud computing (Mukherjee et al., 2021), and classification tasks (Mannino and Koushik, 2000; Das et al., 2017).

GAs make use of a variety of genetic terminologies and concepts, such as gene, chromosome, population, generation, reproduction, selection, mutation, and crossover. We encourage readers to read Goldberg's (1989) text, which serves as an excellent introduction to GAs. We propose three variants of a GA in this research: GA1, GA2, and GA3. GA1 is essentially a standard GA that randomly generates the initial population. The GA2 and GA3 algorithms are, however, distinct in terms of initial population generation and a few other components that will be discussed later. We begin by introducing all of GA1's elements, then we discuss the distinctions between GA1 and the GA2 and GA3 algorithms.

4.4.1.1 GA1

A solution representation is the antecedent to the implementation of any metaheuristic algorithm, and it should be encoded in such a way that it is fully compatible with the objective function and allows for the generation of all feasible solutions without restriction (Whitley, 1994). Additionally, the efficiency of encoding has a direct effect on the operators for exploration and exploitation in metaheuristic algorithms (Talbi, 2009). Since the number of selected features is a decision variable that is unknown, the solution representation should be structured in such a way that it enables the model to search for all feasible solutions. Table 4-2 contains an illustration of a binary representation of chromosomes with *m* features. The length of each chromosome is equal to the number of features in the dataset under consideration. This solution representation is used for all algorithms throughout this paper.

Table 4-2. Binary chromosome representation for feature-selection problem

Feature index	1	2	3	...	*m*-2	*m*-1	*m*
Value	0	1	1	...	1	0	0

GA creates a population of chromosomes that encode solutions. The term “chromosome” refers to a solution in GA. Following that, each random solution is evaluated against a predefined fitness function.

This fitness function could be adjusted depending on the nature of the problem. For regression, balanced, and imbalanced classification tasks, respectively, we used MSE, Accuracy Score, and F-score as fitness functions. The selection approach establishes the parents for the future generation. To identify viable solutions for the reproduction phase, we used a tournament selection technique. The selection technique is repeated until the population reaches the predefined size. Crossover and mutation are two critical components of GAs that contribute to the algorithm's trade-off between exploitation and exploration in the search space. The crossover operator, which blends features from both parents to produce offspring, provides exploitation on the search space. Mutation occurs within a single individual; mutations diversify the search space by introducing tiny alterations to the population's selected individuals. We next describe the crossover and mutation operators in more detail.

GA specifies the number of children that should undergo the crossover procedure using a crossover probability $(i.e., p_c)$. We use single-point crossover in GA1 to generate high-quality offspring in subsequent generations. A point (gene) on the parent organism strings is selected using this crossover operator. Then, if we combine the left hand side of parent one with the right hand side of parent two, child 1 is generated. Child 2 is made up of parent 1's right and parent 2's left sides.

Similar to crossover, GA uses a mutation probability $(i.e., p_m)$ to assist in specifying the number of offspring to choose for mutation. To increase the diversity of a population, the uniform mutation operator is used. For each chromosome in the case of uniform mutation, GA creates random chromosomal lengths from a uniform distribution ranging from zero to one. If the chromosome's i^{th} generated random number is less than the predefined mutation rate $(i.e., p_{mr})$, the chromosome's i^{th} gene gets flipped. To be more precise, if the original value is zero, it flips to one; if the original value is one, it flips to zero. Otherwise, the gene maintains its current value. In Table 4-3, we illustrate the uniform mutation operator numerically. Once crossover and mutation are complete, GA combines the newly created solutions with the initial population and evaluates the fitness functions of all solutions. Then, all answers will be ranked from greatest to worst. Finally, the new population will be produced by ranking the solutions according to population size and eliminating the remaining solutions, a process referred to as elitism. We define the stopping criterion in this study as hitting a predefined number of iterations $(i.e., Max_{it})$. If MSE accuracy (or F-score) is used as the fitness function, the GA selects and reports the smallest fitness (or highest F-score) of an individual as the optimal solution. Algorithm 1 provides the GA1 pseudo-code.

Table 4-3. A numerical example of the uniform mutation operator for a mutation rate of 0.3

Feature Index	1	2	3	4	5	6	7
The original Solution	0	1	0	1	0	0	1
Random Uniform Chromosome	0.27	0.68	0.54	0.34	0.83	0.71	0.17
Mutated Solution	**1**	1	0	1	0	0	**0**

Algorithm 1. The pseudo code of GA1

Initialization: Randomly generate solutions with the size of pop_{size}

Fitness evaluation: Evaluate the randomly generated solutions in terms of fitness function.

Determine best solution: Pick the best obtained solution and record its value.

For iteration=1 to Max_{it}

Select $p_c * pop_{size}$ solutions and implement the one-point crossover operator on them.

Select $p_m * pop_{size}$ solutions and implement the uniform mutation operator on them.

Merge all new generated solution with current solutions, evaluate them, and rank them.

Replace the population by keeping the best solutions of parents and new generated solutions.

Update the best solution.

End

Report the best solution.

4.4.1.2 GA2

GA2 has a similar primary structure to GA1, with a few minor changes. In GA2, an innovative mechanism is used in the initialization step, in addition to random generation. This technique is referred to as using a Probabilistic Solution Generator via Information Fusion (PSGIF). There are supervised feature-selection techniques that find the essential features for reaching the supervised model's objective (e.g., a classification or regression problem); these techniques depend on the accessibility of labelled data. These algorithms fall into several categories: information-theoretic algorithms such as CMIM (Fleuret, 2004); similarity-based algorithms such as fisher score (Peh et al., 2004); sparse-learning-based algorithms such as RFS (Nie et al., 2010); statistical algorithms such as chi-square (Gu et al., 2012), and streaming-based algorithms such as alpha-investing (Zhou et al., 2005). There have been numerous algorithms developed for feature selection, both classical and non-classical. Some of these algorithms require an input such as the number of desirable features or some form of threshold, while others do not require any input at all to generate a score or weight for the feature's relevance. We offer a general framework that is compatible with a wide variety of algorithms, the output of which will be used to determine an aggregated probability associated with each feature.

The output of feature-selection algorithms can be binary (one if a feature is selected versus zero if it is not selected) or score/weight. In either scenario, these outputs are normalized and used as the input to our proposed approach (PSGIF). It is worth noting that in the case of binary output, all selected features have the same normalized weight, which is not always the case in score/weight output. The following paragraph outlines the PSGIF.

For a PSGIF, several inputs need be stated prior to implementation. The user must choose some feature-selection algorithms (at least two). It should be noted that a greater the number of these algorithms is better, since this increases the reliability of the corresponding probability for each feature. Following that, a probability should be defined to specify the fraction of the initial population generated by the PSGIF ($i.e., p_{PSGIF}$). For instance, if the population size is 100 and $p_{PSGIF} = 0.2$, the PSGIF should generate 20 solutions. The following process involves running each of these feature-selection algorithms and then normalizing their output. After that, if the user has domain knowledge, he or she might apply different weights to the output of these algorithms. Otherwise, the un-normalized score of each feature will be calculated by finding the average value of all algorithms' outputs. Then, we normalize these un-normalized scores to obtain the aggregated probability associated with each feature. Now consider that the PSGIF wants to generate a solution. A random integer between one and the number of features should be created. Assume that the number is three. This implies that the resulting solution should contain three selected features. Now, based on their probability, we use roulette-wheel selection (RWS) to choose one of these features at random. By choosing the roulette wheel, we ensure that the features with the highest aggregated probability have a greater chance of being selected. The PSGIF method is repeated using RWS until three features are picked (three is just for this example). Let us explain PSGIF via a numerical example. Assume the user has chosen five feature-selection algorithms (FSAs),

denoted by the abbreviations FSA1 – FSA5. In this example, we assume that the user lacks domain expertise about the feature-selection techniques and hence cannot assign greater or lesser weights to these algorithms, so they all have equal weights (i.e., $w_{fs} = 0.2$). Assume our dataset has ten features, and the normalized outputs of these five algorithms are given in Table 4-4.

Table 4-4. A numerical example to obtain probabilities of 10 features by PSGIF

		Feature Index										
w_{fs}	Algorithm	1	2	3	4	5	6	7	8	9	10	Sum
0.2	FSA1	0	0.25	0	0	0.25	0	0.25	0	0	0.25	1
0.2	FSA2	0.1	0.15	0.03	0.04	0.12	0.02	0.2	0.07	0.06	0.21	1
0.2	FSA3	0.12	0.1	0.05	0.03	0.16	0.04	0.21	0.06	0.07	0.16	1
0.2	FSA4	0.2	0	0.2	0	0.2	0	0.2	0.2	0	0	1
0.2	FSA5	0.15	0.2	0.09	0.01	0.11	0.03	0.1	0.04	0.03	0.24	1
	Probability	0.114	0.14	0.074	0.016	0.168	0.018	0.192	0.074	0.032	0.172	1

Now, using the probabilities obtained, we build cumulatively ten intervals between zero and one in Table 4-5.

Table 4-5. A numerical example to obtain probabilities of 10 features by PSGIF

Feature Index	1	2	3	4	5
Intervals	[0,114)	[0.114,0.254)	[0.254,0.328)	[0.328,0.344)	[0.344,0.512)
Feature Index	6	7	8	9	10
Intervals	[0.512,0.53)	[0.53,0.722)	[0.722,0.796)	[0.796,0.828)	[0.828, 1]

The PSGIF then generates a random number between 1 and 10 (i.e., the number of features) to determine which features are included in the resulting solution. Assume the random number generated is three. Now, RWS generates a random number between zero and one and determines the interval in Table 4-5 to which that number belongs. Assume the random number generated by RWS is 0.38. As seen in Table 4-5, this value belongs to the fifth interval, indicating that PSGIF

has selected the fifth feature. This approach should be repeated until we have three distinctive features. Algorithm 2 provides the pseudo-code for PSGIF.

Algorithm 2. The pseudo code of PSGIF

```
Step 1: Pick some feature selection methods that fall into the filter type (i.e., classifier independent).
Step 2: Normalize the feature importance (binary/score/rank) obtained by each method.
Step 3: Calculate the aggregate probability of each feature.
Step 4: Calculate the probability intervals of the features.
Step 5: Generate a random integer number between 1 and the number of features.
        While number of features < random integer number
            Use Roulette Wheel Selection (RWS) to randomly choose one of the features.
        End
Step 6: Report the generated solution and its selected features.
```

4.4.1.3 GA3

GA3's primary structure (including the initialization phase) is similar to GA2's, with two exceptions. First, GA3 uses PSGIF as a mutation operator in addition to the uniform mutation operator outlined in GA1. Second, we use an affinity function in conjunction with stochastic universal sampling (SUS) to avoid premature convergence and to increase diversity. We next define how the affinity function is used with the universal sampling approach. First, we define a parameter called p_{AF} (percentage of affinity), which is used to specify the minimum percentage of unique solutions that can be found during each GA3 iteration. If unique solutions are insufficient to cover the population's remaining capacity, GA3 allows for repeated solutions. We then use stochastic universal sampling (SUS) to further diversify the solution pool at each iteration (Baker, 1987). SUS is a slightly modified version of the previously stated roulette wheel selection. We use the same roulette wheel with the same proportions, but rather than using a single selection point and repeatedly spinning the wheel until all required individuals are selected, we turn the wheel only once and use several selection points evenly distributed around the wheel. This way, all individuals are selected concurrently. This allows for the selection of weaker individuals of the

population (based on their fitness) and thus mitigates the unfair aspect of fitness-proportional selection approaches.

4.4.2 Regressors and classifiers

The goal of feature selection is to minimize the number of irrelevant characteristics in a data set without sacrificing efficiency. As this paper establishes a broad framework, we apply our proposed algorithms to two distinct types of problems: regression, and classification. We chose three widely used regression forms: linear regression (LNR), Adaboost (ADB), and Support Vector Regression (SVR), as well as three widely used classifiers: Logistic Regression (LGR), Random Forest (RF), and Support Vector Machine (SVM). These models differ on multiple facets including efficiency, ease of implementation, and analytical tractability; each outperforms in some of these areas more than in other areas.

4.4.3 Parameter tuning

Despite the success of evolutionary algorithms (EAs) in tackling a wide variety of real-world problems, they are dependent on a set of input parameters that affect their performance and must be modified. Indeed, determining and selecting the optimal parameters for an EA is a time-consuming procedure that, in some cases, can be as difficult as the optimization problem at hand.

The task of determining the optimal parameter set for an algorithm can be regarded as a nonlinear optimization problem. In this way, each parameter set x may be considered as a candidate solution in the search space, and the objective value returned by executing the algorithm with x, f(x), can be regarded as a measure of x's quality. Thus, the optimization problem can be written in the following manner:

$$x^* = \underbrace{\arg\min/\max\ f(x)}_{x \in R^d} \tag{13}$$

where d denotes the number of input parameters, and the goal is minimizing/maximizing the predefined objective function. The objective function *f* is a "black box" function of unknown analytic form. A major disadvantage of traditional approaches is the large number of evaluations required to adjust parameters, particularly if evaluating a parameter configuration is time-consuming. In this research, we use an offline approach based on Bayesian optimization (BO) motivated by the characteristics of mentioned scenario (Mockus, 1989; Brochu et al., 2010; Snoek et al., 2012). BO is a universal strategy for optimizing expensive-to-evaluate black-box objective functions. This method has been widely used to adjust the parameters for machine-learning models, and we use it not only for classifiers and predictors but also for our developed feature-selection algorithms (Snoeck et al., 2012). The process of parameter tuning was executed separately for each dataset.

4.4.4 Stability of the developed algorithms

Stability is a critical aspect of any feature-selection technique. Stability refers to the robustness of the feature selections indicated by a feature-selection technique across various training sets derived from the same generating distribution (Somol and Novovicova, 1979). Indeed, the stability of a technique is defined as the stability of specific features following resampling for a given data set. Suppose *Y* is the set of all features and |*Y*| is its cardinality. Per Dunne et al. (2002), the average normalized hamming distance (ANHD) is defined as the stability metric for any feature-selection method. This is the formula for ANHD:

$$ANHD(S) = \frac{2}{|Y|\,n(n-1)} \sum_{i=1}^{n-1} \sum_{j=1}^{n-1} HD(m_i, m_j) \qquad (14)$$

where ANHD(S) is the hamming distance between m_i and m_j, which are two binary outcomes obtained from subsets S_i and S_j, respectively, from different samples of a data set originated from *n* runs. HD is calculated as follows:

$$HD(m_i, m_j) = \sum_{i=1}^{|Y|} \left| m_{i,k} - m_{j,k} \right| \quad (15)$$

where $m_j = (m_{j1}, ..., m_{j|Y|})$ (16)

Table 4-6 provides the stability value of developed algorithms.

Table 4-6. The stability value of developed algorithms

		Logistic Regression			Random Forest			Support Vector Machine		
		GA1	GA2	GA3	GA1	GA2	GA3	GA1	GA2	GA3
Classification	1	0.003	0.002	0.002	0.001	0.003	0.001	0.005	0.004	0.003
	2	0.006	0.006	0.005	0.004	0.003	0.003	0.005	0.005	0.005
	3	0.003	0.003	0.003	0.004	0.002	0.002	0.007	0.005	0.003
	4	0.008	0.006	0.007	0.001	0.002	0.002	0.003	0.003	0.005
	5	0.073	0.052	0.061	0.044	0.023	0.015	0.014	0.023	0.034
	6	0.007	0.005	0.004	0.001	0.002	0.006	0.003	0.002	0.002
	7	0.003	0.003	0.005	0.002	0.006	0.005	0.002	0.002	0.003
	8	0.002	0.002	0.004	0.003	0.003	0.002	0.024	0.003	0.008
		Linear Regression			Adaboost			Support Vector Regression		
		GA1	GA2	GA3	GA1	GA2	GA3	GA1	GA2	GA3
Regression	9	0.012	0.023	0.018	0.003	0.005	0.006	0.074	0.065	0.081
	10	0.019	0.021	0.017	0.002	0.003	0.007	0.002	0.006	0.006
	11	0.004	0.003	0.002	0.002	0.006	0.006	0.004	0.003	0.001
	12	0.008	0.009	0.006	0.014	0.043	0.002	0.001	0.005	0.008
	13	0.005	0.004	0.004	0.006	0.004	0.005	0.006	0.002	0.003
	14	0.002	0.002	0.001	0.009	0.001	0.004	0.002	0.004	0.002
	15	0.005	0.003	0.002	0.007	0.007	0.009	0.007	0.002	0.003
	16	0.002	0.002	0.005	0.002	0.005	0.006	0.008	0.002	0.002

4.4.5 Features' contribution in the prediction accuracy

The majority of models developed by researchers are complicated in terms of their relationship between predictors and outcome. It is extremely difficult to understand the relationship between any one predictor and the outcome in such models. One approach for gaining insight into an individual feature in a complicated model is to fix all other selected features to a single value and then examine the impact of altering the desired feature on the outcome. This method, termed a

partial dependence plot, is oversimplified and gives just a small insight into the true impact of the feature on prediction accuracy. In this section, we develop a posterior ensemble algorithm (PEA) to offer a better estimate of the true impact of features on prediction accuracy.

Algorithm 3 provides the pseudocode of a PEA. We elaborate the steps of this algorithm using a numerical example. Suppose we have a classification dataset with ten features, and we have the output of 12 developed algorithms. This output includes the best F-score obtained by each algorithm and the final solution that determines the selected features. Now, let's assume that the best F-score among these 12 algorithms is 0.97. Then, we should filter the solutions that have F-score lower than 0.873 (0.873 = 0.9*0.97). Let's assume that the best F-score obtained for 8 of these algorithms is a value less than 0.873. So, based on Step 3, we remove these 8 solutions and transfer the rest (four solutions) to the final pool. Table 4-7 shows these four solutions with their F-scores and their selected features.

Algorithm 3. The pseudo code of PEA

Step 1: Record the best solution (i.e., selected features) and prediction accuracy (F-score/MSE) obtained by each of the developed algorithms and create an initial pool of solutions.

Step 2: Find the best prediction accuracy obtained for each classifier in the initial pool and add that algorithm to the second pool. We refer to the best obtained value as "*BestValue*".

Step 3: Eliminate solutions with prediction accuracy lower than 0.95**BestValue* and transfer other solutions to the final pool. 0.95 can be replaced with any value between zero and one.

Step 4: Generate the benchmark solution using the following rule:

Include any feature that is selected by at least one algorithm.

Step 5: Remove one of the selected features in the benchmark solution, then train the model with the rest of the features using all of algorithms in the second pool. Finally, record the best obtained prediction accuracy of algorithms and call it "*NewBestValue*".

Step 6: Calculate the difference between *BestValue and NewBestValue* for all the features.

Step 7: Normalize these differences to estimate the true impact (i.e., probability) of each feature.

Step 8: Report the true impact of each feature.

Table 4-7. F-scores and selected features for final pool solutions

		Feature Index									
F-score	Algorithms	1	2	3	4	5	6	7	8	9	10
0.96	Alg1	1	1	1	0	0	0	1	0	0	1
0.94	Alg2	1	1	0	0	0	0	1	0	0	1
0.97	Alg3	0	1	1	0	1	0	1	0	0	1
0.92	Alg4	1	0	1	0	0	0	1	0	0	1

As can be seen in Table 4-7, four features (4, 6, 8, and 9) are not selected by any of these four solutions. Hence, we eliminate these three features based on rule 1 provided in Step 4. The benchmark solution is created by the rest of the features (1, 2, 3, 5, 6, 7, and 10). According to Step 5, Table 4-8 provides the best obtained F-score after removing any of these features.

Table 4-8. New best obtained F-scores by these four algorithms after feature reduction

		Feature Index					
F-score	Algorithms	1	2	3	5	7	10
0.96	Alg1	0.86	0.90	0.78	0.96	0.74	0.84
0.94	Alg2	0.9	0.82	0.94	0.94	0.83	0.77
0.97	Alg3	0.97	0.88	0.85	0.92	0.82	0.81
0.92	Alg4	0.85	0.92	0.83	0.92	0.86	0.79

Step 6 helps us to estimate the prediction power of that feature. According to this step, Table 4-9 provides the difference between the original best F-score and the new best F-scores after removing each feature.

Table 4-9. Difference between the original best F-score and the new best F-scores after reduction

	Feature Index					
Algorithms	1	2	3	5	7	10
Alg1	0.10	0.06	0.18	0	0.22	0.12
Alg2	0.04	0.12	0	0	0.11	0.17
Alg3	0	0.09	0.12	0.05	0.15	0.16
Alg4	0.07	0	0.09	0	0.06	0.13
Sum	0.21	0.27	0.39	0.05	0.54	0.58
Normalized Impact	0.1	0.13	0.19	0.02	0.26	0.29

Finally, the last row of Table 4-9 provides the true impact of each feature after normalization. These values are called S_t.

Table 4-10 shows that feature #10 has the highest prediction among these features. Since features 4, 6, 8, and 9 were not among the selected features by these four algorithms, these features will be classified as redundant, and their prediction power is set to zero (See Table 4-10).

Table 4-10. Normalized true impact (or prediction power) of features obtained by PEA

Feature Index	1	2	3	4	5	6	7	8	9	10
True Impact	0.1	0.13	0.19	0	0.02	0	0.26	0	0	0.29

Now, let's assume that Table 4-11 provides the normalized and aggregate opinions of academicians (in their papers) or practitioners with respect to features. This is the e_t parameter we defined in our developed mathematical model. We consider $ec_t = e_t - c_t$ as a proxy to determine any feature is properly rated, overrated or underrated. The es_t can technically get any value between -1 and 1.

Table 4-11. Normalized and aggregate opinions of experts

Feature Index	1	2	3	4	5	6	7	8	9	10
Subjective Opinion	0.08	0.17	0.08	0.03	0.08	0.02	0.38	0	0	0.16

Table 4-12 shows the difference between the subjective opinion and the prediction power of features for this numerical example. Figure 4-1 demonstrates the classification of features based on the difference between their subjective opinion and e impact. Table 4-13 provides the classification of these 10 features according to guidelines provided by Figure 4-1.

Table 4-12. The difference between subjective opinion and objective importance of features

Feature Index	1	2	3	4	5	6	7	8	9	10
Difference	-0.02	0.04	-0.11	0.03	0.06	0.02	0.12	0	0	-0.13

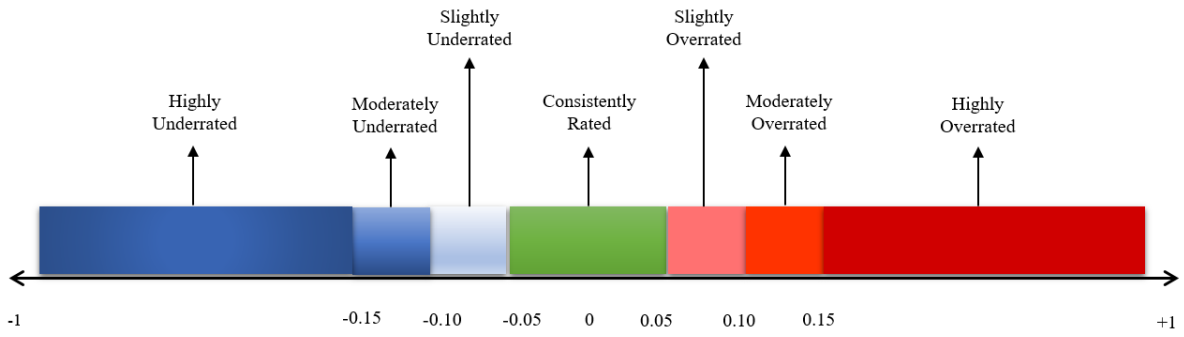

Figure 4-1. Classification of features based on prediction power and subjective opinion

Table 4-13. Classification of features for numerical example

Class	Feature#	Class	Feature#
Consistently Rated	1, 2, 4, 6, 8, 9	Moderately Underrated	7
Slightly Overrated	-	Highly Overrated	-
Slightly Underrated	5	Highly Underrated	-
Moderately Overrated	3, 10		

4.5 Numerical study

This section presents the findings in three subsections: single objective problem, multi-objective problem, and categorization of features.

4.5.1 Single-objective problem

In this section, we compared the obtained value of algorithms where the only objective is the accuracy rate. Besides considering GA1, GA2, and GA3 as feature-selection methods, we tested the classifiers without any feature-selection method; we call it No-Feature-Selection (NFS). Moreover, we used other popular feature-selection methods including LASSO (Tibshirani, 1996), ANOVA, Decision Tree (DT) (Zhou et al. 2021), and CMIM (Fleuret, 2004).

4.5.1.1 Classification datasets

We did the analysis for eight publicly available classification datasets. To prevent repetition, we present the findings here for only one classification dataset: the Wine dataset. The results for the other seven datasets are shown in Appendix A. As indicated earlier, this numerical analysis uses LGR, RF, and SVM classifiers. Figures 4-2, 4-3, and 4-4 illustrate the confidence intervals for algorithms using LGR, RF, and SVM classifiers, respectively, that were tested on the Wine dataset. As these three figures show, GA2 or GA3 surpass the average accuracy score of the other algorithms.

However, when it comes to statistical significance, we should analyze each one independently. As seen in Figure 4-2, GA1, GA2, DT, and CMIM are statistically superior to the other algorithms

when LGR is used as the classifier. However, there is no statistical difference between them. GA3 statistically outperformed all five benchmark algorithms when using RF as the classifier. When SVM is used as a classifier, GA3 statistically outperforms all other algorithms except DT.

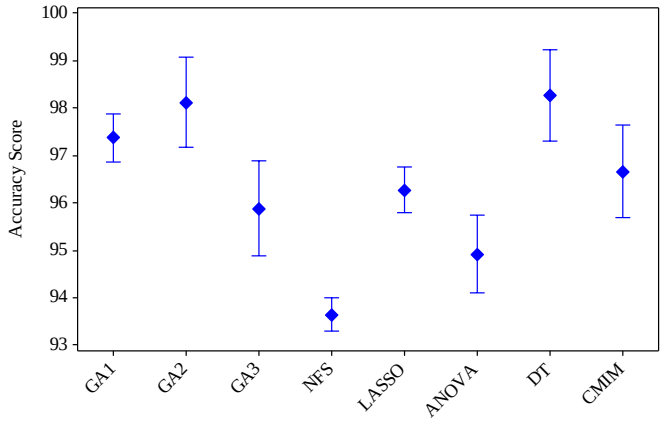

Figure 4-2. 95% CI of accuracy score for Wine dataset with LGR as classifier

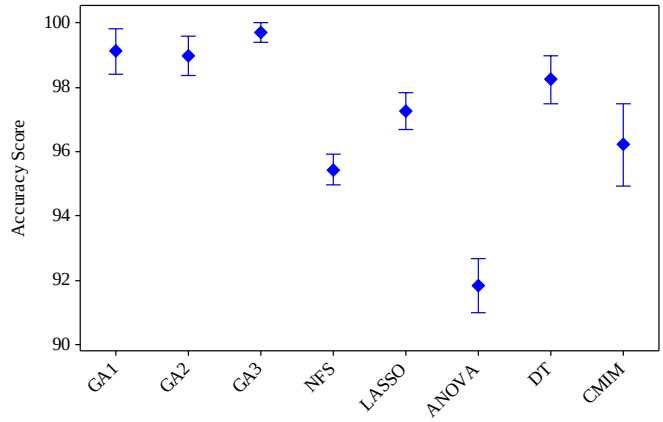

Figure 4-3. 95% CI of accuracy score for Wine dataset with RF as classifier

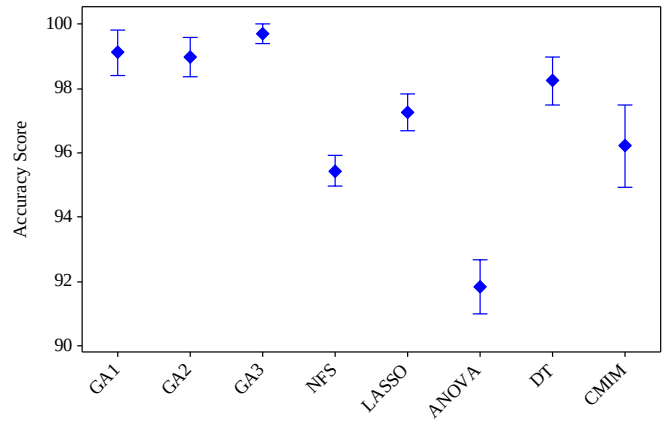

Figure 4-4. 95% CI of accuracy score for Wine dataset with SVM as classifier

The statistical significance of GA variants in classification datasets is summarized in Table 4-14 in comparison to five benchmark algorithms: NFS, LASSO, ANOVA, DT, and CMIM. For instance, there is a statistical difference between at least one of the GA variants and LASSO when LGR is used as a classifier in the Wine dataset. However, as seen in the first row of Table 14, there is no meaningful statistical difference between GA variants and CMIM. These findings show that GA variants dominate NFS and ANOVA variants in the vast majority of cases (i.e., 91.6 percent and 95.8 percent). However, GA variants dominated DT and CMIM in about one-third of the cases, and there was no statistical evidence for the remaining cases.

Finally, Table 4-15 provides the average accuracy gap of all algorithms in all eight classification datasets. As seen in this table, all GA variants are better on average compared to other benchmark algorithms. Interestingly, GA2 performs even better than GA3 in terms of the average. This indicates that, on average, exploration works better than focusing more on exploitation.

4.5.1.2 Regression datasets

The analysis for regression part was conducted on eight publicly available regression datasets. To prevent repetition, we show the findings here for only one of them: the Bike Sharing dataset. The results for the other seven datasets are included in Appendix B. As noted previously, this numerical study uses ADB, LNR, and SVR as regressors. Figures 4-5, 4-6, and 4-7 demonstrate the confidence intervals of MSE for algorithms using ADB, LNR, and SVR regressors, respectively, that were tested on the Bike Sharing dataset. GA3 statistically outperforms all benchmark algorithms when ADB is used as the regressor, as seen in Figure 4-5. When LNR is used as the regressor, GA3 statistically outperforms all benchmark algorithms except LASSO (see Figure 4-6). When we use SVR as the regressor instead of LNR, GA3's performance is nearly the same, except for the point that it statistically dominates GA1.

Table 4-14. Summary of statistical significance comparison for classification datasets

		NFS	LASSO	ANOVA	DT	CMIM
Wine	LGR	✓	✓	✓	-	-
	RF	✓	✓	✓	✓	✓
	SVM	✓	✓	✓	-	-
Breast Cancer	LGR	✓	✓	✓	✓	✓
	RF	✓	✓	✓	-	✓
	SVM	✓	✓	✓	✓	✓
Lung Cancer	LGR	✓	-	-	-	-
	RF	-	-	✓	-	✓
	SVM	✓	✓	✓	✓	✓
Parkinson	LGR	✓	-	✓	✓	-
	RF	-	-	✓	-	-
	SVM	✓	-	✓	-	-
Spambase	LGR	✓	-	✓	-	-
	RF	✓	-	✓	-	-
	SVM	✓	✓	✓	-	-
Arrythmia	LGR	✓	✓	✓	-	✓
	RF	✓	✓	✓	-	✓
	SVM	✓	-	✓	-	-
Hill-Valley	LGR	✓	-	✓	-	-
	RF	✓	✓	✓	✓	✓
	SVM	✓	✓	✓	✓	-
Sonar	LGR	✓	✓	✓	✓	-
	RF	✓	✓	✓	-	-
	SVM	✓	-	✓	-	-
Overall		91.6%	58.3%	95.8%	33.3%	37.5%

Table 4-15. Summary of average accuracy gap for classification datasets

	GA1	GA2	GA3	NFS	LASSO	ANOVA	DT	CMIM	Average
LGR	6.47	5.57	5.86	11.42	7.82	10.00	6.56	6.80	7.56
RF	1.08	0.55	0.54	6.79	3.23	7.83	1.54	3.22	3.10
SVM	3.07	2.60	2.64	9.49	4.45	8.79	3.15	3.41	4.70
Average Gap	3.54	2.90	3.01	9.23	5.17	8.87	3.75	4.47	

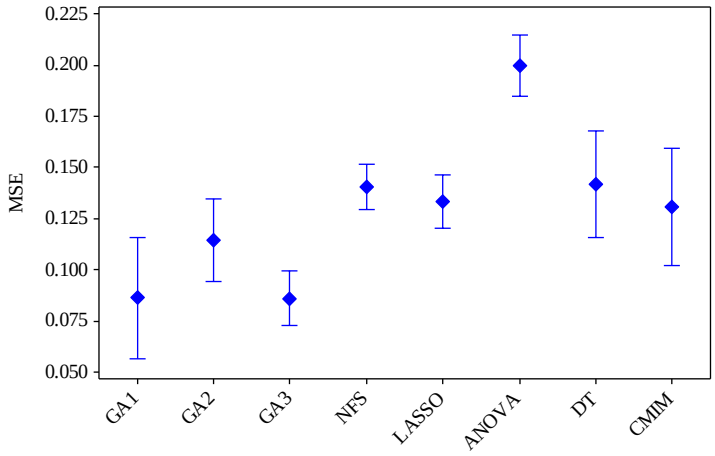

Figure 4-5. 95% CI of MSE for Bike-Sharing dataset with ADB as regressor

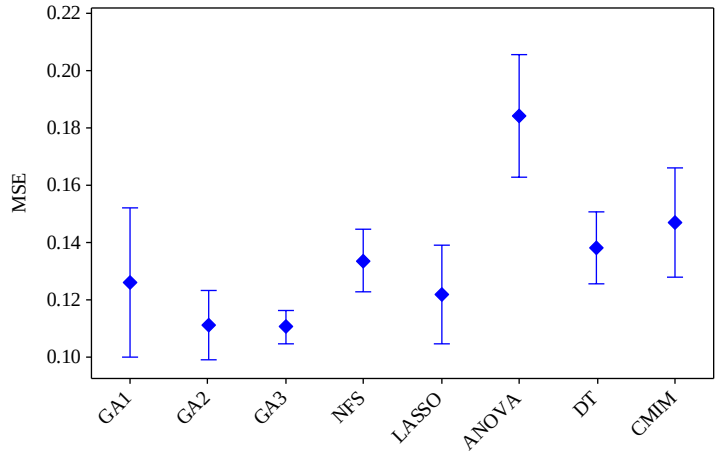

Figure 4-6. 95% CI of MSE for Bike-Sharing dataset with LNR as regressor

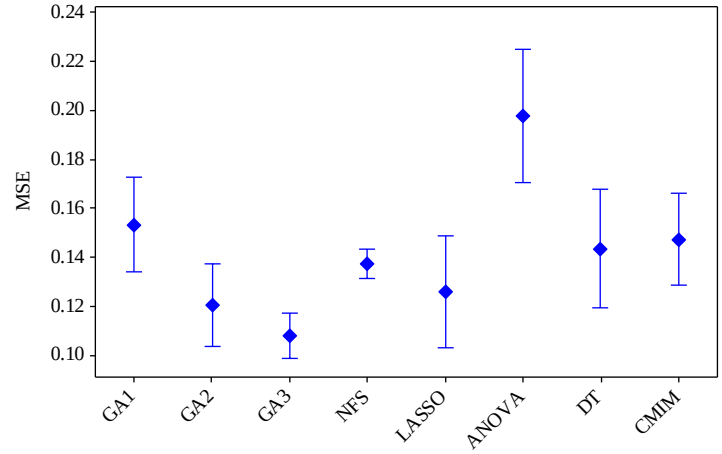

Figure 4-7. 95% CI of MSE for Bike-Sharing dataset with SVR as regressor

The statistical significance of GA variants in regression datasets is summarized in Table 4-16 in comparison to five benchmark algorithms: NFS, LASSO, ANOVA, DT, and CMIM. The performance of GA variants versus NFS and ANOVA in regression datasets is very similar, and GA variants dominate these two algorithms in most cases. However, when LASSO is used as the feature-selection algorithm, GA variants could outperform LASSO in just 29.16% of the cases. We can argue the same for the comparison between GA variants and CMIM, since GA variants are statistically better than CMIM in one-third of the cases.

Finally, Table 4-17 provides the average accuracy gap of all algorithms in all eight regression datasets. As seen in this table, all GA variants are better on average compared to other benchmark algorithms. The best performance in terms of average MSE goes to GA3. After GA variants, LASSO has the best performance with an average MSE of 0.032. ADB exhibits the best performance among the three regressors with an average MSE of 0.033.

Table 4-16. Summary of statistical significance comparison for regression datasets

		NFS	LASSO	ANOVA	DT	CMIM
AutoMPG	ADB	-	-	-	-	-
	LNR	✓	✓	✓	-	-
	SVR	✓	-	✓	-	-
Bike Sharing	ADB	✓	✓	✓	✓	✓
	LNR	✓	-	✓	✓	✓
	SVR	✓	-	✓	✓	✓
Gas Sensor	ADB	✓	✓	✓	✓	-
	LNR	✓	-	✓	✓	✓
	SVR	✓	-	✓	-	-
Stock Portfolio	ADB	✓	-	✓	✓	-
	LNR	✓	-	✓	✓	-
	SVR	✓	-	✓	✓	-
Red Wine	ADB	✓	✓	✓	✓	-
	LNR	✓	-	✓	✓	-
	SVR	✓	✓	✓	✓	✓
Life Expectancy	ADB	✓	-	✓	✓	-
	LNR	✓	-	✓	✓	✓
	SVR	-	-	✓	-	-
Blog Feedback	ADB	✓	✓	✓	✓	-
	LNR	✓	-	✓	✓	-
	SVR	✓	-	✓	✓	-
Bodily Expression	ADB	✓	-	✓	✓	-
	LNR	✓	-	✓	✓	✓
	SVR	✓	✓	✓	✓	✓
Overall		91.6%	29.16%	95.8%	79.16%	33.3%

Table 4-17. Summary of average MSE gap for regression datasets

	GA1	GA2	GA3	NFS	LASSO	ANOVA	DT	CMIM	Average
ADB	0.010	0.013	0.005	0.038	0.031	0.087	0.047	0.035	0.033
LNR	0.028	0.020	0.012	0.048	0.032	0.105	0.040	0.043	0.041
SVR	0.024	0.021	0.012	0.039	0.032	0.103	0.046	0.036	0.039
Average Gap	0.021	0.018	0.010	0.042	0.032	0.098	0.044	0.038	

4.5.2 Multi-objective problem

The Lung Cancer dataset was used to analyze our proposed multi-objective model in this section. Since one of our model's key parameters was the subjective importance of features (e_t), we estimated it by examining 100 articles published in this domain. The Lung Cancer dataset contains 23 variables classified either as risk factors or symptoms.

For this estimation, the following four steps were taken.

I. Fifty papers were identified using the keywords "Risk Factors for Lung Cancer."
II. Fifty papers were selected using the keywords "Lung Cancer Symptoms."
III. The frequency of use of each of the 23 features was determined in the 100 papers.
IV. The frequency counts of usage have been normalized and transformed to probabilities.

We followed all the aforementioned steps to determine the subjective importance of these 23 features. (See Table 4-18.)

The review of the literature for these 100 papers indicated that, from the perspective of medical experts, age, gender, smoking status, and passive smoking are the most influential risk factors for lung cancer. Additionally, the symptoms most often reported in lung cancer research are dry cough, chest pain, weight loss, and fatigue.

Table 4-18. Estimated subjective importance of features of Lung Cancer dataset

Feature	Age	Gender	Air Pollution	Alcohol Use	Dust Allergy	Occupational Hazards
Probability	14.1	12.3	2.2	2	0.4	2.6
Feature	Genetic Risk	Chronic Lung Disease	Balanced Diet	Obesity	Smoking	Passive Smoker
Probability	3.4	1.8	0.6	4	12.5	8.6
Feature	Chest Pain	Coughing of Blood	Fatigue	Weight Loss	Shortness of Breath	Wheezing
Probability	6.5	1.8	5.2	6	4	1
Feature	Swallowing Difficulty	Clubbing of Finger Nails	Frequent Cold	Dry Cough	Snoring	
Probability	0.8	2.4	0.2	7.2	0.4	

Following the estimation of subjective importance, we develop an experiment design to draw insights from the model. In this section, we outline four important parameters: the Forced Variables Threshold (FVT), the Sacrifice Percentage in Accuracy (SPA), and the Weight Allocation Approach (WAA). In FVT, we consider two distinct levels: 8% and 12%. We use this threshold to determine which features must be selected by the algorithm (when $f_t = 1$). For SPA, we consider two levels: 1% and 2%. We examine three approaches for weight allocation: Accuracy Oriented (AO), Equally Distributed (ED), and Subjective Oriented (SO). We evaluate three distinct weight sets for AO and SO and a single weight set for ED, resulting in a total of seven distinct weight sets. Table 4-19 shows the weights allocated to w_f, w_o, and w_c. Finally, three alternative values for the multiplier of the penalty function (i.e., w_p) are considered: 0 (i.e., no penalty), 1, and 2. This experiment design generates 84 potential scenarios. All these scenarios were implemented using the model.

Table 4-19. The weights assigned to different parts of the objective function in each approach

WAA	WS#	w_f	w_o	w_c
AO	1	0.7	0.2	0.1
	2	0.7	0.1	0.2
	3	0.8	0.05	0.15
ED	4	0.5	0.25	0.25
SO	5	0.4	0.4	0.2
	6	0.3	0.4	0.3
	7	0.3	0.2	0.5

Here, we highlight some key insights. First, there are six unique solutions for all these scenarios. In other words, one of these six solutions is the optimal solution in multiple scenarios. This demonstrates that there are some solutions that are dominant in the solution space. With an FVT of 8% or 12%, three features are always required to be zero: Age, Gender, and Smoking. This shows that there are some non-dominated solutions in the solution space which dominate the rest

of solutions independent of objective functions' assigned weights. Apart from these three features, the algorithm picked the "passive smoking" feature in each of these six unique optimal solutions. This demonstrates that both current smoking status and passive smoking are strong predictors of lung cancer. The other frequently selected features among these six unique solutions are "Dry Cough", "Weight Loss" and "Fatigue". These three features are selected in five out of six solutions. Five of these six solutions have an accuracy rate of 97.53 percent to 98.76 percent, while the least accurate solution has an accuracy rate of 94.4 percent. As predicted, the scenario with the highest subjective component weights (total 0.7) results in the lowest accuracy rate. This indicates that the maximum required sacrifice in accuracy rate to gain more explainability is around 5% (94.4 %). While the findings from this numerical analysis are not generalizable to other datasets, it demonstrates that there may be some opportunity to compromise accuracy in order to obtain greater explainability.

4.5.3 Estimation of the contribution of each feature

This section implements our suggested method on the Lung Cancer dataset, to quantify the contribution of each feature. Table 4-20 summarizes the objective importance of features that were obtained by posterior ensemble algorithm. As previously noted, the difference between subjective and objective importance (ec_t) is used as a proxy for classifying features. The difference value for each of these features is presented in Table 4-21. This table indicates that 18 features are consistently rated. The most overrated features are "Gender" and "Age", respectively. The "Balanced diet" and "Alcohol Use" are the most underrated features.

Table 4-20. Estimated objective importance of features of Lung Cancer dataset

Feature	Age	Gender	Air Pollution	Alcohol Use	Dust Allergy	Occupational Hazards
Probability	5.35	2.01	5.75	6.18	4.21	1.72
Feature	Genetic Risk	Chronic Lung Disease	Balanced Diet	Obesity	Smoking	Passive Smoker
Probability	5.64	3.09	5.64	7.80	6.61	2.20
Feature	Chest Pain	Coughing of Blood	Fatigue	Weight Loss	Shortness of Breath	Wheezing
Probability	6.97	5.67	3.71	7.99	2.36	1.77
Feature	Swallowing Difficulty	Clubbing of Finger Nails	Frequent Cold	Dry Cough	Snoring	
Probability	2.87	2.92	3.09	2.90	3.55	

Table 4-21. Difference between subjective and objective importance of features of Lung Cancer dataset

Feature	Age	Gender	Air Pollution	Alcohol Use	Dust Allergy	Occupational Hazards
Probability	8.75	10.29	-3.55	-4.18	-3.81	0.88
Feature	Genetic Risk	Chronic Lung Disease	Balanced Diet	Obesity	Smoking	Passive Smoker
Probability	-2.24	-1.29	-5.04	-3.80	5.89	6.40
Feature	Chest Pain	Coughing of Blood	Fatigue	Weight Loss	Shortness of Breath	Wheezing
Probability	-0.47	-3.87	1.49	-1.99	1.64	-0.77
Feature	Swallowing Difficulty	Clubbing of Finger Nails	Frequent Cold	Dry Cough	Snoring	
Probability	-2.07	-0.52	-2.89	4.30	-3.15	

4.6 Conclusion

Feature selection has become a challenging issue in recent years, particularly in supervised machine-learning disciplines such as classification and regression. In this paper, we examined the supervised feature-selection problem by directly incorporating experts' opinions with prediction accuracy and complexity. This study makes several contributions to the literature.

As one of our algorithmic contributions, we develop a probabilistic solution generator using an information fusion (PSGIF) method that serves as an ensemble operator, gathering data from many

filter approaches in order to improve the genetic algorithm's exploration and exploitation process. Additionally, we use Bayesian optimization to tune the parameters of our method in order to reduce the algorithm's processing time, which contributes to the scalability of our proposed method. Using sixteen publicly available classification or regression datasets, we compare the performance of our proposed algorithms to that of several well-known algorithms. We use three variants of genetic algorithms (GA1, GA2, and GA3), the No Feature Selection technique (NFS), LASSO, ANOVA, and CMIM for the feature selection. Three variants of genetic algorithms (i.e., GA1-GA3), the No Feature Selection technique (NFS), LASSO, ANOVA, and CMIM are used for feature selection. We consider three well-known classifiers: logistic regression, random forest, and support vector machines, as well as three well-known regressors: AdaBoost (ADB), Linear Regression (LNR), and Support Vector Regressor (SVR). The analysis shows that GA versions outperform other benchmark algorithms in terms of average accuracy rate for classification datasets and average MSE for regression datasets.

Second, we develop a multi-objective mathematical model that accounts for the trade-off between complexity, explainability, and prediction accuracy. We allow the decision-maker (or domain expert) to indicate a percentage of accuracy loss in exchange for possibly increasing explainability by limiting the number of selected features and selecting those that are more influential from the experts' perspective.

Our proposed multi-objective model was analyzed using the Lung Cancer dataset. We quantified the subjective importance of lung cancer features by reviewing one hundred studies published in this medical field and assessing risk factors and symptoms. According to the literature, the most important risk factors for lung cancer are age, gender, smoking status, and passive smoking, and

the symptoms most often reported in lung cancer research are dry cough, chest discomfort, weight loss, and fatigue.

After determining the subjective importance of features, we developed a posterior ensemble approach for computing the accuracy-contribution degree of all features. As an outcome of this method, we determine features that academics regularly underrate, overrate, or consistently rate. We discovered that “Gender” and “Age” are the most overrated features, respectively, by comparing their subjective and objective values. Additionally, Balanced Diet appears to be significantly underrated, with “Alcohol Use” being the next most underrated feature.

We have limited ourselves to supervised learning in this study. However, the techniques and model provided here may be expanded to support unsupervised learning as well. Enhancing the algorithm's performance through the development of innovative ensemble models deserves additional investigation. Another avenue for study would be to apply our multi-objective model to real-world situations and evaluate the model's impact on experts' attitudes toward future decision-making.

References Cited

Agarwal, P., & Owzar, K. (2014). Next generation distributed computing for cancer research. Cancer informatics, 13, CIN-S16344.

Andersen, C. M. , & Bro, R. (2010). Variable selection in regression –A tutorial. Journal of Chemometrics, 24 (11–12), 728–737 .

Baker, James E. (1987). “Reducing Bias and Inefficiency in the Selection Algorithm”. *Proceedings of the Second International Conference on Genetic Algorithms and Their Application*. Hillsdale, New Jersey: L. Erlbaum Associates: 14–21.

Bertolazzi, P., Felici, G., Festa, P., Fiscon, G., & Weitschek, E. (2016). Integer programming models for feature selection: New extensions and a randomized solution algorithm. European Journal of Operational Research, 250(2), 389-399.

Bertolazzi, P., Felici, G., Festa, P., Fiscon, G., & Weitschek, E. (2016). Integer programming models for feature selection: New extensions and a randomized solution algorithm. European Journal of Operational Research, 250(2), 389-399.

Chandrashekar, G. , & Sahin, F. (2014). A survey on feature selection methods. Computers & Electrical Engineering, 40 (1), 16–28.

Coussement, K., Benoit, D. F., & Antioco, M. (2015). A Bayesian approach for incorporating expert opinions into decision support systems: A case study of online consumer-satisfaction detection. Decision Support Systems, 79, 24-32.

Das, A. K., Das, S., & Ghosh, A. (2017). Ensemble feature selection using bi-objective genetic algorithm. *Knowledge-Based Systems*, 123, 116-127.

De Caigny, A., Coussement, K., & De Bock, K. W. (2018). A new hybrid classification algorithm for customer churn prediction based on logistic regression and decision trees. European Journal of Operational Research, 269(2), 760-772.

Elhamifar, E., & Vidal, R. (2013). Sparse subspace clustering: Algorithm, theory, and applications. IEEE transactions on pattern analysis and machine intelligence, 35(11), 2765-2781.

Emary, E., Zawbaa, H. M., & Hassanien, A. E. (2016). Binary grey wolf optimization approaches for feature selection. Neurocomputing, 172, 371-381.

Fleuret, F. (2004). Fast binary feature selection with conditional mutual information. *Journal of Machine learning research*, 5(9).

Fleuret, F. (2004). Fast binary feature selection with conditional mutual information. Journal of Machine learning research, 5(9).

Garey, M. R., & Johnson, D. S. (1979). Computers and intractability (Vol. 174). San Francisco: *freeman*.

Gu, Q., Li, Z., & Han, J. (2012). Generalized fisher score for feature selection. arXiv preprint arXiv:1202.3725.

Gunavathi, C., & Premalatha, K. (2015). Cuckoo search optimisation for feature selection in cancer classification: a new approach. *International journal of data mining and bioinformatics*, 13(3), 248-265.

Holland, J. H. (1975). Adaptation in natural and artificial systems. Ann Arbor, MI: *University of Michigan Press* .

Kira, K., & Rendell, L. A. (1992). A practical approach to feature selection. In Machine learning proceedings 1992 (pp. 249-256). Morgan Kaufmann.

Li, A. D., He, Z., Wang, Q., & Zhang, Y. (2019). Key quality characteristics selection for imbalanced production data using a two-phase bi-objective feature selection method. European Journal of Operational Research, 274(3), 978-989.

Li, J., Cheng, K., Wang, S. , Morstatter, F. , Trevino, R. P. , Tang, J. , & Liu, H. (2017). Feature selection: A data perspective. ACM Computing Surveys, 50 (6), 94:1–94:45

Mannino, M. V., & Koushik, M. V. (2000). The cost-minimizing inverse classification problem: a genetic algorithm approach. *Decision Support Systems*, 29(3), 283-300.

Mao, Q., & Tsang, I. W. H. (2012). A feature selection method for multivariate performance measures. IEEE transactions on pattern analysis and machine intelligence, 35(9), 2051-2063.

Mukherjee, A., Sundarraj, R., & Dutta, K. (2021). Time-preference-based on-spot bundled cloud-service provisioning. *Decision Support Systems*, 113607.

Nie, F., Huang, H., Cai, X., & Ding, C. (2010). Efficient and robust feature selection via joint ℓ2, 1-norms minimization. *Advances in neural information processing systems*, 23.

Oosterlinck, D., Benoit, D. F., & Baecke, P. (2020). From one-class to two-class classification by incorporating expert knowledge: Novelty detection in human behaviour. European Journal of Operational Research, 282(3), 1011-1024.

Oreski, S., Oreski, D., & Oreski, G. (2012). Hybrid system with genetic algorithm and artificial neural networks and its application to retail credit risk assessment. Expert systems with applications, 39(16), 12605-12617.

PEH, D. (2007). RO Duda, PE Hart, and DG Stork, Pattern Classification, New York: John Wiley & Sons, 2001, pp. xx+ 654, ISBN: 0-471-05669-3. *Journal of Classification*, 24(2), 305-307.

Pérez-Castillo, R., Ruiz, F., & Piattini, M. (2020). A decision-making support system for Enterprise Architecture Modelling. Decision Support Systems, 131, 113249.

Pudil, P., Novovičová, J., & Kittler, J. (1994). Floating search methods in feature selection. Pattern recognition letters, 15(11), 1119-1125.

Ramirez-Nafarrate, A., Araz, O. M., & Fowler, J. W. (2021). Decision assessment algorithms for location and capacity optimization under resource shortages. *Decision Sciences*, 52(1), 142-181.

Sampath, S., Gel, E. S., Kempf, K. G., & Fowler, J. W. (2021). A generalized decision support framework for large-scale project portfolio decisions. *Decision Sciences*.

Serrano-Silva, Y. O., Villuendas-Rey, Y., & Yáñez-Márquez, C. (2018). Automatic feature weighting for improving financial Decision Support Systems. *Decision Support Systems*, 107, 78-87.

Shmueli, G. (2010). To explain or to predict? *Statistical Science*, 25(3):289–310.

Song, L., Bedo, J., Borgwardt, K. M., Gretton, A., & Smola, A. (2007). Gene selection via the BAHSIC family of algorithms. *Bioinformatics*, 23(13), i490-i498.

Talbi, E. G. (2009). Metaheuristics: from design to implementation (Vol. 74). *John Wiley & Sons*.

Tang, J., Alelyani, S., & Liu, H. (2014). Feature selection for classification: A review. In Data Classification: Algorithms and Applications (pp. 37–64) . CRC Press. https: //doi.org/10.1201/b17320.

Tarca, A. L., Carey, V. J., Chen, X. W., Romero, R., & Drăghici, S. (2007). Machine learning and its applications to biology. *PLoS computational biology*, 3(6), e116.

Taylor, A., Steinberg, J., Andrews, T. S., & Webber, C. (2015). GeneNet Toolbox for MATLAB: a flexible platform for the analysis of gene connectivity in biological networks. Bioinformatics, 31(3), 442-444.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society: Series B (Methodological), 58(1), 267-288.

Verbeke, W., Dejaeger, K., Martens, D., Hur, J., & Baesens, B. (2012). New insights into churn prediction in the telecommunication sector: A profit driven data mining approach. *European journal of operational research*, 218(1), 211-229.

Verboven, S., Berrevoets, J., Wuytens, C., Baesens, B., & Verbeke, W. (2020). Autoencoders for strategic decision support. *Decision Support Systems*, 113422.

Whitley, D. (1994). A genetic algorithm tutorial. *Statistics and computing*, 4(2), 65-85.

Zhou, H., Zhang, J., Zhou, Y., Guo, X., & Ma, Y. (2021). A feature selection algorithm of decision tree based on feature weight. *Expert Systems with Applications*, 164, 113842.

Zhou, J., Foster, D., Stine, R., & Ungar, L. (2005, August). Streaming feature selection using alpha-investing. *In Proceedings of the eleventh ACM SIGKDD international conference on Knowledge discovery in data mining* (pp. 384-393).

Appendix A

Chapter 4: 95% CI of Accuracy Score Figures for Classification Datasets

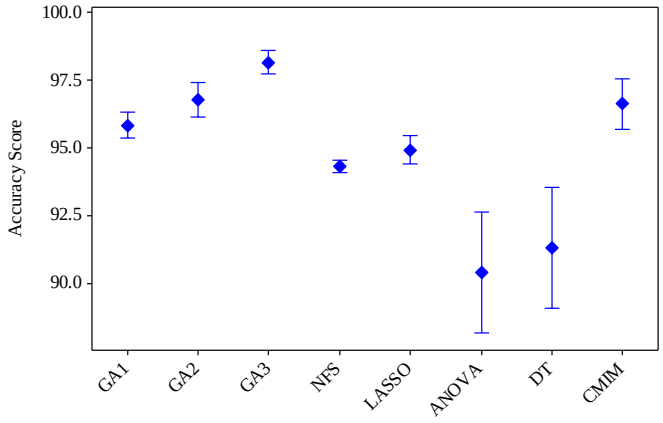

Figure A-1. 95% CI of accuracy score for Breast Cancer dataset with LGR as classifier

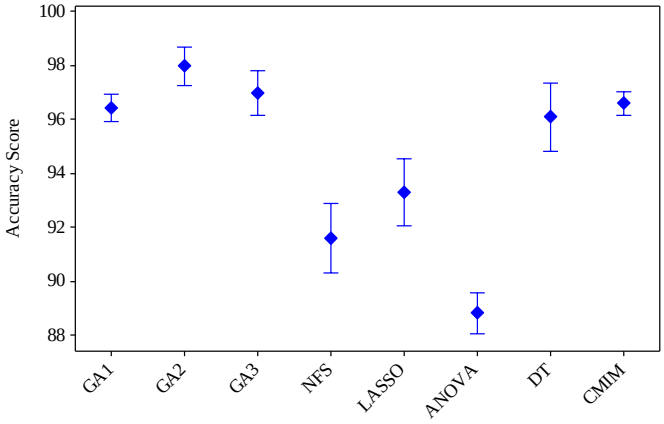

Figure A-2. 95% CI of accuracy score for Breast Cancer dataset with RF as classifier

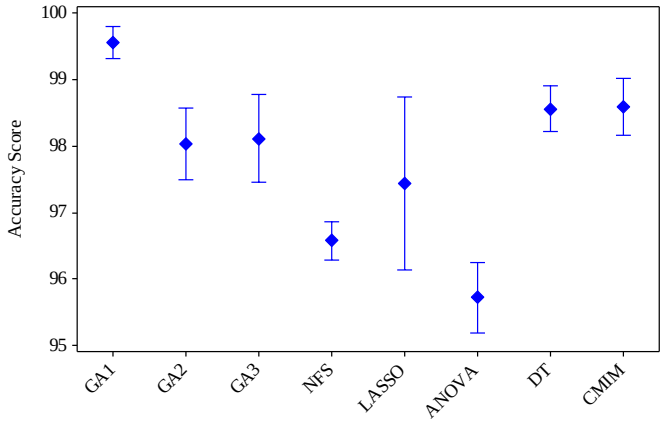

Figure A-3. 95% CI of accuracy score for Breast Cancer dataset with SVM as classifier

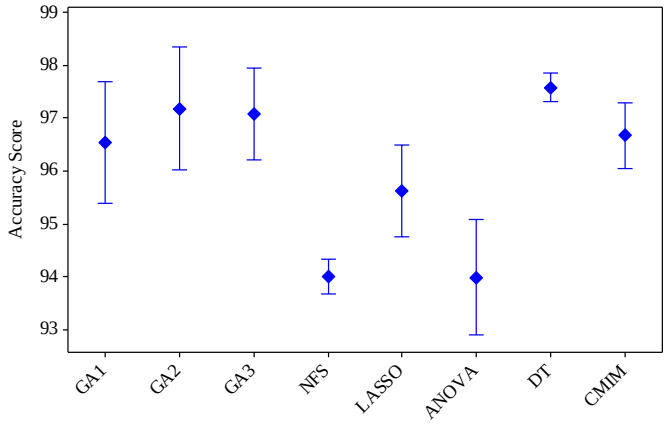

Figure A-4. 95% CI of accuracy score for Lung Cancer dataset with LGR as classifier

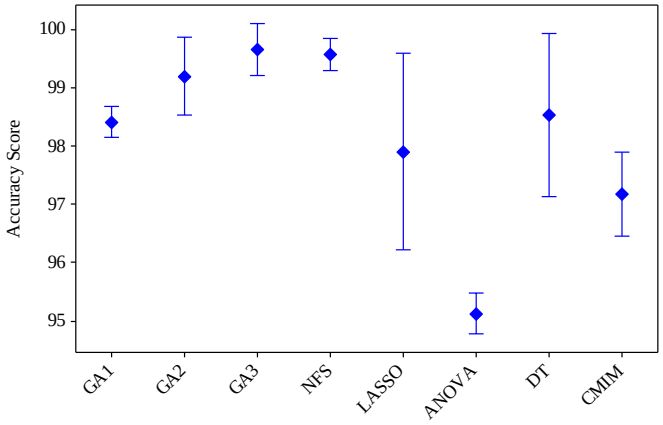

Figure A-5. 95% CI of accuracy score for Lung Cancer dataset with RF as classifier

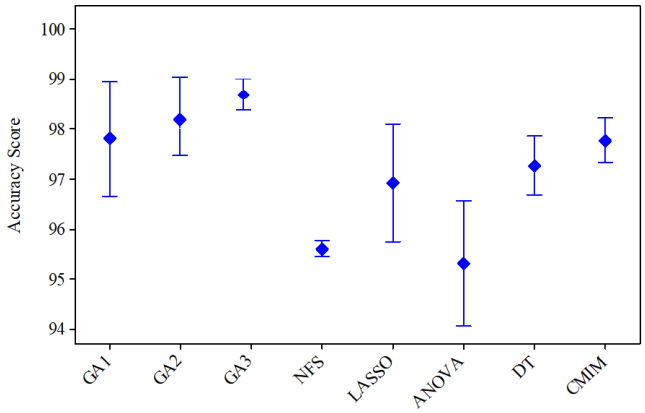

Figure A-6. 95% CI of accuracy score for Lung Cancer dataset with SVM as classifier

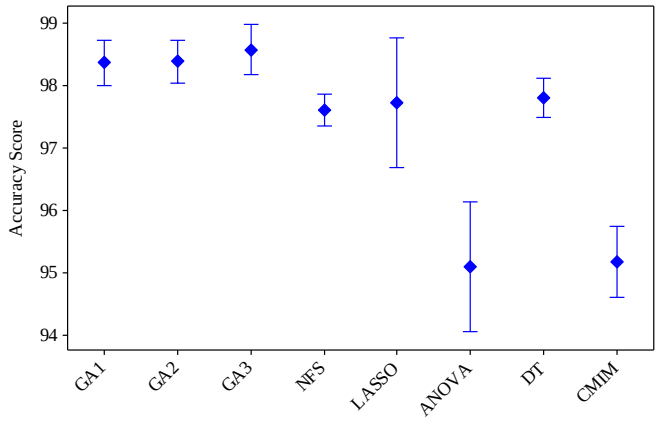

Figure A-7. 95% CI of accuracy score for Parkinson dataset with LGR as classifier

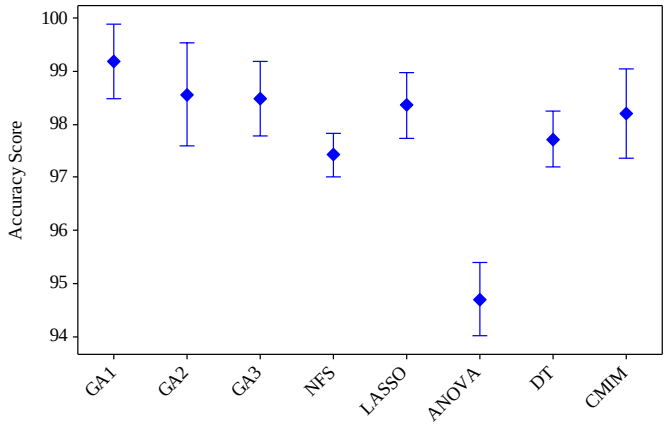

Figure A-8. 95% CI of accuracy score for Parkinson dataset with RF as classifier

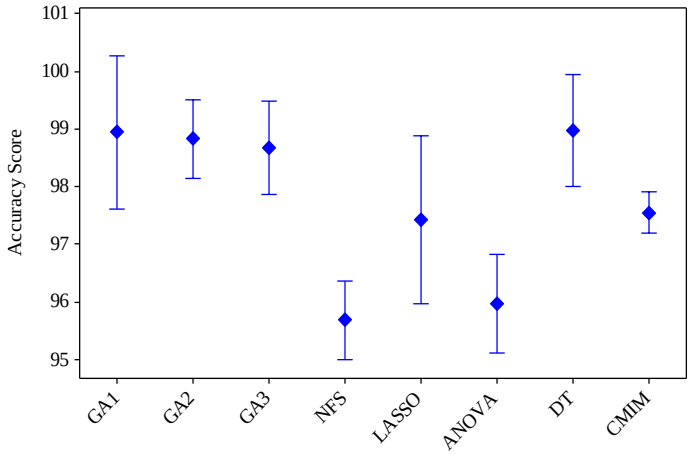

Figure A-9. 95% CI of accuracy score for Parkinson dataset with SVM as classifier

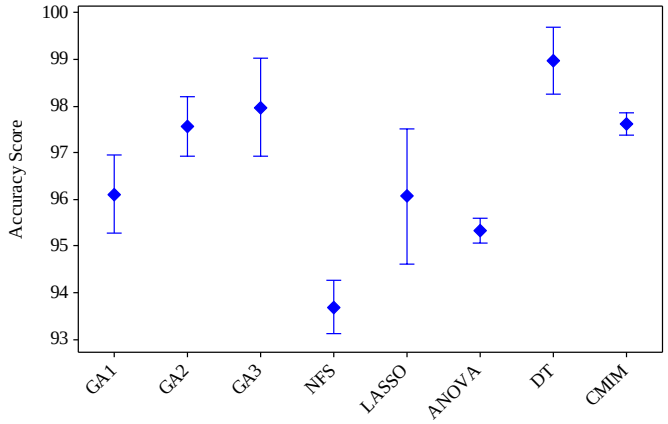

Figure A-10. 95% CI of accuracy score for Spambase dataset with LGR as classifier

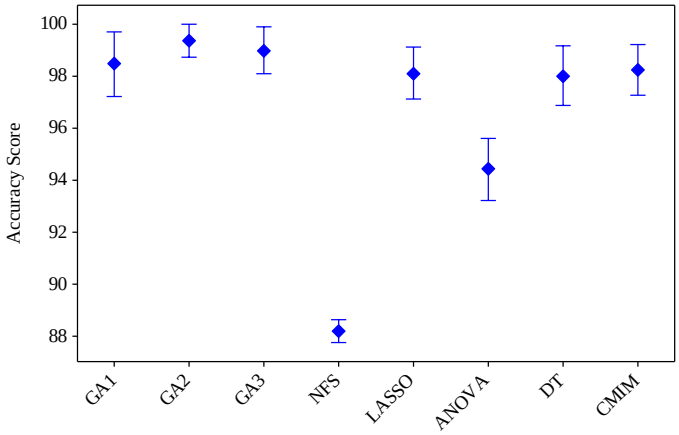

Figure A-11. 95% CI of accuracy score for Spambase dataset with RF as classifier

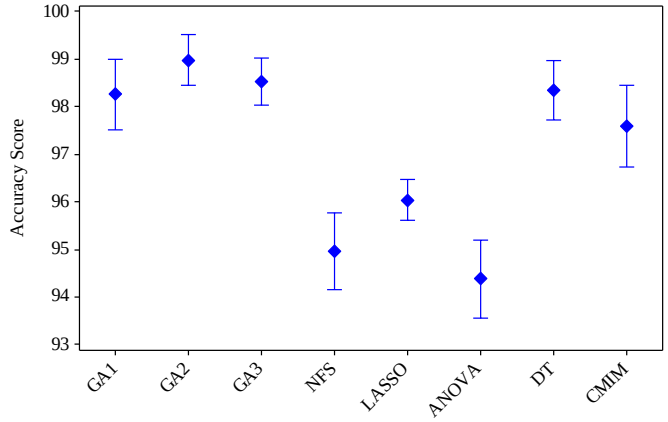

Figure A-12. 95% CI of accuracy score for Spambase dataset with SVM as classifier

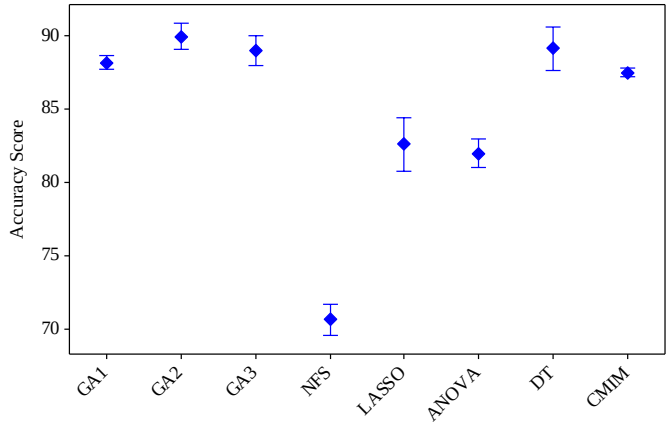

Figure A-13. 95% CI of accuracy score for Arrythmia dataset with LGR as classifier

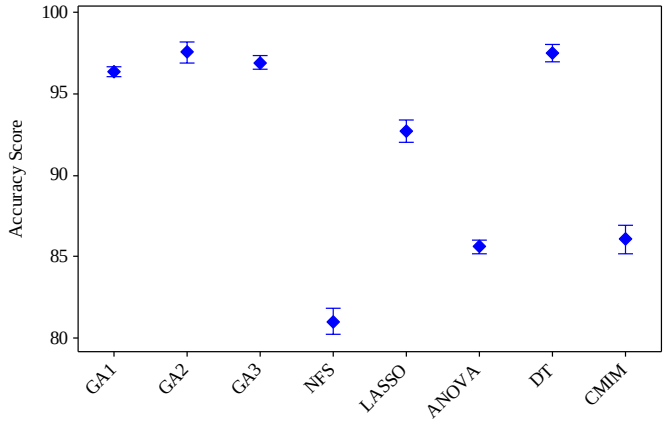

Figure A-14. 95% CI of accuracy score for Arrythmia dataset with RF as classifier

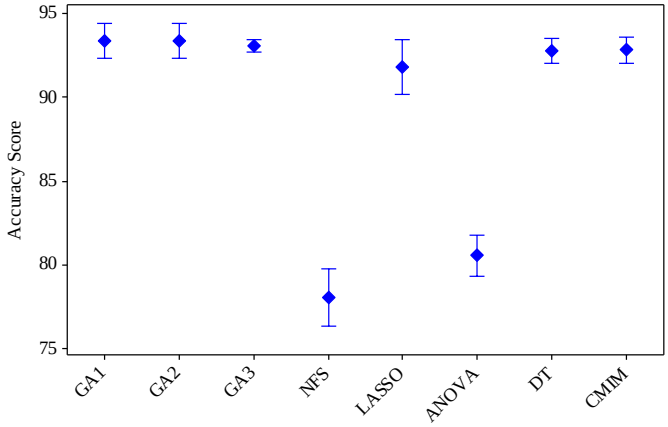

Figure A-15. 95% CI of accuracy score for Arrythmia dataset with SVM as classifier

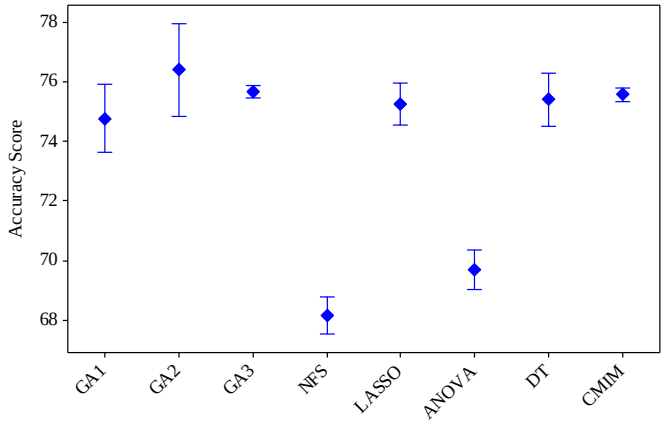

Figure A-16. 95% CI of accuracy score for Hill-Valley dataset with LGR as classifier

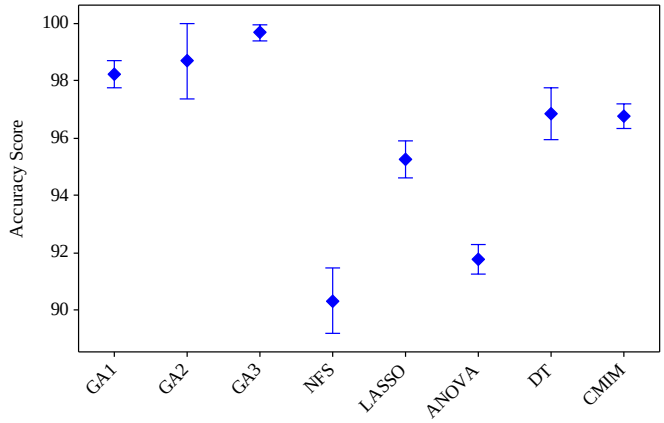

Figure A-17. 95% CI of accuracy score for Hill-Valley dataset with RF as classifier

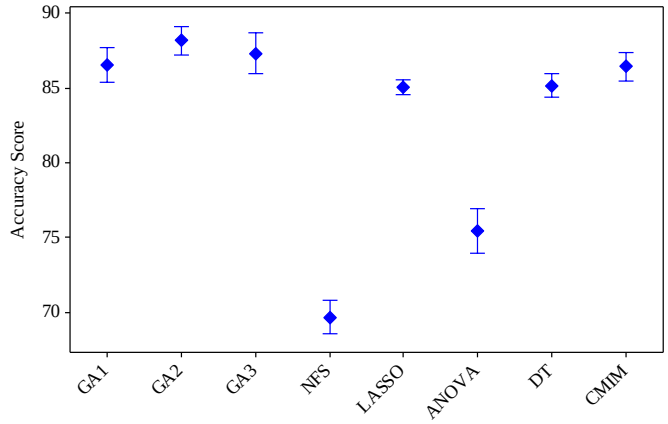

Figure A-18. 95% CI of accuracy score for Hill-Valley dataset with SVM as classifier

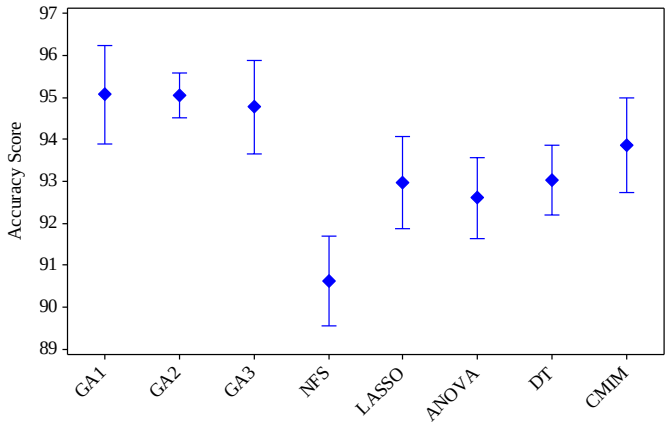

Figure A-19. 95% CI of accuracy score for Sonar dataset with LGR as classifier

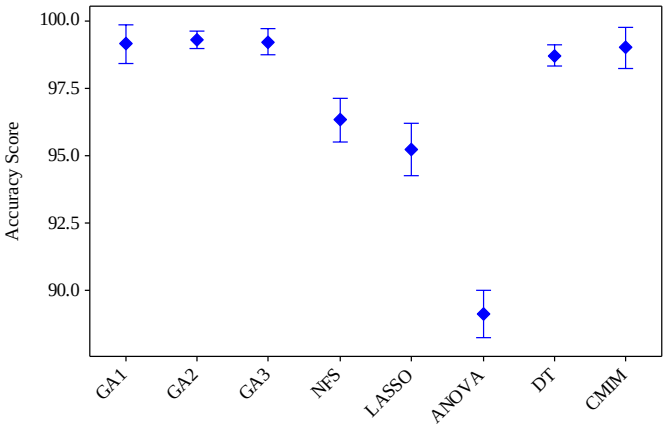

Figure A-20. 95% CI of accuracy score for Sonar dataset with RF as classifier

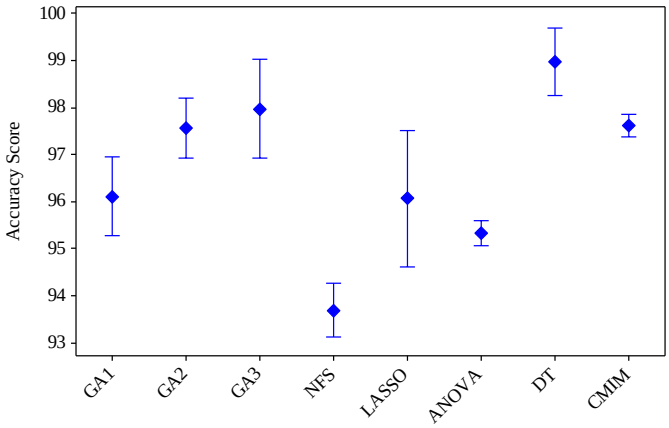

Figure A-21. 95% CI of accuracy score for Sonar dataset with SVM as classifier

Appendix B

Chapter 4: 95% CI of MSE Figures for Regression Datasets

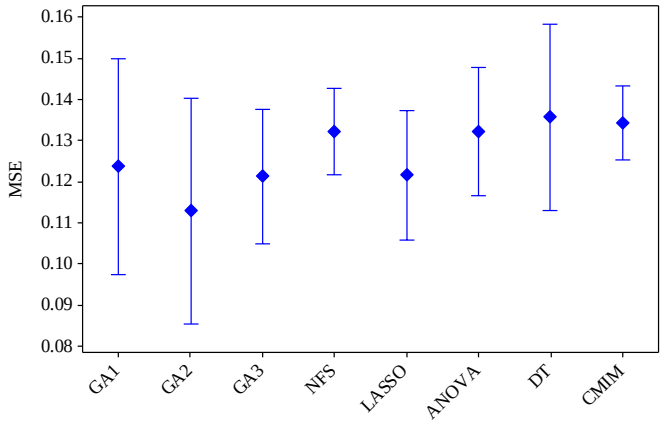

Figure B-1. 95% CI of MSE for Auto MPG dataset with ADB as regressor

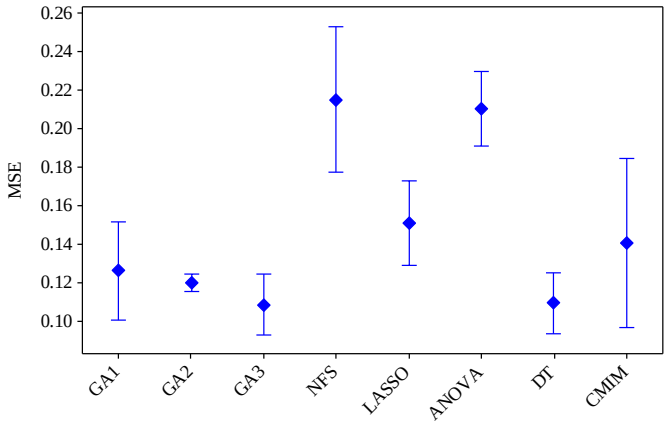

Figure B-2. 95% CI of MSE for Auto MPG dataset with LNR as regressor

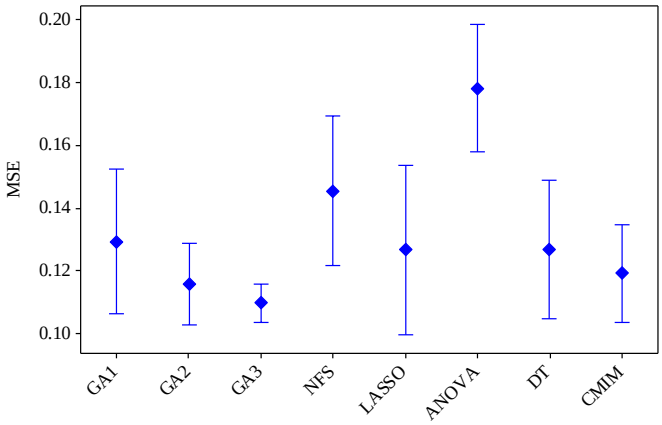

Figure B-3. 95% CI of MSE for Auto MPG dataset with SVR as regressor

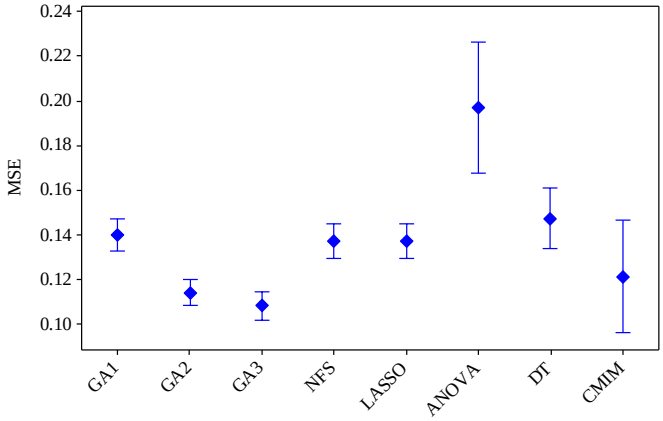

Figure B-4. 95% CI of MSE for Gas Sensor dataset with ADB as regressor

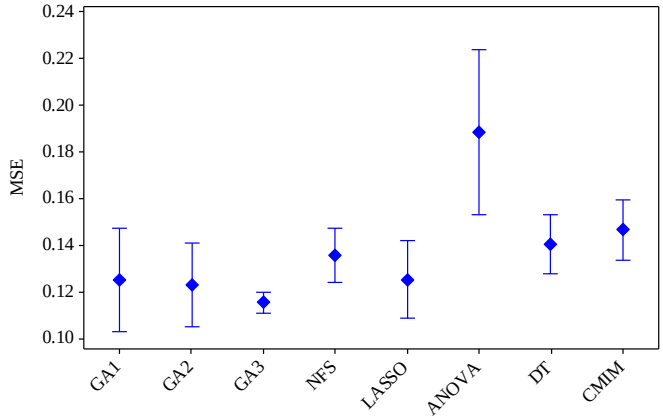

Figure B-5. 95% CI of MSE for Gas Sensor dataset with LNR as regressor

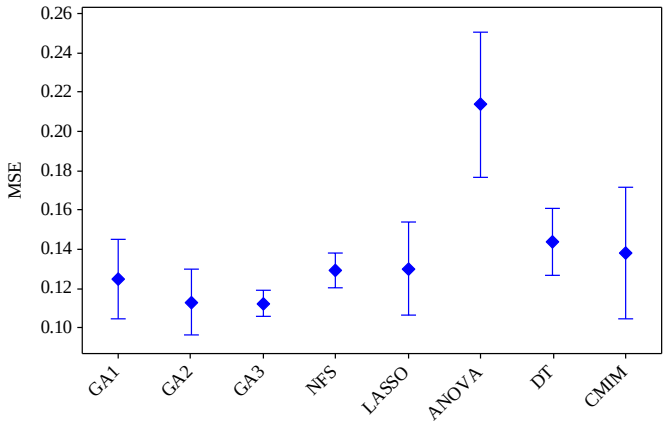

Figure B-6. 95% CI of MSE for Gas Sensor dataset with SVR as regressor

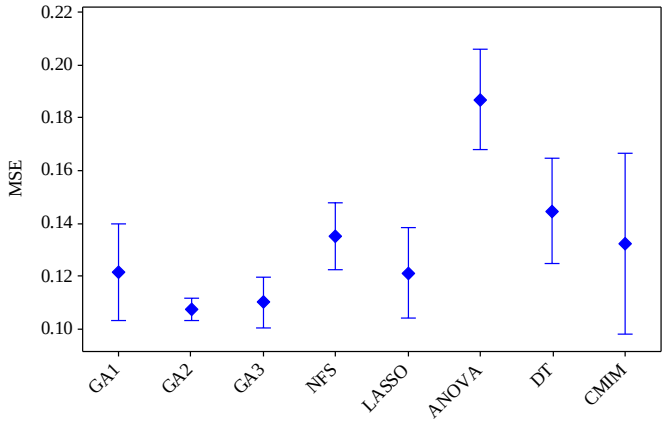

Figure B-7. 95% CI of MSE for Stock portfolio dataset with ADB as regressor

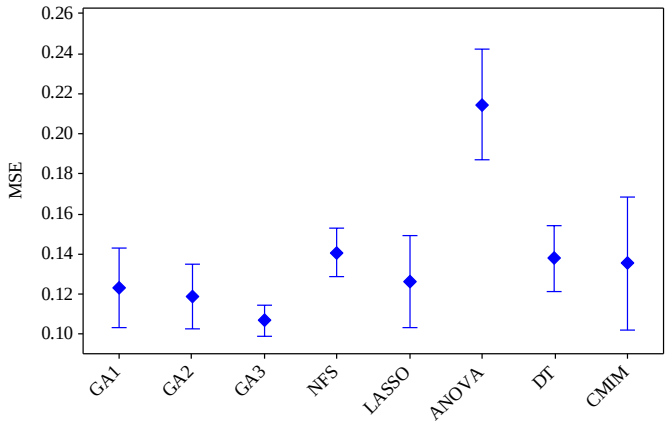

Figure B-8. 95% CI of MSE for Stock portfolio dataset with LNR as regressor

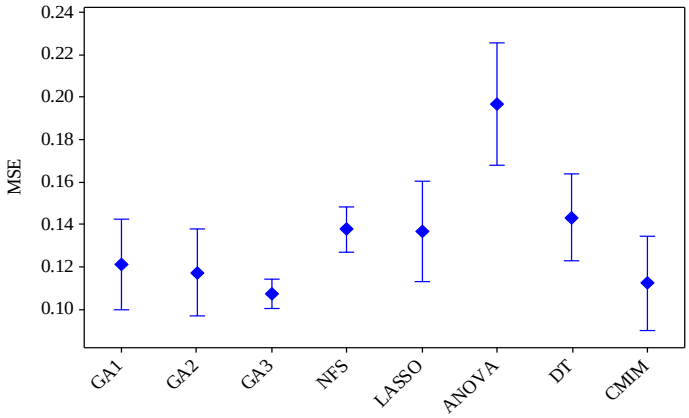

Figure B-9. 95% CI of MSE for Stock portfolio dataset with SVR as regressor

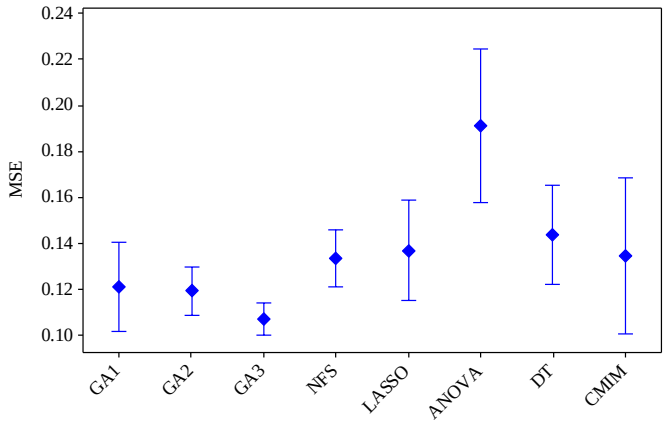

Figure B-10. 95% CI of MSE for Red Wine dataset with ADB as regressor

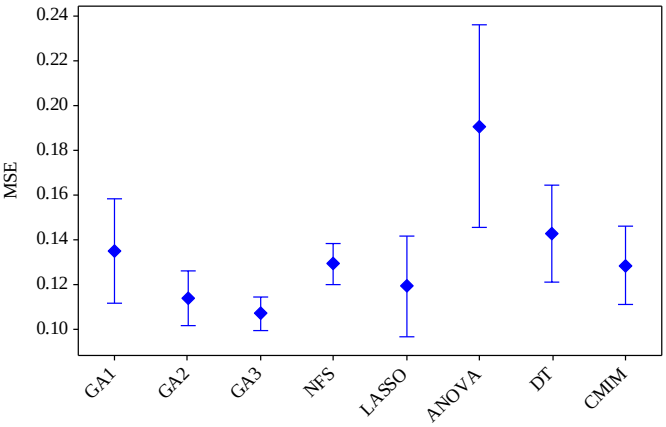

Figure B-11. 95% CI of MSE for Red Wine dataset with LNR as regressor

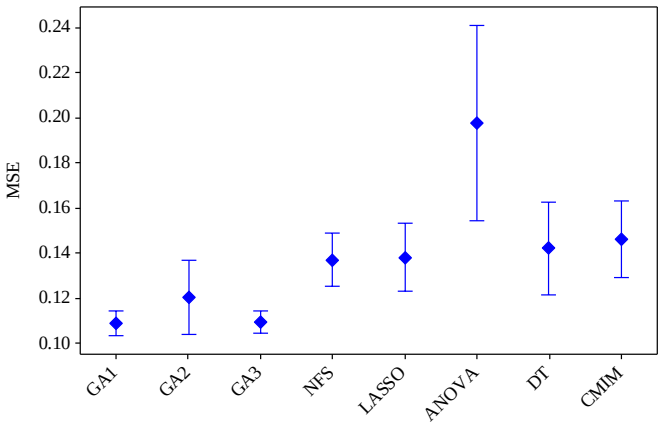

Figure B-12. 95% CI of MSE for Red Wine dataset with SVR as regressor

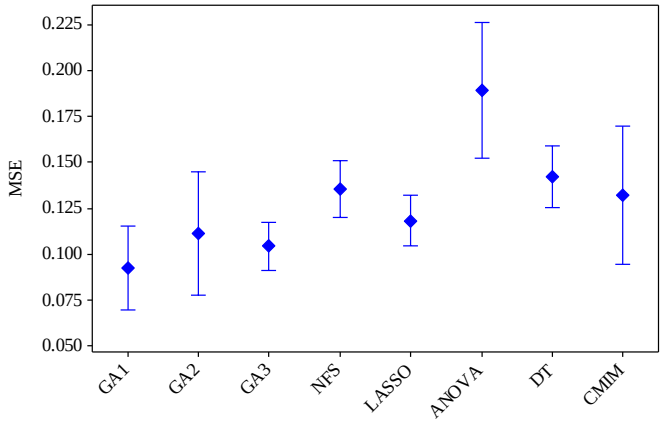

Figure B-13. 95% CI of MSE for Life expectancy dataset with ADB as regressor

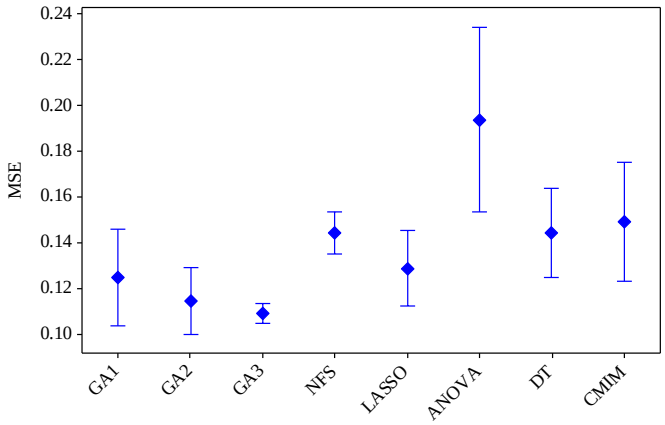

Figure B-14. 95% CI of MSE for Life expectancy dataset with LNR as regressor

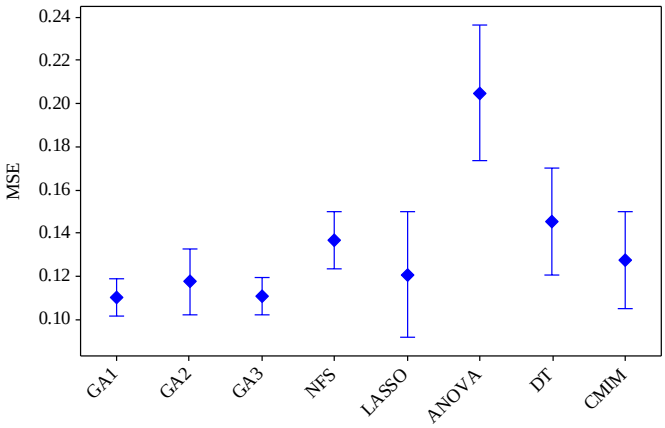

Figure B-15. 95% CI of MSE for Life expectancy dataset with SVR as regressor

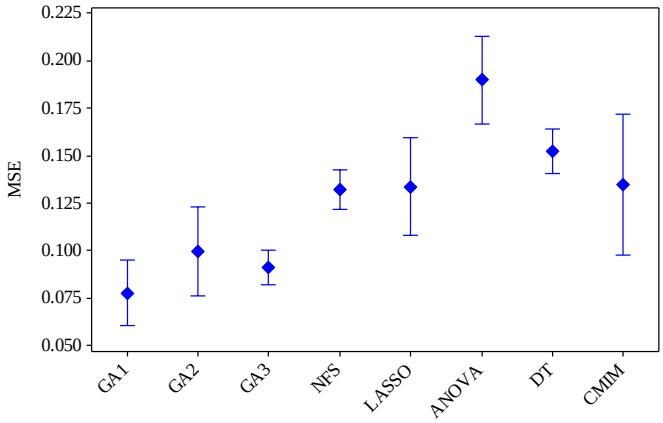

Figure B-16. 95% CI of MSE for Blog Feedback dataset with ADB as regressor

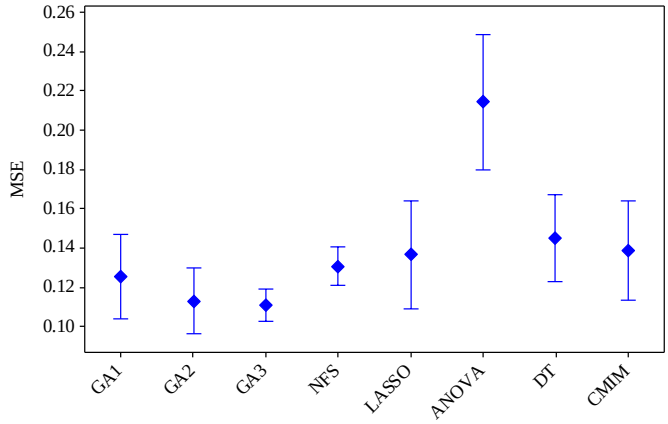

Figure B-17. 95% CI of MSE for Blog Feedback dataset with LNR as regressor

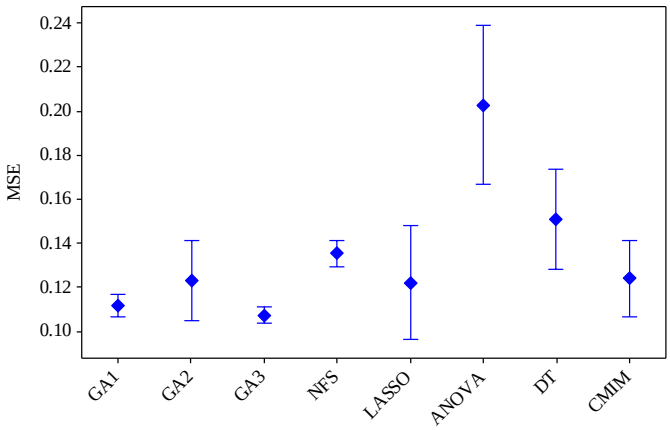

Figure B-18. 95% CI of MSE for Blog Feedback dataset with SVR as regressor

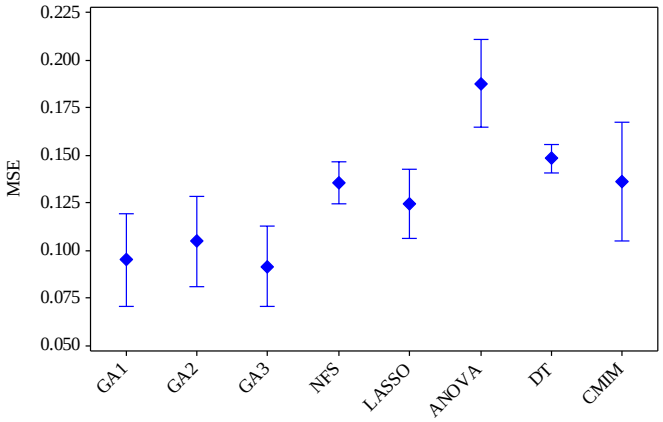

Figure B-19. 95% CI of MSE for Bodily expression dataset with ADB as regressor

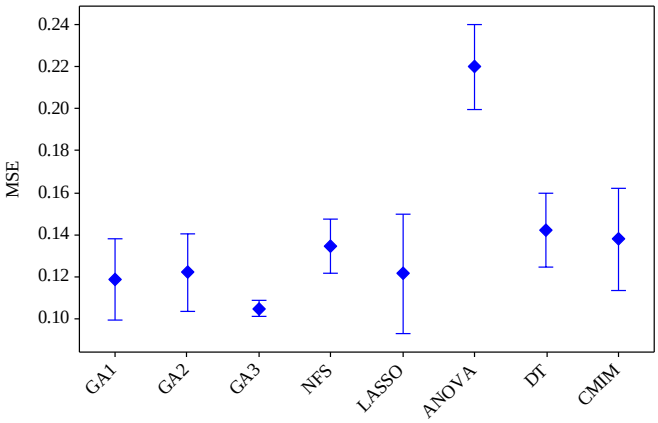

Figure B-20. 95% CI of MSE for Bodily expression dataset with LNR as regressor

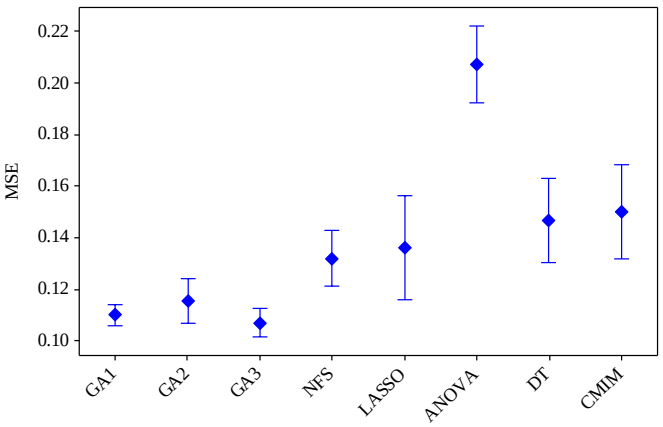

Figure B-21. 95% CI of MSE for Bodily expression dataset with SVR as regressor