

ISSN (Print): 2077-7973  
ISSN (Online): 2077-8767  
DOI: 10.6977/IJoSI.202303\_7(5)

# International Journal of Systematic Innovation



VOL. 07 NO.05  
March, 2023

Published by the Society of Systematic Innovation

---

***Opportunity Identification  
&  
Problem Solving***

# The International Journal of Systematic Innovation

---

**Publisher:**

The Society of Systematic Innovation

**Editorial Team:**Editor-in-Chief:

Sheu, Dongliang Daniel (National Tsing Hua University, Taiwan)

Executive Editor:

Deng, Jyhjeng (DaYeh University, Taiwan)

Editorial Team Members (in alphabetical order):

- Cavallucci, Denis (INSA Strasbourg University, France)
- Feygenon, Oleg (Algorithm Technology Research Center, Russian Federation)
- Filmore, Paul (University of Plymouth, UK)
- Kusiak, Andrew (University of Iowa, USA)
- Lee, Jay (University of Cincinnati, USA)
- Litvin, Simon (GEN TRIZ, USA)
- Lu, Stephen (University of Southern California, USA)

- Mann, Darrell (Ideal Final Result, Inc., UK)
- De Guio, Roland (INSA Strasbourg University, France)
- Sawaguch, Manabu (Waseda University, Japan)
- Shouchkov, Valeri (ICG Training& Consulting, Netherlands)
- Song, Yong-Won (Korea Polytechnic University, Korea)
- Yoo, Seung-Hyun (Ajou University, Korea)
- Yu, Oliver (San Jose State University, USA)
- Zhang, Zhinan (Shanghai Jiao Tong University)

Assistant Editor:

- Maggie Sheu
- Fish Shih

**Editorial Office:**

- The International Journal of Systematic Innovation
- 6F, # 352, Sec. 2, Guanfu Rd, Hsinchu, Taiwan, R.O.C., 30071
- e-mail: [editor@systematic-innovation.org](mailto:editor@systematic-innovation.org)
- web site: <http://www.IJoSI.org>

## INTERNATIONAL JOURNAL OF SYSTEMATIC INNOVATION

## CONTENTS

March 2023 VOLUME 7 ISSUE 5

## FULL PAPERS

- Structural Simulation of Devices Based on Patent Description.....  
... Alexandr Bushuev, Sergey Chepinskiy, Weijie Lin, Botao Zhang, Jian Wang 1-9
- New Model for Creating Innovative Solutions in Continuous Improvement  
Environments.....Vasco F. C. V. Soares, Helena V. G. Navas 10-29
- Feature Selection Using Binary Particle Swarm Optimization Algorithm to Predict  
Repurchase Intention from Customer Reviews .....  
.....Dimas Adrianto, Dedy Suryadi 30-45
- Robustified Principal Component Analysis for Feature Selection in EEG Signal  
Classification..... R. John Martin 46-55
- A Novel Underwater Packet Scheduling based on Modified Priority Backpressure  
and Peak Age of Information approach .....  
..... A Caroline Mary, A V Senthil Kumar, H R Chennamma 56-63
- Application of Text Mining in PTT Forum in Analysis of Consumer Preference for  
Online Shopping Platforms.....Wen-ni Shih, Yu-sen Lin 64-79

# Structural simulation of devices based on patent descriptions

Alexandr Bushuev<sup>1\*</sup>, Sergey Chepinskiy<sup>1</sup>, Weijie Lin<sup>2</sup>, Botao Zhang<sup>2</sup>, Jian Wang<sup>1,3</sup>

<sup>1</sup>Faculty of Control Systems and Robotics, ITMO University, Saint Petersburg, Russia

<sup>2</sup>School of Automation, Hangzhou Dianzi University, Hangzhou, China

<sup>3</sup>HDU-ITMO Joint Institute, Hangzhou Dianzi University, Hangzhou, China

\* Corresponding author E-mail: bushuev@inbox.ru

(Received 8 November 2022; Final version received 3 March 2023; Accepted 8 March 2023)

## Abstract

The problem of the interaction of patent law and TRIZ is considered in this paper. A su-field analysis is applied from TRIZ tools. The European claim of the invention is used to describe the device. The claim is divided into technical features for structural simulation. Binary relationships are introduced between features to build hierarchy levels. As a result, an oriented graph is obtained, the nodes of which are features, and the branches show the subordination of features. The elements are at the top level, and the connections between them are at the middle level. The properties of elements and links are at the lower level. The physical operating principle of the device from the patent description is used for the numerical evaluation of the structure. At the middle level of the hierarchy, features of interaction between elements are replaced by mechanical, thermal, electric, and magnetic fields from the Su-field analysis. Further, the inventive fields are replaced by the dimensions of the physical quantities in which the fields are measured. For example, the dimensions of ampere, volt, watt, and coulomb are used for electric fields, depending on the specific design of the device. The dimensions of physical quantities are given in the Bartini basis with two basic units time T and space L. Two-dimensional diagonal matrices are introduced to describe the weights of graph branches. A weighted and oriented graph mathematically describes the structure of the claim. The purpose of the simulation is to calculate the resource intensity of the claim structure, as well as a numerical comparison of the novelty of the claims with the prototype. The novelty coefficient is determined by the degree of asymmetry of the new solution and the prototype. The symmetric part is the inventive features included in the restrictive part of the claims. The inventive features included in the characteristic part form the asymmetry of the claims.

*Keywords: claims in the invention, novelty, simulation, Su-field analysis*

## 1. Introduction

An important problem of technical design is the evaluation of inventive solutions. To compare inventive solutions, they must be formalized according to certain rules. One of the well-known methods of comparing formalized solutions is patent examination of inventions. The most structured part of the patent is the claim. Witz and Geisel (2017) indicate the claim has three parts: preamble, transitional phrase, and claim body. The claim body contains features, i.e. components, their connections, and characteristics. In Russia and China, the European form of the claims is used, which has a restrictive and distinctive part. The restrictive part contains the features of the invention common to the prototype and the new solution. The distinctive part contains only the features of a new solution. The features of the invention are used for the mathematical patent model, according to which

different indicators of the technical solution are evaluated.

In the work (Bushuev and Chepinskiy, 2007a), a probabilistic mathematical model is proposed, according to which the level of development of the technical system is estimated. A chronological sequence of inventions  $x_k$  is introduced, in which the invention  $x_{k-1}$  is the prototype for the invention  $x_k$ ,  $k=0, 1, 2, \dots$ . Then each invention can be considered as a state of a single-scale queuing system with waiting, which receives a random stream  $S_i$  of invention features. The features of the restrictive part come to the kernel for maintenance  $Ker\ x_k$ , and the features of the distinctive part form a queue  $Que\ x_k$ . The probability  $p_{ik}(Ker\ x_k) \mid S_i \in Ker\ x_k$  and the probability  $p_{jk}(Que\ x_k) \mid S_j \in Que\ x_k, i \neq j$  are entered. It is shown for  $k \rightarrow \infty$   $p_{jk}(Que\ x_k) \leq 0.5$  and  $\lim p_{ik}(Ker\ x_k) = 1$ . The product of the probabilities of several features included in the kernel gives the kernel density  $p(k)$ . The product of the

probabilities of several features included in the core gives the core density  $p(k)$ . The graph  $p(k)$  gives a discrete S-curve (Altshuller, 1999), according to which the level of technology development is estimated.

Weidong et al. (2020) propose a graph-based probabilistic patent evaluation model. In the model, the textual parts are combined with some structured parts of patents. The patent value is initially determined by the internal features of the patent. Given a patent  $o$ , the patent value  $v_o$  is initially formed by the features from the patent and exhibits a prior probability distribution.  $v_o \sim p(v_o | D_o)$  where  $o \in VO$  and  $D_o$  denotes some features extracted from the structured and unstructured parts of  $o$ .  $VO = \{o_n\}$  where  $o$  denotes an object to be valued. Next patent values are changed by the values of the nodes that are associated with the patents.

In the work (Bushuev and Chepinskiy, 2007b), a structural model of the claims in the form of a graph is proposed. The nodes of the graph are the features of the device, and the branches are binary relations between the features. The structural scheme determines the strength of the claims  $F = 1/n$ , where  $n$  is the number of nodes of the graph. In the refined estimation of the strength of the formula, node weight functions are used, depending on the number of branches of the node.

The considered methods of stimulation have a disadvantage. The features of the claims are considered equivalent since the claims represent the device in a static stationary state. The importance of the features is found in the dynamic state. The physical operating principle of the device is presented in the patent description, but not in the claims. The dynamic action of the device represents an unstructured part of the patent description. In TRIZ (Goldovsky and Weinerman, 1990), a simulation of the structure by an oriented graph consisting of substances  $S_i$  and fields  $F_j$  is proposed. Mechanical, thermal, electric, and magnetic fields are used. In the graph  $F_1 \rightarrow S_1 \rightarrow F_2 \rightarrow S_2 \rightarrow \dots \rightarrow F_i \rightarrow S_j$ , the direction of energy conversion is shown by arrows. The nodes of the graph are not equivalent, but their numerical weight is missing. Differential equations are used to simulate a dynamic graph. In (Zaripova et al, 2015), for any node of the graph, the differential equation of thermodynamics  $dQ = PdE$  is used, where  $dQ$  is the differential of the generalized work,  $P$  is the generalized force, and  $dE$  is the generalized coordinate. The graph model turns out to be dynamic, but it is redundant for patent protection of the device.

In TRIZ, the use of the theory of dimensions of physical quantities is also known. The inputs and outputs of the nodes of the graph are encoded by the dimensions of physical quantities on one or another basis. For example, in (Coatanéab et al, 2015) the MLT-basis (mass-length-time) is considered. The topology of a technical system is represented by a graph, the nodes of which are variables of three types. The first type includes variables at the input of the device. The second type includes variables at the output of the device. The third type includes intermediate variables that form chains of transformations from inputs to outputs. The branches of the graph are constructed according to the expert assessment of the cause-and-effect relationships. The graph matrix in the MLT-basis is used to check the reliability of the structural model and find nodes with a violation of cause-and-effect relationships. In (Bushuev and Kudriavtseva, 2019), the LT-basis (length-time) is used to numerically estimate the resource intensity of the graph. Shibayama et al (2021) propose a numerical evaluation of scientific documents on semantic text analysis. The novelty of the document is assessed by the frequency of references in the cited literature. However, such an estimate has weak validity for patents, since a patent can have only one reference to a prototype.

The work aims to obtain numerical estimates for comparing a new solution and a prototype according to their patents. Let's pose the following search problem.

## 2. Problem statement

The new technical solution is specified in the patent description with the claim. Therefore, the description and claim of the prototype invention are known. We will assume that the claim has features from three levels of hierarchy. The highest level of the hierarchy includes features of the presence of structural elements. The middle level of the hierarchy includes features of a connection between elements and their mutual arrangement. The lower level of the hierarchy includes features that define the shape of the element and the form of the relationship between the elements, as well as features that define the parameters and other characteristics of the element.

The set of features is denoted by  $\{D_i\}$ , where  $i$  is the number of the feature,  $i = 1, 2, \dots$ . We introduce the binary relation  $D_i R D_j$ , where  $R$  means that the attribute  $D_i$  does not exist without the attribute  $D_j$ ,  $i \neq j$ . The binary relation establishes the subordination of



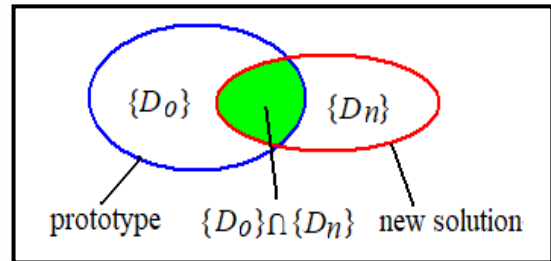
features, which is indicated by arrows in the graph, and the nodes in the graph are the features of the device.

Let's explain the concept of hierarchy levels in the following example. The claim is written as "the device has block A mounted on block B". Here we have three features: block A, block B, and mounted on. Let's exclude the mounted on feature. Then we will get a claim like "the device has block A and block B". This is a logically valid expression that can be used in the restrictive part of the claim. The indication of the relative position of blocks A and B can be indicated in the distinctive part of the claim for the operability of the device. If we exclude the features block A and block B, then we get the claim "the device has a mounted on. This is a logically incorrect expression. In patent rules, it is customary to exclude all connections of an element with other elements if this element is excluded. Mathematically, these rules are confirmed by graph theory. There are disconnected graphs in which some nodes do not have branches. The simplest structure model can consist only of nodes. However, a model consisting only of branches or links does not exist. Thus, elements, blocks, and nodes form the highest level of the hierarchy of the model, and the connections between them form the middle level.

The lowest level of the hierarchy is formed by features characterizing the internal properties of elements, and blocks, as well as the connections between them. For example, the device has a round block A mounted rigidly on block B. The features "round" and "rigidly" are at the lowest level of the hierarchy. They disappear if features of a higher level are excluded from the claim. In graph theory, the weight of nodes and branches is a feature of the lowest level of the hierarchy. Such a graph is called weighted or colored when each node is assigned its color.

Some of the  $D_i$  features are common to the new solution and the prototype. Common features are included in the restrictive part of the claims for a new solution. We denote the set of features of the prototype by  $\{D_o\}$ , and we denote the set of features of the new solution by  $\{D_n\}$  where  $o$  and  $n$  are integers denoting the feature number. Then the restrictive part of the claims of the new solution will be equal to the intersection  $\{D_o\} \cap \{D_n\}$  of the sets. We will consider the intersection as the symmetric part of the new solution-prototype pair. The set of features of the new solution-prototype pair is shown in Fig. 1 where the symmetrical part is indicated in green. The distinctive part of the new solution forms the set  $\{D_n\} - \{D_o\} \cap \{D_n\}$ . If  $\{D_o\} = \{D_n\}$ , then  $\{D_o\} \cap \{D_n\} =$

$\{D_o\} = \{D_n\}$ , and such an invention has no novelty  $\{D_n\} - \{D_o\} \cap \{D_n\} = \{\emptyset\}$ . In this case, the new solution and the prototype are completely symmetrical. Therefore, the inventor needs a minimum of information to get a new solution from a known prototype.



**Fig. 1. The set of features of the new solution-prototype pair**

This information is that the new solution and the prototype are completely symmetrical. A new solution turns out to be asymmetric when the inventor has to generate more new information so that this solution can be reproduced. Thus, we will evaluate the degree of novelty by the magnitude of the asymmetry. The paper (MacCormac, 1998) discusses in more detail the use of symmetry and asymmetry in science and technology.

### 3. Defining the similarity function

Brown (2021) considers the similarity function  $S(A, B)$  of two objects  $A$  and  $B$  as a function  $f$  of three arguments

$$S(A, B) = f(A \cap B, A \setminus B, B \setminus A), \quad (1)$$

where  $A \cap B$  are features belonging to  $A$  and  $B$ ,  $A \setminus B$  are features belonging to  $A$  but not belonging to  $B$ ,  $B \setminus A$  are features belonging to  $B$  but not belonging to  $A$ . For patent features, you can write  $A \cap B = \{D_o\} \cap \{D_n\}$ ,  $A \setminus B = \{D_o\} - \{D_n\}$ ,  $B \setminus A = \{D_n\} - \{D_o\}$  where object  $A$  is a prototype, in object  $B$  is a new solution. Eq. (1) for the degree of symmetry  $E$  is written as

$$E = f(\{D_o\} \cap \{D_n\}) / [f(\{D_o\} \cap \{D_n\}) + \alpha f(\{D_o\} - \{D_n\}) + \beta f(\{D_n\} - \{D_o\})], \quad (2)$$

where  $\alpha$  and  $\beta$  are some feature weights. It is necessary to take two steps to find the function  $f$  and the coefficients  $\alpha$  and  $\beta$ . The first step is called structural simulation, and the second step is called dimensional simulation.

### 3.1 Structural simulation

We will show a structural simulation of some Di features using a simple example. Let the claim of an optical device (Fig. 2) be given: a radiation source on the optical axis of which a photodetector is installed. Let's make a block diagram of the claim. The scheme has two upper-level features D1 - the radiation source and D2 - the photodetector.

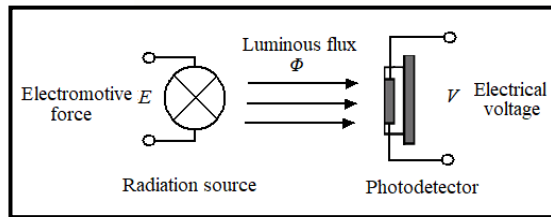


Fig. 2. The design of the optical device

Installation on the optical axis D3 is a feature of a middle level, since D3 R D1 and D3 R D2. A structural simulation of the claim is shown in Fig. 3, where the graph of features and its nodes are shown on the left, and an equivalent Su-field structure is shown on the right. The nodes of the Di graph are placed in circles; the directions of subordination are indicated by arrows. To obtain an equivalent Su-field structure, the description of the work and the design of the device in Fig. 2 are used. Top-level features D1 and D2 are replaced with full Su-fields consisting of three elements  $F \rightarrow S \rightarrow F$ . Such a structure has problems to detect (Petrov, 2014). The feature of the middle level D3 is replaced by the field F2, since the radiation field passes along the optical axis between the radiation source S1 and the photodetector S2. Such a Su-field is called an incomplete Su-field.

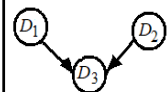
	The structure of the features of the claims	Equivalent Su-field structure
Graph nodes	$D_1$	$F_1 \rightarrow S_1 \rightarrow F_2$
	$D_2$	$F_2 \rightarrow S_2 \rightarrow F_3$
	$D_3$	$F_2$
Graph		$F_1 \rightarrow S_1 \rightarrow F_2 \rightarrow S_2 \rightarrow F_3$

Fig. 3. Structural simulation of an optical device, F1 - electric field, S1 - radiation source, F2 - radiation field or electromagnetic field, S2 - photodetector, F3 - electric field.

The final Su-field structure represents the complex Su-Field  $F_1 \rightarrow S_1 \rightarrow F_2 \rightarrow S_2 \rightarrow F_3$ . If there is no D3 feature in the claim, then the Su-field structure contains two completely unrelated Su-Fields  $F_1 \rightarrow S_1 \rightarrow F_2$  and  $F_2 \rightarrow S_2 \rightarrow F_3$ . If there is no D2 feature in the claim, then the D3 feature disappears, and only the D1 Su-Field remains  $F_1 \rightarrow S_1 \rightarrow F_2$ . As you can see, the F2 field does not disappear. It defines the operational function of the radiation source.

Let's assume that the new solution has one distinctive feature D4. This feature means that the radiation source is monochromatic with a wavelength of 700 nm. The structural model of the new solution is shown in Fig. 4.

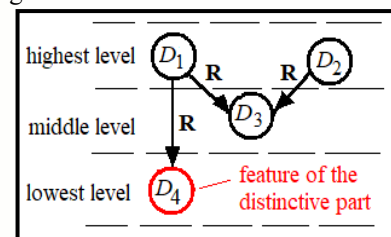


Fig. 4. Hierarchical structural simulation of a new solution, D4 is a feature of the distinctive part

The D4 feature is at the lowest level of the hierarchy since  $D_4 R D_1$ . The feature of the lower level of the hierarchy is introduced intentionally to show the limitations of Su-field analysis. When Su-fields are constructed, the internal properties of substances are not taken into account. Therefore, the equivalent Su-field structure of the new solution  $F_1 \rightarrow S_1 \rightarrow F_2 \rightarrow S_2 \rightarrow F_3$  coincides with the structure of the prototype. The wavelength characterizes the radiation field F2 which replaces the feature D3 of optical communication between D1 and D2. However, the claim describes the design of the device in static, i.e. in the off state. Fields do not exist in static so the D4 feature is subordinate to the D1 feature.

### 3.2 Dimensional simulation

The transition from the structural model of the claims to an equivalent Su-field structure also involves the use of a description of the operation of the device. The physical operating principle of the device allows you to determine the types of fields in Su-field analysis. However numerical estimates of the features are needed to compare the prototype and the new solution. Su-Field analysis does not allow making such a comparison. Indeed, it is impossible to answer how much an electric field is better than a mechanical

pressure field or a magnetic field is better than a thermal field. The paper (Bushuev and Kudriavtseva, 2019) shows how inventive fields of different types of energy can be numerically compared. In invention problems, the fields differ from each other in the physical quantities by which they are measured. For example, the electric field can be measured in units of electric voltage, current, charge, and field strength. Consider the radiation source D1 in Figure 2 with the Su-Field F1→S1→F2. The input electric field F1 is measured in EMF units, i.e. in volts. The output radiation field F2 is measured in units of luminous flux, i.e. lumens. In the Bartini system of kinematic quantities (Bartini, 2005), the EMF has a volt dimension [L2T-2], and the lumen has the dimension joule/c [L5T-5], where length L and time T are the basic units. We introduce matrices for the input and output values

$$EMF = \begin{bmatrix} L^2 & 0 \\ 0 & T^{-2} \end{bmatrix}, \Phi = \begin{bmatrix} L^5 & 0 \\ 0 & T^{-5} \end{bmatrix}, \quad (3)$$

where EMF is the electromotive force at the input of the radiation source,  $\Phi$  is the luminous flux at the output. We find the transfer matrix W1 of the radiation source from Eq. (3)

$$W_1 = \Phi EMF^{-1} = \begin{bmatrix} L^5 & 0 \\ 0 & T^{-5} \end{bmatrix} \begin{bmatrix} L^2 & 0 \\ 0 & T^{-2} \end{bmatrix}^{-1} = \begin{bmatrix} L^3 & 0 \\ 0 & T^{-3} \end{bmatrix} \quad (4)$$

The input value of the photodetector will be the illumination E measured in lux, and the output value is the electrical voltage V in volts. The illumination E in the Bartini system has the dimension of surface power [L3T-5]. Therefore, the transfer matrix W2 of the photodetector is equal to

$$W_2 = VE^{-1} = \begin{bmatrix} L^2 & 0 \\ 0 & T^{-2} \end{bmatrix} \begin{bmatrix} L^3 & 0 \\ 0 & T^{-5} \end{bmatrix}^{-1} = \begin{bmatrix} L^{-1} & 0 \\ 0 & T^3 \end{bmatrix} \quad (5)$$

The transfer matrix W3 of the feature D3 is equal to

$$W_3 = E\Phi^{-1} = \begin{bmatrix} L^3 & 0 \\ 0 & T^{-5} \end{bmatrix} \begin{bmatrix} L^5 & 0 \\ 0 & T^{-5} \end{bmatrix}^{-1} = \begin{bmatrix} L^{-2} & 0 \\ 0 & T^0 \end{bmatrix} \quad (6)$$

In Eq. (6), it is assumed that the feature D3 cuts out a part of the spherical luminous flux  $\Phi$ , limited by the aperture S of the photodetector. The transfer matrix W3 is the inverse matrix of the surface S. In the LT-basis, the surface has dimension [L2T0] or m2. The dimensional simulation scheme of the prototype is shown in Fig. 5. The features of the claims Di are given by the transfer matrices Wi. The transfer matrices differ

from each other in exponent m and n with basic units Lm and Tn where m and n are integers.

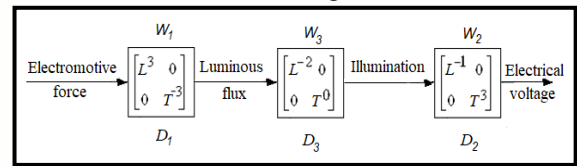


Fig. 5. Graph of the dimensional simulation of the prototype

Let's determine the resource intensity RI of the transfer matrices forming the nodes of the graph for dimensional simulation

$$RI_i = \sqrt{m^2 + n^2} \text{ for } W_i = \begin{bmatrix} L^m & 0 \\ 0 & T^n \end{bmatrix},$$

$$RI_1 = \sqrt{(3)^2 + (-3)^2} = 4.24,$$

$$RI_2 = \sqrt{(-1)^2 + (3)^2} = 3.16,$$

$$RI_3 = \sqrt{(-2)^2 + (0)^2} = 2.0 \quad (7)$$

The total intensity of the prototype is equal to

$$RI_o = \sum_{i=1}^3 RI_i = 9.40. \quad (8)$$

From a physical point of view, the resource intensity of the node Wi means the time and space resources spent on converting the input value into the output value. If a block (or node) spends little resources on conversion, the more it is ideal. The ideal end result (IER) is obtained without spending resources. Therefore, the resource intensity RI is zero for an ideal end result. Space and time are the most important resources for resolving contradictions in the algorithm for solving inventive problems. They define the operational zone and the operational time of the conflict. All other resources, such as weight, speed, pressure, electrical voltage, temperature, and others, can be obtained from the resources of time and space using dimensional theory. The technology of obtaining dimensions for resources is considered in the work (Bushuev A. 2017). The work is based on the system of kinematic quantities by Bartini (2005).

Let's imagine the Su-field F1→S1→F2 of the radiation source as a physical effect to simulate the distinctive feature D4. This physical effect converts the electromotive force EMF at the input into a luminous flux  $\Phi$  and the wavelength  $\lambda$  at the output. Litvinov et al (2022) consider the stimulation of multidimensional physical effects. In our example, a physical effect with two outputs has two transfer matrices for dimensional simulation. The first matrix W1 is already defined for the prototype in Eq. (4). Similarly, we define the second transfer matrix W41 for feature D4:



$$W_{41} = \lambda EMF^{-1} = \begin{bmatrix} L^1 & 0 \\ 0 & T^0 \end{bmatrix} \begin{bmatrix} L^2 & 0 \\ 0 & T^{-2} \end{bmatrix}^{-1} = \begin{bmatrix} L^{-1} & 0 \\ 0 & T^2 \end{bmatrix} \quad (9)$$

where  $\lambda$  is the dimensional matrix of the wavelength. Note that the specific value of the wavelength  $\lambda = 800 \text{ nm}$  is not simulated. The feature of a monochromatic light source is simulated here only. Taking into account the numerical values of the parameters of the blocks and inputs and outputs is considered in the work (Bushuev et al, 2021).

Thus, the radiation source model contains two matrices  $W1$  and  $W41$ . The matrix  $W1$  stands in the energy channel  $\Phi$ , and the matrix  $W41$  stands in the information channel  $\lambda$ . The model of the radiation source is outlined with a blue dashed line in Fig. 6. Since the radiation source is monochromatic, the photodetector must receive radiation of the same wavelength. Therefore, the sensing element of the photodetector must have a maximum spectral

characteristic at the transmitted wavelength. A physical effect with two inputs has this characteristic. The surface power or illumination  $E$  of the photocrystal is the first input. The second input is the wavelength  $\lambda$  of the radiation. The electrical voltage  $V$  is the only output of the physical effect. The dimensional model of such a physical effect is outlined with a green dashed line in Fig. 6. Let's define the transfer matrix  $W42$  from the wavelength  $\lambda$  to the electrical voltage  $V$

$$W_{42} = V\lambda^{-1} = \begin{bmatrix} L^2 & 0 \\ 0 & T^{-2} \end{bmatrix} \begin{bmatrix} L^1 & 0 \\ 0 & T^0 \end{bmatrix}^{-1} = \begin{bmatrix} L^1 & 0 \\ 0 & T^{-2} \end{bmatrix} \quad (10)$$

The matrix  $W43$  is a unit matrix that does not change the dimension of the input quantities. The matrix is designed to produce a single photodetector output  $V$ . The matrices  $W41$  and  $W42$  simulate the distinctive feature  $D4$ , the structure of which is outlined with a red dashed line in Fig. 6.

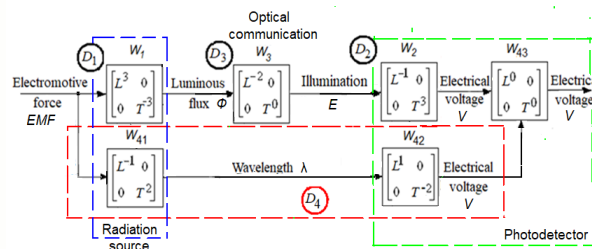


Fig. 6. Dimensional simulation graph for a new solution

Let's determine the resource intensity of the transfer matrices for the distinguishing feature

$$RI_{41} = \sqrt{(-1)^2 + (2)^2} = \sqrt{5}, RI_{42} = \sqrt{(1)^2 + (-2)^2} = \sqrt{5} \quad (12)$$

The total intensity of the distinguishing feature is equal to

$$RI_4 = RI_{41} + RI_{42} = 2\sqrt{5} = 4.47 \quad (13)$$

The total intensity of the new solution is equal to where  $w_i$  is the weighting coefficient of the feature  $i$ ,  $w_i = \alpha = 3$  for the highest level of the hierarchy,  $w_i = \beta = 2$  for the middle level of the hierarchy,  $w_i = \gamma = 1$  for the lowest level of the hierarchy, The distribution of features by hierarchy levels is shown in Fig. 4.

Then the similarity function for the features from the restrictive part of the claims is equal to

$$f(\{D_o\} \cap \{D_n\}) = \alpha RI_1 + \beta RI_2 + \alpha RI_3 = 3\sqrt{18} + 2\sqrt{10} + 3 \cdot 2 = 25.05.$$

Then the similarity function for the features from the distinctive part of the claims is equal to

$$RI_n = RI_o + RI_4 = 9.40 + 2\sqrt{5} = 13.87 \quad (14)$$

The resource intensity from equations (7) and (13) will be used to calculate the similarity function  $f$  in equation

(2). Let the similarity function  $f$  of the set of features  $D_i$  be equal to

$$f(\{D_i\}) = \sum w_i RI_i,$$

$$f(\{D_n\} - \{D_o\}) = \gamma RI_4 = 1 \cdot 4.47 = 4.47.$$

We find the degree of symmetry  $E$  of the prototype-new solution pair from Eq. (2)

$$E = \frac{f(\{D_o\} \cap \{D_n\})}{f(\{D_o\} \cap \{D_n\}) + f(\{D_o\} - \{D_n\}) + f(\{D_n\} - \{D_o\})} = \frac{25.05}{25.05 + 4.47} = 0.849 \quad (15)$$

The set  $\{D_o\} - \{D_n\}$  is empty, since all the features of the prototype are included in the new solution and  $f(\{D_o\} - \{D_n\}) = 0$ . In general, the set  $\{D_o\} - \{D_n\}$  may be nonempty. Next numerically we find the novelty coefficient

$$N=1- E= 1- 0.849= 0.151.$$

The novelty coefficient is in the range  $0 < N < 1$ , the greater the  $N$ , the greater the novelty of the invention. If the novelty coefficient is  $N=0$ , then all the features of the new solution coincide with the features of the prototype. A patent for an invention is not issued. The Ideal End Result (IER) is obtained when  $N=1$ . It is impossible to get the IER, since there is always at least

#### 4. Conclusion

Structural simulation based on the patent description of inventions makes it possible to establish a link between TRIZ and patent law. Substance-field analysis of the claims allows you to build models of the operating physical principle of devices. The costs of time and space resources for design are based on the model. The exponents  $m$  and  $n$  for  $L_m$  and  $T_n$  show the number of integrators and differentiators in time and space in a dimensional simulation of a dynamic model. The complexity of the model depends on the number of integrators and differentiators. The complexity of the model is numerically equal to the resource intensity. The low resource intensity with a large novelty coefficient corresponds to the high quality of the design. Therefore, it is possible to set the task of designing a graph of the physical operating principle of the device with a minimum length and a maximum novelty coefficient. This problem is partially solved by the example of flowmeters (Litvinov et al, 2021). The optimal synthesis of the physical principle of action based on dimensional simulation is considered in the work (Bushuev et al, 2021).

#### Acknowledgments

This research was supported by Zhejiang Provincial Natural Science Foundation of China under Grant No.

#### References

Altshuller, G. S. (1999). *The Innovation Algorithm: TRIZ, systematic innovation, and technical creativity*. Worcester, MA: Technical Innovation Center. ISBN 978-0-9640740-4-0.

Bartini, R. O. (2005). Relations Between Physical Constants, Progress in Physics, October, v.3, 34-40.

Brown, D. C. (2021). A Brief Review of Approaches to Design Novelty Assessment. Computer Science Technical Report

one common feature. The name of the invention is a common feature of the prototype and the new solution. The name defines the function of the device. The value of the patent is modeled by novelty, inventive step, and other value determining parameters (Reitzig, 2003). Therefore, numerical estimation of novelty helps to find the value of the patent.

The considered method can be used to compare different projects of technical systems at the first stage of design. To do this, it is enough to describe the design solution according to the patent rules. The claim includes the most essential features of the physical principle of operation of the device. At the first stage of design, different variants of the operating physical principle of the device are compared.

Simulation of methods of action on material objects is an unsolved problem. The features of the method in the claims are divided into three groups: the presence of actions; sequence of actions; the mode of performing actions, devices, and tools necessary to perform actions. The dynamics of features in claims complicate dimensional modeling. For example, there is such a feature as steel heats up at a rate of 100 degrees Celsius per minute in the claims. The dimension of the heating rate in the Bartini system is  $L_5 T_5$ . Hence, we can find the transfer matrix for this action. However, the numerical value of 100 degrees is not described in dimensional simulation.

LGG19E070008 and was supported by Key Research and Development Project of Zhejiang Province under Grant No. 2019C04018.

Series. Worcester Polytechnic Institute. Massachusetts 01609-2280.

Bushuev, A. B. & Chepinskiy, S. A. (2007a). *Discrete mathematics of inventive problems*. Proceedings of the conference «Simulation Modeling: Theory and Practice» IMMOD-2007. Saint Petersburg. V.1. 93-97. (in Russian)

Bushuev, A. B. & Chepinskiy, S. A. (2007b). *Structural and patent analysis of technical systems*. Proceedings of the conference «Theory and practice of inventive problem solving» TRIZ fest 07. Moscow. 240-246 [http://www.trizland.ru/trizba/pdf-articles/TRiz\\_fest\\_2007\\_referats.pdf](http://www.trizland.ru/trizba/pdf-articles/TRiz_fest_2007_referats.pdf) (in Russian)

- Bushuev, A. B. (2017) *Numerical Estimation of the Energy Information Circuits of Measurement Devices*. Meas Tech 60, 857–862.  
<https://doi.org/10.1007/s11018-017-1283-3>
- Bushuev, A. B. & Kudriavtseva, V. A. (2019). *Simulation of the Block Diagrams of the Information Energy Converters*. Proceedings of the International Conference on Innovative Applied Energy (IAPE'19), UK, Oxford. No. 272. p. 40.
- Bushuev A. B., Boikov V. I., Bystrov S. V., Grigoriev V. V. & Mansurova O. K. (2021). *Synthesis of Optimal Information and Energy Schemes of Measuring and Converting Devices*. Mekhatronika, Avtomatizatsiya, Upravlenie.;22(10):518-526. (In Russ.) <https://doi.org/10.17587/mau.22.518-526>
- Coatanéab E., Rynnänen L., Caloniusb O., Mokammelb F. & Riitahuhtab A. (2015). *Systematic search and ranking of physical contradictions using graph theory principles: Toward a systematic analysis of design strategies and their impacts*. World Conference: TRIZ FUTURE, TF 2011-2014. Procedia Engineering 131. 1165 – 1182.
- Goldovsky, B. I. & Vainerman, M. I. (1990). *Rational creativity*. Publishing “Rechnoy transport”, Moscow. (in Russian)
- Litvinov Y. V., Bushuev A. B., Litvinov E. Y. (2021) *Graphic Synthesis of the Operating Physical Principle of Control and Measuring Devices*. Wave Electronics and its Application in Information and Telecommunication Systems (WECONF 2021). pp. 9470754.  
<https://doi.org/10.1109/WECONF51603.2021.9470754>
- Litvinov Y., Bushuev A. & Nuyya O. (2022) *Simulation of graphs of physical effects for information and energy circuits*. J. Phys.: Conf. Ser. 2373 092003 DOI 10.1088/1742-6596/2373/9/092003
- MacCormac, E. R. (1998). *Symmetry and asymmetry in science and technology*. Duke University Medical Center, Durham, NC. V.4, № 2.
- Petrov, V. (2014). *Theory of inventive problem solving – TRIZ: textbook on the discipline "Algorithms for solving non-standard problems"*. Moscow State University of Economics, Statistics, and Informatics (MESI). (in Russian)
- Reitzig, M. (2003). *What determines patent value?: Insights from the semiconductor industry*. Research Policy. V. 32, Issue 1, Jan, Pages 13-26.  
[https://doi.org/10.1016/S0048-7333\(01\)00193-7](https://doi.org/10.1016/S0048-7333(01)00193-7)
- Shibayama, S., Yin, D. & Matsumoto, K. (2021). *Measuring novelty in science with word embedding*. PLoS ONE 16(7): e0254034.  
<https://doi.org/10.1371/journal.pone.0254034>
- Weidong, L., Xin, L. & Wenbo, Q. (2020). *Probabilistic graph-based valuation model for measuring the relative patent value in a valuation scenario*. Pattern Recognition Letters. V. 138, 204-210.
- Witz, J. & Geisel, K. (2017). *Claim Drafting Workshop. Invention-Con 2017. The Place for Inventors, Markers & Entrepreneurs*. USPTO's Inventors Conference. August 11-12. Alexandria, VA. 1-38.
- Zaripova, V., Petrova, I., Kravets, A. & Evdoshenko, O. (2015). *Knowledge bases of physical effects and phenomena for method of Energy-Informational Models by means of ontologies*. «Creativity in Intelligent Technologies and Data Science». First Conference, CIT&DS. V. IV. Volgograd, Russia: Springer, 224-237.

**AUTHOR BIOGRAPHIES**


**Alexandr B. Bushuev** is an Associate Professor at **ITMO University**, Faculty of control system and robotics in Russia since 1986. A. Bushuev received his Ph.D. degree in Automatic Control in 1980. He studied at the University of Technical Creativity under V. Petrov from 1985-1986, majoring in teacher and developer of TRIZ. A. Bushuev is the author of 10 textbooks for students on TRIZ, and 38 inventions, a member of the World Industrial Property Organization, patent expert.



**Sergey A. Chepinskiy** is an Associate Professor at **ITMO University**, School of Computer Technologies and Control, Faculty of control system and robotics in Russia since 2007. Sergey received his Ph.D. degree in System analysis, Control, and Processing of information in Technical systems from ITMO University. He is also the Founder of RoboEd company. He is currently the Overseas Expert of «111 Center» of Hangzhou Dianzi University. His areas of interest include Systematic Innovation including TRIZ, research activities on motion control of underactuated systems and mobile robots, teaching, and supervision of master students and s.Ph.D. students.



**Weijie Lin** is an Associate Professor at **Hangzhou Dianzi University**, School of Automation, Faculty of electrical engineering in China since 2005. He received his Ph.D. degree in electrical engineering from Zhejiang University in China. His areas of interest include methods of motor driving and technologies of electric vehicle driving systems.



Botao Zhang received a Ph.D. degree in Control Engineering from East China University of Science and Technology, Shanghai, China, in 2012. He is presently an Associate Professor at the School of Automation, Hangzhou Dianzi University, Hangzhou, China. His current research interests include machine vision, intelligent perception, and navigation of mobile robots.



Jian Wang was born in Zhejiang Province, China, in 1980. He received a master degree in computer science in Saint-Petersburg State University of Information Technologies, Mechanics and Optics (ITMO University, Russian) in 2006, where he received a Ph.D. degree in Methods and systems of protection information, and information security in 2011. Now, associate professor at Hangzhou Dianzi University (China) and ITMO University (Russia). His research includes Digital image processing, mode recognition, mobile robot navigation, nonlinear adaptive control, etc.

## New Model for Creating Innovative Solutions in Continuous Improvement Environments

Vasco V. Soares<sup>1</sup>, Helena V. G. Navas<sup>2</sup>

<sup>1</sup> Department of Mechanical and Industrial Engineering, NOVA School of Science and Technology, Universidade NOVA de Lisboa, 2829-516 Caparica, Portugal

<sup>2</sup> UNIDEMI, Department of Mechanical and Industrial Engineering, NOVA School of Science and Technology, Universidade NOVA de Lisboa, 2829-516 Caparica, Portugal

vf.soares@campus.fct.unl.pt; 2 hvgn@fct.unl.pt

(Received 29 July 2022; Final version received 1 January 2023; Accepted 20 February 2023)

### Abstract

Nowadays the competition factor between companies has been essential in their path to innovation. So, firms try to evolve through continuous improvement methodologies that can significantly improve their activities' efficiency as well as gain more and more the trust of customers. The Lean philosophy, the FMEA methodology, and the TRIZ methodology can help companies to achieve these goals.

This article is intended to propose a new model called Continuous Improvement Integrated Model with Innovation and Management (CI-IMIM). It promotes a strong alliance between the areas of continuous improvement and innovation as it is specialized in the creation and prioritization of innovative solutions. The FMEA methodology (Failure Mode and Effects Analysis) focuses on the analysis and prioritization of problems depending on the underlying risk, the TRIZ (Inventive Problem-Solving Theory) contradictions matrix (CM) allows the creation of differentiating solutions from the establishment of technical contradictions, the GUT Matrix (Severity, Urgency and Trend) helps in prioritizing solutions, the Brainstorming focuses on screening solutions through feedback gathered from workers' point of view and the Lean philosophy works as the conducting wire of the entire model. This model is then applied in a real automobile company. Digital Kanban, Visual Management, and 5S stand out here as Lean tools derived from the proposed model. It is also important to mention that persuasion and workers' motivation capacity are crucial.

Conclusions show a significant improvement of 8 out of 10 KPIs. This proves the practical viability of the new model. The mudas reduction of 14,1% and the 2,2% PCE improvement are KPI improvement examples. For future PDCA cycles, regular follow-up meetings about the studied KPIs, a bigger task informatization in the company, and the application of the Lean tools Mizusumashi, Andon e Heijunka are suggested.

*Keywords: FMEA, KPI, Lean, TRIZ*



## 1. Introduction

Nowadays the competitiveness factor between companies/organizations has been decisive in its path to innovation and differentiation. This factor is characterized by the leadership dispute in terms of created value through critical success factors like cost, innovative technology, and product/service customization (Holweg, 2008). The automotive industry stands out as one of the main target sectors of this competitiveness improvement. It is considered the backbone of Gross Domestic Product (GDP) of the countries (Kaitwade, 2020).

In the second half of the twentieth century, the Toyota Production System (TPS) appeared in Japan as a new industrial management philosophy that combines the organizational management Taylorism principles with elements like Just in Time (JIT) and Heijunka. The Lean Manufacturing (LM) concept stands out here (Holweg, 2008).

In twenty first century, there are many successful evidences of the Lean philosophy application, especially in the USA and Germany (Clarke, 2005). However, the effect of the worldwide SARS-CoV-2 outbreak has negatively aggravated the automobile sector since 2020. Environmental implications (air pollution), trade wars (US-China), and tax increases inherent in the sector are also problem sources even though the automobile sector has been adapting better and better to this new reality by using new technologies (Industry 4.0) and continuous improvement methodologies (Clarke, 2005; Kaitwade, 2020).

Besides innovation, the growing use of continuous improvement philosophies with an emphasis on the Lean philosophy is crucial in the company's way to differentiation. Lean production is an essential philosophy in creating value through the removal of waste and improving operational performance. Although Lean Manufacturing is used by weight in a manufacturing environment, Lean Services have been increasingly applied since the beginning of the 21st century. However, a profound change in mentality is required for the implementation of Lean Thinking in services, and there is often some resistance to change in organizations (Andrés-López et al., 2015). Additionally, Six Sigma has proven to be a useful philosophy for quality improvement in any industry. Consequently, productive income improves as well as customer satisfaction (Raisinghani et al., 2005).

Thus, the alliance between continuous improvement philosophies and innovation has enormous potential, whatever the organization might be. This alliance is about to be tested in a real case study in the improvement of a car's dealer internal processes namely in the after-sales department.

## 2. Lean philosophy

The TPS has been the Lean paradigmatic basis since the Toyota Motor Company foundation. The Lean Thinking success has been outrageous as the productivity, reliability, and profitability indexes have been improving (Shang & Low, 2014).

2 models stand out: TPS house and Toyota Way. The TPS house describes Lean as a culture with the following goal: best quality, the lowest cost, the shortest lead time, the best safety, and the highest morale through stable, pull, and standardized processes. The Toyota Way values the people's role in their environment and continuous improvement where people are the most important element in the operational performance improvement even better than the improvement methodologies and techniques suitable to increase production efficiency. In this way, both models are complementary (J. Liker, 2004; J. K. Liker & Morgan, 2006).

*Mudas* (literally translated as waste) are concerned with all types of activities that consume resources and contribute to Lead Time improvement without adding value. This means operational waste. There are 7 wastes according to Ohno: Overproduction, Waiting, Transportation, Motion, Overprocessing, Stock, and Defects. Additionally, some authors have been considering the existence of new waste. This waste either refers to the goods/services that do not satisfy the customers' needs or to the people's sub utilization which means people's skills are disused (J. Liker, 2004; Womack & Jones, 1996). Here Lean Thinking appears not just as the antidote for waste but also as a valuable line of thinking which can be divided into 5 principles (Womack & Jones, 1996): Value, Value Stream, Flow, Pull, and Perfection. As can be seen in Table 1, *mudas* can be characterized from Lean Manufacturing and Lean Services perspectives.

Many Lean tools can be applied in the continuous improvement of a company/organization for instance: PDCA (Plan, Do, Check, Act) cycle, VSM (Value Stream Mapping), Kanban, *Mizusumashi*, etc.

**Table 1.** Characterization of Lean Services *mudas* [adapted from (Andrés-López *et al.*, 2015; Bonaccorsi *et al.*, 2011)]

<i>Muda</i> in the Services sector	LM Analogy	Example	Cause	Corrective Action
Overproduction	Overproduction	Processing items before being required	Poor planning	Levelling
Delay	Waiting	Pending requests	Poor coordination	Flow
Transportation/ Motion	Transportation	Looking for data and information	Poor office housekeeping	Layout change
	Motion			
Duplication	Overprocessing	Repeated unnecessary details	Excessive bureaucracy	Digitalization
Lack of standardization	Stock	Fluctuating lead times	Demand fluctuations	Visual Control
Lack of customer's focus	Defects	Poor attention to the customer	Lack of motivation	Planned breaks
Obsolescence	Defects	Error or incomplete work	Disorder	5S
Miscommunication	Defects	Transparency inexistence	Lack of bonds between workers	Standardization
Under-utilized resources	Under-utilized resources	Limited responsibility	Inefficient management	Use of workers' skills
Failure Demand	Inefficient goods/services	Rejected suggestions	Lack of training	Worker's training
Resistance to change	Inefficient goods/services	Rejected suggestions	Lack of motivation	Awards given to workers

### 3. Innovation

Although there is not a single definition for innovation, this concept is related to creating something new and different.

According to Tohidi & Jabbari (2012), innovation appears as a way of introducing new products/services in the current market as well as new production processes, new supply sources, or even radical changes in a certain industrial structure. It is important to refer that the main goals for innovation are based on new and disruptive technologies which help in the improvement of some process flexibility, quality, and environmental performance. The reduction in energy and raw material consumption is also a goal (Livotov *et al.*, 2019; Tohidi & Jabbari, 2012).

The implementation of these new innovative technologies relies on the capacity of solving a lot of contradiction factors that appear in conflict with each other. The TRIZ methodology (Theory of Inventing Problem Solving) emerges as one of the best ways to solve these types of problems most effectively and efficiently. Since it was established by Altshuller, the practical application of this methodology has proven to be the most organized and suitable invention and creative thinking methodology for knowledge-based innovation (KBI). TRIZ tools like the inventive algorithm ARIZ, the TRIZ contradiction matrix (CM), and the 76 standard solutions enable users to easily generate innovative solutions for their problems (Livotov *et al.*, 2019).

So, along with continuous improvement methodologies, innovation has a great role in the growth, survival, and success of any organization.

### 4. New model proposal

This section is intended to illustrate the creation of a new model specialized in the generation of innovative solutions with a corresponding action plan by prioritizing problems and solutions and involving the company's employees in the referred solutions. The new model is named CI-IMIM which means Continuous Improvement Integrated Model with Innovation and Management.

Before showing the fundamentals of CI-IMIM model it is important to understand the reason for its creation (4.2 and 4.3 subchapters).

#### 4.1 Characterization of existing models

Nowadays, there are many models and methodologies that integrate different types of tools within the areas of continuous improvement and innovation. These models are intended to take advantage of the complementarity characteristics of each one of the different tools applied together (Bariani *et al.*, 2004; Toivonen, 2015). The purpose of this combination of tools in this type of model is related to the proposal of innovative solutions in environments of continuous improvement that can solve real problems in a short period of time and that can prevent potential failures in future scenarios (Toivonen, 2015).

The Lean Six Sigma philosophy and the ecological perspective can and should also be considered in the improvement of processes (Wang & Chen, 2010; Yen & Chen, 2005).

Several methodologies/tools have become evident: the FMEA methodology, the FMECA methodology (Failure Mode, Effects & Criticality Analysis), the Brainstorming technique, the PDCA cycle, the 5Whys technique, the Pareto Diagram, the GUT matrix, the Ishikawa Diagram, the Eisenhower Matrix, the QFD (Quality Function Deployment), the FTA (Fault Tree Analysis), the Kano Model and other quality management such as control charts or scatter diagrams (Costa, 2018; Dias *et al.*, 2020; Ng *et al.*, 2017). The FMEA methodology is the one that stood out the most.

The alliance between the FMEA methodology and the TRIZ methodology is the most frequent in the attempt to create differentiating solutions. There are several examples that prove this alliance: according to Yen & Chen (2005) the application of TRIZ CM was fruitful in the search for solutions related to environmentally unsustainable failure modes shown by

FMEA; in the case of Vysotskaya & Dmitriev (2021), the requirements and basic parameters of the processes under study begin by being identified and analyzed either by the FMEA or by the QFD and later the TRIZ tools emerge as solutions to the technical contradictions highlighted before in the same case study.

The alliance of the areas of continuous improvement and innovation is also very important in the operational improvement of an organization. According to the approach of Wang & Chen (2010), the integration of the Lean Six Sigma philosophy with the TRIZ methodology helped to improve the process in the banking services sector. From a DMAIC cycle, the processes mapped through a VSM and a SIPOC were analyzed to determine the adjacent problems. Before applying the FMEA, a Cause-Effect Matrix, and a Pareto Diagram made it possible to highlight the root causes of the mentioned problems. Then, the application of the TRIZ methodology was useful in the development of solutions, mainly through MC. The combination of Lean philosophy with TRIZ can also be successfully observed through the creation of a continuous improvement model intended to create ideal solutions in a Portuguese company in the food industry. Here several Lean and management tools were used such as the Brainstorming technique, the 5W technique, the PDCA cycle, and the Kano model. Depending on the nature and complexity of the problem detected, the application of TRIZ problem-solving tools could be useful through the application of the 40 inventive principles, MC or Substance-Field Analysis. The 5S and line balancing can also be used as solutions (Dias et al., 2020).

## 4.2 Gaps

As noted in the previous subchapter, there is an immensity of models that interconnect various tools to get complementary benefits from them. Most of them have some gaps. In this sense, many of the models and/or methodologies of literature are limited.

The main gaps verified in the previous models are

**Table 2.** Description of CI-IMIM model parts

New Model part	Topic	Tool/Methodology
I	Problems numerical prioritization	FMEA (First part of FMEA card)
II	Improvement solutions proposal	TRIZ (contradiction matrix)
	Improvement solutions triage and prioritization	GUT
III	Improvement solutions action plans	5W1H
	Worker's feedback	Brainstorming
IV	Result analysis of taken actions	FMEA (Last part of FMEA card)

the following 6 (Ng et al., 2017; Sutrisno & Lee, 2011; Toivonen, 2015):

1. Difficulty in mapping processes;
2. Inefficiency in the analysis of failures and risks;
3. Psychological inertia;
4. Unknowledge of customer perspectives and needs;
5. Proposal of expensive improvement solutions;
6. Technical contradictions are often without effective resolution.

Although many of the models mentioned above manage to overcome some of these gaps, there is no single model that has the joint capacity to overcome all these limitations. Considering, for example, the methodology of Dias et al (2020), it manages to provide an integration of Lean tools (e.g., 5S) with management tools (e.g., 5W) and innovation (TRIZ). In this sense, this methodology overcomes the difficulty gaps in the mapping processes, psychological inertia, neglect of customer perspectives, and technical contradictions that are often without effective resolution. However, the inefficiency in the analysis of failures and risks and the proposal of solutions for costly improvement emerge as the biggest limitations of the methodology. The application of the FMEA methodology and the 5W1H technique would, by hypothesis, make this methodology even richer. Considering now the perspective of Costa (2018), the model incorporates the FMEA methodology, the 5W2H technique, and the GUT matrix, which reinforces the detailed and prioritized analysis of failures and risks and the mapping of processes. Despite this detailed analysis, the solution generation process does not follow any pattern or algorithm for creating innovative solutions, and feedback from internal and/or external customers is not considered.

### 4.3 New model structure

This model combines a mix of integrated tools which provide an intuitive and differentiating line of reasoning. In this way, it is possible to create innovative and differentiating solutions that directly involve the company's employees.

Each of the tools/methodologies used in the new model is allocated to its conceptual intervention area (topic) and part of the model (Table 2).

Through this model, it becomes possible to create solutions with action plans efficiently and innovatively. Additionally, the gaps evidenced in the previous subchapter cease to exist.

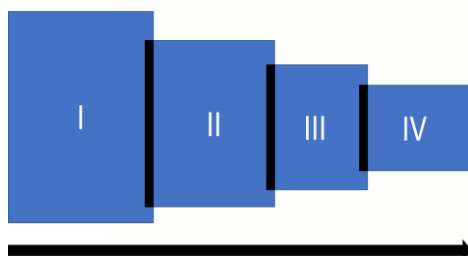
As CI-IMIM focuses on the phase of problem analysis and planning of improvement solutions (Plan), on the development of action plans for solutions and their implementation (Do), on the monitoring and analysis of the results (Check), and serves as a starting point for the discussion of results and the follow-up of future actions (Act), it seems to fit into a PDCA cycle.

Thus, it becomes possible to create solutions with action plans associated with their implementation efficiently and innovatively.

Although the model does not have any original Lean tool in its structure, the culture of the CI-IMIM model resides in the so-called Lean Thinking. In this sense, the culture of waste reduction is implicit in the model, namely during the process of generating innovative ideas via TRIZ methodology. Consequently, the solutions generated can be associated with Lean analytical tools, for example the application of VSM along with CI-IMIM model can help to overcome the process mapping so common gap.

### 4.4 New model fundamentals

Figure 1 schematically represents the new model over 4 parts sequentially.



**Figure 1.** Sequence of CI-IMIM model parts

Each of the parts is characterized by a table of characteristic indicators. The last column(s) of the table

of each part correspond(s) to the first column(s) of the following part table. For example, the last 2 columns of the table in Part I (Table 4) correspond to the first 2 columns of the table in Part II (Table 5). In this way, the CI-IMIM parts relationship can be called a cascade-shaped relationship. This connexion between parts of CI-IMIM favors its use in practice.

#### a. Part I

CI-IMIM is initialized with Part I where it starts by considering the problems associated with the respective root causes (inputs). These problems (failure modes) are characterized with the service stage and department in which they are inserted and the type of agents/employees of the company who are involved in the occurrence of the problem in the respective service stage.

Then, the problems associated with the root causes are numerically prioritized through the calculation of the associated RPN. This calculation follows the criteria of the FMEA chart according to Geum et al (2011).

The severity index (S) measures the severity of the failure effect caused by the associated failure mode, the occurrence index (O) measures the probability the root-cause responsible for the associated failure mode, and the detection index (D) measures the probability that the implemented means of control detect the root cause or the effect of the associated failure mode before it reaches the customer. The quantitative and qualitative scales can be found in Table 3. The classification of indices S, O, and D is between coefficients 1 and 9. According to the NPR classification, problems with an NPR greater than or equal to 200 are, by convention, considered priority

**Table 3.** Qualitative and quantitative scales of S, O and D [adapted from (Geum et al., 2011)]

Coefficient	Level	Criteria (S)	Criteria (O)	Criteria (D)
1	NO	No effect	Failure unlikely; History shows no failure	High quality available detection means
2	VERY SLIGHT	Customer not annoyed; Very slight effect on system performance	Rare number of failures likely	Proven Means of Detection
3	SLIGHT	Customer slightly annoyed; Slight effect on system performance	Very few failures likely	Simulated means of detection
4	MINOR	Customer experiences minor nuisance; Minor effect on system performance	Few failures likely	Detection means only tested at an early stage
5	MODERATE	Customer experiences some dissatisfaction; Moderate effect on system performance	Occasional number of failures likely	Detection means created in pre-simulation
6	SIGNIFICANT	Customer experiences discomfort; System performance degraded but operable and safe	Medium number of failures likely	Detection means compared to other similar systems
7	MAJOR	Customer very dissatisfied; System performance severely affected but functioning and safe	Moderately number of failures likely	Detection means compared to other type of components
8	SERIOUS	Customer very dissatisfied; System inoperable but safe	High number of failures likely	Disapproved or unreliable means of detection
9	EXTREME	Potential hazardous effect; System able to stop service (potentially null performance)	Very high number of failures likely	No known techniques available



**Table 4.** Part I from CI-IMIM model

Service Stage	Department(s)	Agent(s)	Failure Mode	Failure Effect	S	Failure Cause	O	Current Control Measures	D	RPN	
										Index	Classification
Type of worker's service	Departments (macro level) in which the corresponding service stage problem occurs	Elements of the company involved in the occurrence of the problem in the respective service stage	How the service fails	Consequent result of the occurrence of the failure mode	Severity of the failure effect caused by the associated failure mode	System failure which causes a certain failure mode	Likelihood of the root cause causing the corresponding failure mode	Ways of detecting current service failures	Likelihood that the implemented control means will detect the root cause or the effect of the corresponding failure mode before it reaches the customer	Risk Priority Number resulting from the product between S, O and D	Problem priority order

problems and, therefore, move on to the next part (Part II). The only exception is related to problems whose NPR is less than 200, but where the S index is maximum (G=9) so the associated problem also passes to the next part. In case of a tie of NPR's, the order of classification is arbitrary, and it is usual to follow the order in which the problems were initially presented.

This Part (Part I) can be seen in the Table below (Table 4) with the indicators "Service Stage", "Department(s)", "Agent(s)", "Failure Mode", "Failure Effect", "S", "Failure Cause "O", "Current Control Measures", "D" and "RPN".

### b. Part II

After calculating all the RPNs, the problems with a RPN greater than 200 are the problems which continue to be analyzed in Part II as they have the highest priority (priority problems).

Then, the improvement action proposal phase takes place in CI-IMIM. This phase is concerned with the creation of innovative solutions using the matrix of contradictions (CM), one of the main innovation tools of the TRIZ methodology. This tool allows the creation of innovative solutions through Inventive Principles defined according to certain Engineering Parameters of the referred methodology. According to Altshuller, there are 40 inventive principles and 39 technical

parameters. These are defined from the identification of contradictions between parameters which means the identification of a parameter that is intended to improve and another that will have to worsen, contradicting the potential effect from the other.

Thus, the new model focuses on the creation of recommended corrective actions according to the TRIZ methodology CM.

After the proposal of solutions, these are sorted and prioritized with the help of the GUT matrix. This tool measures 3 indexes: the severity of the impact the project may have on the company if it is not carried out soon (G), the urgency of carrying out the project with a deadline (U) and the tendency of the problem to worsen over time if the solution is not implemented (T). This decision support tool fits into the triage and prioritization of improvement actions. The criteria used, by convention, in the selection of improvement solutions is to consider solutions with a GUT index greater than or equal to 45. Afterwards, these are classified in order of priority.

This Part (Part II) can be seen in the Table below (Table 5) with the indicators "RPN", "TRIZ Engineering Parameters", "Chosen TRIZ Principles", "Recommended Corrective Actions", "G", "U", "T" and "GUT" (includes "Index" and "Classification").

**Table 5.** Part II from CI-IMIM model

RPN									GUT	
Index	Classification	TRIZ Engineering Parameters	TRIZ Invention Principles	TRIZ Principles Chosen	Recommended Corrective Actions	G	U	T	Index	Classification
Risk Priority Number resulting from the product between S, O and D	Problem priority order	Definition of 2 engineering parameters of the TRIZ methodology: an improving feature and a worsening feature depending on the type of problem	Definition of TRIZ inventive principles associated with the considered engineering parameters; it is done through the intersection of these parameters in the matrix of contradictions	Choosing the most suitable TRIZ inventive principle(s) to solve the problem	Improvement solutions design according to the respective TRIZ inventive principle(s)	Severity of impact that the solution could have on the company if it is not carried out soon	Urgency to carry out the solution considering the time factor	Problem tendency to get worse over time if the solution is not implemented	Product result between G, U and T	Improvement solution priority order



### c. Part III

The referred improvement actions are then subject to action plans that define the description of what the solution is ("What?"), the reason for its creation ("Why?"), in which department the employee works ("Where?"), when does it act ("When?"), who is involved ("Who?") and how to carry out the action in practice ("How?"). Here in Part III, the 5W1H management tool will be used. This can be seen in Table 6.

Before implementing the solutions, some of the company's employees will give their feedback about the referred solutions. Involving people is, therefore, a

Brainstorming and Kaizen meetings will be held to assess the acceptance of the company's workers involved in the solutions described so far. Consequently, the solutions with positive feedback move on to the next part (Part IV) which is the implementation itself. Solutions that may be considered useful in the long term do not move on to the next part of the new model since the solutions to be implemented only concern short medium-term improvement actions.

This Part (Part III) can be seen in Table 6 with the following indicators: "GUT" (includes "Index" and "Classification"); "Sorted Corrective Actions" (incorporates "What?", "Why?"; "Where?" "When?" "Who?" "How?" and "How much does it cost?") and "Feedback" (includes "Approval" and "Decision").

**Table 6.** Part III from CI-IMIM model

GUT		Sorted Corrective Actions						Feedback	
Index	Classification	What?	Why?	Where?	How?	Who?	When?	Approval	Decision
Product result between G, U and T	Improvement solution priority order	Description of the solution	The reason for the solution	Department or worker where the solution operates	Way of performing the solution in practice	Persons in charge involved in the solution	Improvement period	Worker opinion regarding solutions	Approval of the solution (😊/☹️)

critical piece in the acceptance of project solutions.

### d. Part IV

After having decided on the solutions that will be implemented, Part IV (Table 7) takes place in the model. Here, the respective results are measured. It is necessary to remember which modes, effects, and causes of failure justify the implementation of such improvement solutions. Control measures are also illustrated and updated here.

Finally, the implemented actions will be (as in Part I) subject to a numerical evaluation by calculating a new RPN (RPN'). The criteria used to calculate the S, O, and D indices in Part I are repeated here in Part IV. Then, the RPN' will be compared with the previous RPN to verify the positive or negative impact of the

action taken to solve the corresponding problem. This phase concerns the final part of what is usually present in the FMEA chart. Additionally, there is a last column with a value which corresponds to the difference between the previous RPN and the RPN' along with an arrow depending on whether the NPR has gone up or down.

This Part (Part IV) can be seen in the Table 7 with the following indicators: "GUT" and "Result of Taken Actions" (includes "S'", "O'", "D'", "RPN'" and "Comparison with previous NPR").

**Table 7.** Part IV from CI-IMIM model

Feedback		Implemented Solution	Respective Failure Mode	Respective Failure Effect	Respective Failure Cause	Updated Controls Detection	Result of Actions Taken				
Approval	Decision						R'	P'	N'	RPN'	Comparison with previous NPR
Worker opinion regarding solutions	Approval of the solution (😊/☹️)	Proposed solution that is going to be implemented and that is valued by workers	How the service fails	Consequent result of the occurrence of the failure mode	System failure which causes a certain failure mode	Updated ways of detecting current service failures	Severity of the failure effect caused by the associated failure mode after solution	Likelihood of the root cause causing the corresponding failure mode after solution	Likelihood that the implemented control means will detect the root cause or the effect of the corresponding failure mode before it reaches the customer after solution	Risk Priority Number resulting from the product between S, O and D	Comparison of Risk Priority Numbers before and after implemented solutions (↑/↓)

### 4.5 Practical Application Instruction

As CI-IMIM model practical application is very promising, it built a generic methodology that represents the practical application instruction along with CI-IMIM model (Figure 2). Its specific steps are included in detail in Figure 3.

prioritization, solutions' action plans and solution's monitoring and implementation. As it can be seen in Figure 3, these steps correspond to stages 6 to 10. However, if it is intended to integrate the model in a practical application scenario is also important to first understand the path to this identification of problems as well as it is essential to understand the path after

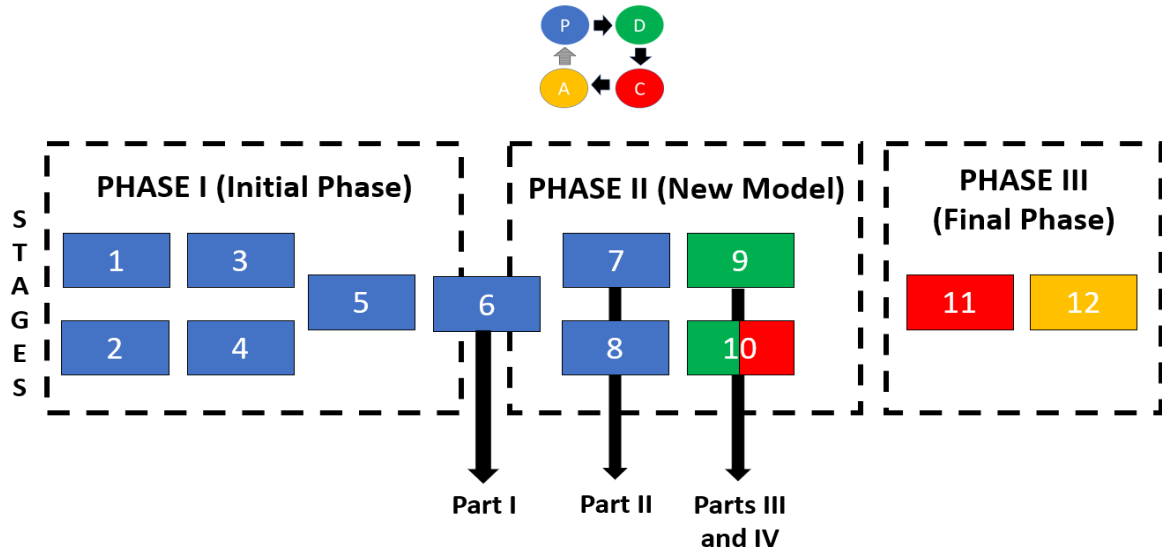


Figure 2. CI-IMIM practical application instruction

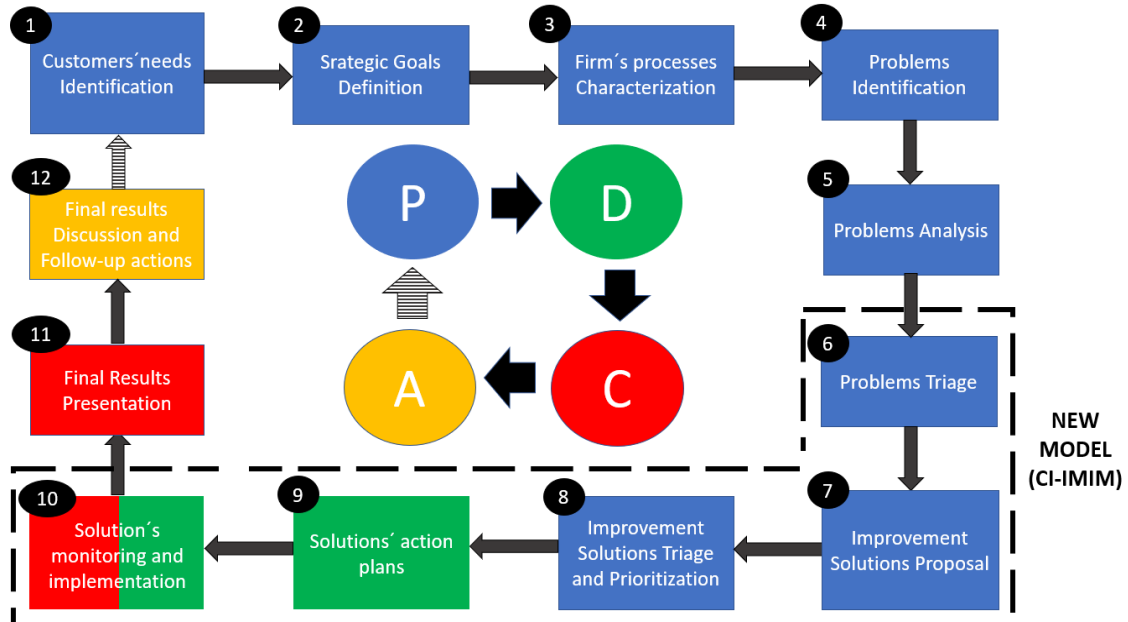


Figure 3. Description of practical application instruction stages

Through the previous explanation of CI-IMIM parts, it becomes clear that the new model includes the following 5 stages: problems triage, improvement solutions proposal, improvement solutions triage and

discovering the improvement solutions via CI-IMIM model.

The methodology shown in Figure 2 is organized into 3 Phases (I, II, and III) with 12 Stages (1 to 12). As

it was referred, the new model parts (I, II, III, and IV) correspond to stages 6 to 12 from Phase II. In addition, this methodology is based on a PDCA cycle. As it can be seen, the “Plan” includes the first 8 stages while the “Do” encompasses another 2 stages (9 and 10), the “Check” incorporates the 11th stage, and the “Act” corresponds to stage 12. Besides, the “Check” stage is either done simultaneously with the “Do” phase or at the end.

In general, the “Plan” step includes the customers’ needs identification, the strategic goals definition, the internal processes characterization, and the problems and solutions identification, analysis, and prioritization. The “Do” step involves the solutions’ action plan and implementation whereas the “Check” step is related to the solutions monitoring and results presentation. The final step “Act” corresponds to the discussion of the obtained results and the follow-up actions. The new model incorporates part of the “Plan”, and all “Do” and “Check” steps. This can be seen in Figure 3.

So, the integration of the CI-IMIM model in case studies proves to be an asset in decision making regarding innovative solutions in continuous improvement environments, which allows for solving a certain number of problems of any company or organization.

## 5. CI-IMIM in practice

After the theoretical explanation of CI-IMIM, it is time to apply the new model in a real case study. In this way, a case study is about to be carried out in a Portuguese car maintenance company. So, the validation of the new model is about to be studied here.

This case study occurred between March and July (2021). Although this case study followed the referred practical application instruction, this article aims to study Phase II in more detail as new model parts correspond to stages 6 to 10.

### 5.1 CI-IMIM background

In Phase I, it is important to have an initial notion of the most critical aspects for the customer in terms of service quality. In this context, a Critical to Quality Tree (CTQ) tree was built, representing the 3 After-Sales macro departments (MECHANICS, COLLISION and PARTS). This CTQ (Figure 4) served as the basis for translating the customers’ broad needs (internal and/ or external) into specific, actionable, and

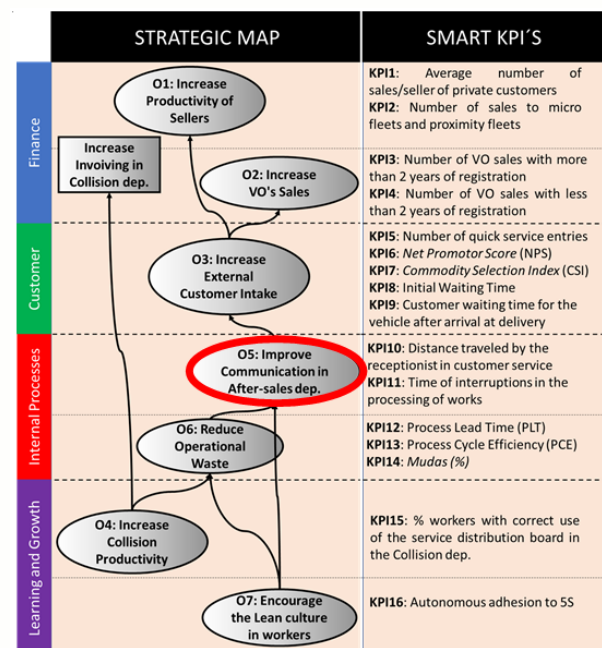
measurable performance requirements. It is possible to see that the Voice of the Customer (VOC) has 5 different CTQ aspects. All CTQs have a direct relation



with external customers except the CTQ “Lack of interdepartmental communication” which is directly

**Figure 4.** Customers’ needs identification via CTQ tree associated with internal customers.

After absorbing the most critical needs of the company's customers, it becomes essential to define the strategic objectives which the company is intended to achieve. In this way, a Balanced Scorecard (Figure 5) is built where a strategic map and SMART (Specific, Measurable, Attainable, Realistic and Time-Bound) KPI's are shown. The strategic map interrelates 7 different strategic goals organized according to their point of view: Financial, Customer, Internal Business Processes and Learning and Growth. Besides, there are



**Figure 5.** Balanced Scorecard of the company

specific KPI's related with the referred perspectives and goals. After meeting with the top management, it was decided to focus attention on the 5th Strategic Objective regarding the improvement of interdepartmental communication in the After-Sales department. There are 10 relevant SMART KPIs (KPI5 to KPI16 except KPI15). All these KPIs are directly related to the work of the receptionists.

As previously mentioned, the observation focused on the MECHANICS department (Reception, Car Repair, and Car Wash). Reception is the first department for direct contact with external customers. 3 receptionists, the Head of Car Repair department, and the mechanics are the active company's agents in Reception. In customer service situations there can be 3 different situations: scheduled appointments (priority), unscheduled appointments, and no appointments. These appointments typically concern overhauls, repairs, and car rentals, among others. The customer's request is organized through the customer file which incorporates the Repair Order (RO) and other relevant information about the appointment and the process.

In the Car Repair department, there are 8 mechanics and 1 Head of Car Repair department. Both interact actively in this department as well as with the receptionists and the Parts Clerk at MECHANICS. Each mechanic starts by cleaning the place of work and starts a new service after having gone to pick up the vehicle at the park. In the intervention itself, mechanics follow a preventive maintenance plan for overhauls and corrective maintenance for repairs.

The Car Wash department is coordinated by an outsourced company with 3 washers. Both the receptionist and the mechanic can access them here if they need to know the location in a queue or if a vehicle is ready or to transport a vehicle to/from this department.

The identification of problems occurred through internal documents, which were carefully analyzed, and the following Lean tools: direct observation (including Gemba Walks), VSM, Spaghetti Diagram, and 5S Checklist. Surveys and SIPOC were also used. Some of the Current State VSM results are presented in Table 8. The results indicated 37,9% of muda if it is

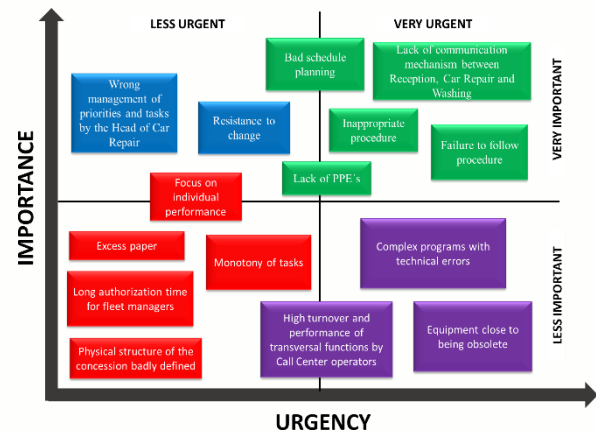
**Table 8.** NVA and VA time in Reception (Current State VSM)

Non-Value-Added Time (NVA)		Value-Added-Time (VA)
External NVA	Internal NVA	
06:37:33	02:27:43	14:54:43
72.91 %	27.09 %	

considered the total time of analysis of 24 daily hours (8 hours per receptionist). Here, 7% was a necessary waste. After identifying all the problems in MECHANICS (34 in total), it was time to analyze them and distinguish causes and effects.

The problems identified were then deeply analysed through the Ishikawa Diagram and the 5Why's technique. In this way, it was possible to get all root causes of the problems in the company.

After 15 root causes had been identified, it was



**Figure 6.** Eisenhower Matrix for identified root causes

useful to decide the most important root causes for the company according to O5. So, it was necessary to carefully choose the root causes to fight by allocating them in an Eisenhower Matrix (Figure 6). The criterion for choosing the root causes to fight was centered on choosing the most important causes for the company's performance. Thus, the priority quadrants are the top quadrants: Quadrant II (+ Important and + Urgent) and Quadrant I (+ Important and – Urgent). The causes allocated in Quadrant II are the following: "Lack of compliance with procedures", "Inappropriate procedure", "Lack of communication mechanism between Reception, Car Repair, and Car Wash", "Lack of PPE" and "Bad schedule planning". These are the 1st priority. Those with 2nd priority correspond to those in Quadrant I: "Resistance to change" and "Wrong management of priorities and tasks by the Head of Car Repair". The next phase (Phase II) was responsible for the application of a new model where the problems associated with the most important root-causes are studied.

## 5.2 CI-IMIM application

Phase II focuses on all the steps involved in CI-IMIM. Before applying the new model to this case study, it is important to point out that the inputs to the model are based on the corresponding problems of the previously most important root causes.

Since the new model is intended to provide a creative and innovative vision of solutions organized by priorities involving people in them, this model will be applied to help achieve the strategic objective of improving interdepartmental communication in APV (O5).

After using the Eisenhower matrix to qualitatively sort out the most important root causes for the strategic objective in question (O5), the corresponding problems will be deeply analyzed to be prioritized quantitatively. In this way, the inputs of the CI-IMIM are the problems associated with the previously screened root causes.



As can be seen in Table 9, the service stages are communication, worker procedures, schedule planning,

calculated from the product between the indices G, O, and D. Then, the problems are sorted in order of

**Table 9.** Part I from CI-IMIM model: application

Eisenhower Quadrant	Service Stage	Department(s)	Agent(s)	Failure Mode	Failure Effect	S	Failure Cause	O	Current Controls Measures	D	RPN	
											Index	Classification
1 <sup>st</sup> Priority (+ IMPORTANT and + URGENT)	Communication	MECHANICS	Receptionists and mechanics	Wrong timing in customer phone call for vehicle delivery	Long waiting time to pick up the vehicle after indication of the Reception that vehicle is ready	6	Lack of communication mechanism between Reception, Car Repair and Washing	7	No	9	378	3
			Receptionists, mechanics and washers	Lack of knowledge of the repair/wash status and location of the vehicle	Long distances traveled by workers when delivering the vehicle	6		8	No	9	432	2
				Misplacement of keys on the keychain	5	Failure to follow procedure		6	No	9	270	8
	Storage of vehicle keys	MECHANICS	Receptionists	Keys located on OR trays	Confusion when searching for keys and other documents	4	Inappropriate procedure	3	No	9	108	17
	Processing of works	MECHANICS	Head of Car Repair, mechanics and receptionists	Unnecessary approval of OR by the Head of Car Repair	Delays in processing works	6		9	No	9	486	1
			Mechanics and receptionists	Receptionists' poor understanding of documents filled in by mechanics		6	Yes (Receptionists usually confirm the documents from mechanics)	3	108	18		
	Billing Confirmation	MECHANICS	After-Sales responsible and receptionists	Lack of confirmation of email data regarding ORs provided by receptionists	Delay in receptionists' tasks	4	7	Yes (Receptionists check this situation in their emails)	5	140	16	
	Car delivery	MECHANICS	Receptionists and washers	Vehicle poorly washed or delivered without being washed	Poorly performed service	6	Failure to follow procedure	5	Yes (sometimes it is detected, and the vehicle goes back to Washing or it even doesn't come back)	5	150	15
	Start of mechanical operation	MECHANICS	Receptionists and mechanics	Lack of work inside the vehicle	Delays in the beginning of the mechanic's service at the Car Repair	4		3	Yes (Mechanic detects this when looking for a vehicle to start a new service)	2	24	22
	Warranty treatment	MECHANICS	Receptionists	Outdated Symptom Code	Return of warranties	7		3	Yes (Receptionists or Head of Car Repair validate warranties before sending it to the Warranty Center)	3	63	20
	Planning with customer	MECHANICS	Receptionists	Lack of prior planning	More time-consuming bureaucratic details	6		7	No	9	378	4
				Lack of agile allocation of vehicle delivery time	Misallocation of vehicle return hours	6	5	No	9	270	9	
	Customer service	MECHANICS	Receptionists	Disinfection performed at the end of customer reception instead of occurring immediately after inspection	Lack of focus on newcomers	4	Bad schedule planning	8	No	9	288	7
				High number of daily and pending tasks	Lack of focus on newcomers	3		4	No	9	108	19
					Stress	5		5	No	9	225	12
				High number of missed calls	5	6	Yes (by recording missed calls on the phone)	2	60	21		
	Car Repair dep. works	MECHANICS	Mechanics	Lack of safety in the work environment	Negative impact on worker health	9	Lack of PPE's	2	No	9	162	14*
	Organization of tasks	MECHANICS, COLLISION and PARTS	Receptionists, mechanics and Parts clerk	Disorganization and clutter of the work environment	Loss of information and/or lack of pride	5	Resistance to change	8	Yes (in the departments where 5S is implemented there are standards to be followed)	6	240	11
PARTS		Parts Clerk and mechanics	Mechanics await a quote at the parts counter	Inefficient communication between Car Repair and Parts Clerk at the parts counter	6	5		No	9	270	10	
2 <sup>nd</sup> Priority (+ IMPORTANT and - URGENT)	Head of Car Repair depart. role	MECHANICS	Receptionists depend on the Head of Car Repair in their high number of pending tasks	Frequent interruptions from receptionists to the Head of Car Repair	5	Wrong management of priorities and tasks by the Head of Car Repair	8	No	9	360	5	
			Head of Car Repair depends on the receptionists	Frequent interruptions from Head of Car Repair to the receptionists	5		7	No	9	315	6	
		MECHANICS and COLLISION	Heads of car Repair and receptionists	Poor interpersonal relationship between macro departments	Inefficient communication between macro departments		5	4	No	9	180	13

task organization, and the role of the Head of Car Repair department. Although the study focuses on MECHANICS, the problems under analysis also encompass agents from COLLISION and PARTS as they are related to MECHANICS. The agents involved are generally receptionists, mechanics, the Head of Car Repair department, Parts Clerks in MECHANICS, and washers. The problems' RPNs under analysis are

priority according to the numerical value of the calculated NPR. In case of a tie in NPR, the first tiebreaker is the Eisenhower Quadrant in question, otherwise, the ranking order does not matter.

Consider, as an example, the failure mode “Wrong timing in customer phone call for vehicle delivery”. This failure mode occurs when the receptionists make this phone call right after receiving the customer file from the mechanic. So, the agents involved are the receptionists and the mechanics. Because of this failure mode, the failure effect corresponds to a long waiting time to pick up the vehicle after the indication of the reception that the vehicle is ready. Then, the indices S, O, and D are calculated with the results of 6, 7, and 9. These values mean that this problem triggers a discomfort to the customer that usually occurs moderately in which there are no means of detection used by the company to prevent such occurrence of

failure. The RPN index has a high value of 378 which, compared to other NPRs, makes this problem the 3rd most serious problem. So, this problem moves to the next part of the model (Part II), in 3rd place.

From the 22 problems identified in Part I, only 13 went on to Part II (60% of the problems identified). It should be noted that all 1st Priority problems (+ Important and + Urgent) affect only the MECHANICS department.

Steps 7 and 8 correspond to Part II of the CI-IMIM. As there are 9 priority problems, these will be analyzed via TRIZ CM which helps in the proposal of innovative solutions for the company. In this way, the inventive principles that matter are responsible for

**Table 10.** Part II from CI-IMIM model: application

RPN									GUT	
Index	Classification	TRIZ Engineering Parameters	TRIZ Invention Principles	TRIZ Principles Chosen	Recommended Corrective Actions	G	U	T	Index	Classification
486	1	Improving “Loss of time” (25); worsening “Difficulty of detecting and measuring” (37)	18, 28, 32 e 10	[28] Mechanics substitution	Change procedure (work approval by receptionist)	4	5	3	60	8
				[10] Preliminary action	Training of receptionists regarding the approval of works derived from the car Repair	4	3	3	36	16
432	2	Improving “Loss of time” (25); worsening “Ease of operation” (33)	4, 28, 10 e 34	[4] Asymmetry	Motivate workers to contact each other (transparency of information)	4	2	2	16	25
				[28] Mechanics substitution	Digital means to receive vehicle status information (eg app, excel, etc.)	5	5	4	100	1
				[10] Preliminary action	Create a virtual vehicle control board from reception to delivery (eg app, excel, etc.)	4	4	4	64	6
				[34] Discarding and recovering	Washing paper sheet used in the service distribution board (Car Repair) instead of being thrown away to the trash	4	4	4	64	7
378	3	Improving “Adaptability” (35); worsening “Device complexity” (36)	15, 29, 37 e 28	[15] Dynamics	Energizing receptionist: he/she only makes the phone call after getting definitive information about the vehicle being ready; he/she moves into the Car Repair and/or Washing departments	4	5	2	40	15
				[28] Mechanics substitution	Digital means to receive vehicle status information (e.g., app, excel, etc.)	5	5	4	100	2
378	4	Improving “Loss of information” (24); worsening “Loss of time” (25)	24, 26, 28 e 32	[24] Mediator	Plan in a shared way with other receptionists (teamwork)	3	4	4	48	10
				[26] Copying	Print an expected sheet of information to be needed the following day per customer	2	3	2	12	26
				[28] Mechanics substitution	Create ready-made inputs for future works	3	3	2	18	21
360	5	Improving “Reliability” (27); Worsening “Adaptability” (35)	13, 35, 8 e 24	[32] Colour changes	Planned ROs with different colors from unplanned ROs	2	2	2	8	29
				[35] Transformation of properties	Building a priority framework for the Head of car Repair to streamline their tasks in the right way	4	3	2	24	18
				[8] Anti-weight	Distribute and differentiate team priorities from individual ones	3	3	2	18	22
315	6	Improving “Loss of time” (25); worsening “Ease of operation” (33)	4, 28, 10 e 34	[24] Mediator	Training receptionists on some topics	4	4	3	48	11
				[28] Mechanics substitution	Planning the Head of Car Repair’s dep. agenda in accordance with the receptionists’ planning	5	4	4	80	5
288	7	Improving “Adaptability” (35); worsening “Ease of operation” (33)	4, 28, 10 e 34	[10] Preliminary action	Building a priority framework for the Head of car Repair to streamline their tasks in the right way	4	3	3	36	17
				[10] Preliminary action	Perform Disinfection soon after inspection in the presence of the customer	4	3	4	48	12
270	8	Improving “Reliability” (27); worsening “Ease of operation” (33)	15, 34, 1 e 16	[15] Dynamics	Dynamizing washer: washer puts the keys in the right place, informing (computer-aided with a beep) the receptionist of the vehicle ready	5	5	4	100	3
				[34] Discarding and recovering	Key placement must be done in the right place on the keychain	3	3	2	18	23
				[1] Segmentation	Division into 2 subtasks: place key in the keychain and notify the receptionist personally	3	3	2	18	24
270	9	Improving “Reliability” (27); worsening “Loss of time” (25)	10, 30 e 4	[10] Preliminary action	Construction of a task board with tasks for each receptionist	5	4	3	60	9
270	10	Improving “Adaptability” (35); worsening “Loss of time” (25)	10, 30 e 4	[10] Preliminary action	Hire/allocate someone to be always available to mechanics (Dynamic Parts clerk)	5	5	4	100	4
240	11	Improving “Loss of information” (24); worsening “Loss of time” (25)	24, 26, 28 e 32	[24] Mediator	Put post-sits on pending works	3	4	2	24	19
				[28] Mechanics substitution	Conduct weekly 5S internal audits at the Reception	5	3	3	45	14
225	12	Improving “Loss of energy” (22); worsening “Productivity” (39)	28, 10, 29 e 35	[10] Preliminary action	Define working hours with more breaks and social spirit among workers	2	3	2	12	27
				[29] pneumatics and hydraulics	Have more comfortable chairs at the Reception	2	3	4	24	20
162	14*	Improving “Object-generated harmful factors” (31); worsening “Device complexity” (36)	19, 1 e 31	[19] Periodic action	Make the use of PPE’s recurring in the Car Repair	3	4	4	48	13
				[1] Segmentation	Have convertible/dismountable PPE’s	3	1	3	9	28
				[31] Porous materials	Buy PPE of porous material	3	1	2	6	30

suitable improvement solutions. In this sense, there can be so many improvement solutions for the same problem depending on the number of invention principles applicable to the problem. However, not all inventive principles are applicable, so it is important to select them before moving on to the solution. After the improvement solutions have been noted, they are then sorted using the GUT matrix that classifies them according to the G, U, and T indexes. Finally, the improvement solutions are screened by importance. All of these features can be seen in Table 10.

Consider the example of the problem “Unnecessary approval of RO by the Head of Car Repair department” according to Part I. This issue is associated with the “Inappropriate procedure” root cause. To propose a differentiating solution that valued less wasted time, it was decided to worsen the convenience of use, on the other hand. Thus, the application of the matrix of contradictions of the TRIZ methodology suggested 4 inventive principles:

mechanical vibration (18); mechanical substitution (28); color changes (32), and preliminary action (10). It is easy to see that mechanical vibration and color changes correspond to physical inventive principles meaningless to the associated problem. Thus, these are disposable and the other 2 were chosen for improvement proposals. According to the principles of mechanical substitution and preliminary action, respectively, the solutions were to change the procedure (Change procedure – work approval by receptionist) and training of receptionists regarding the approval of works derived from the Car Repair department. Both solutions are quite valid even though when they are classified according to the GUT matrix, only the first one is considered a priority since it has a GUT index of  $60 \geq 45$ . After all GUT indexes have been calculated, it was found that this solution occupied the 8th position in terms of the priority of solutions.

**Table 11.** Part III from CI-IMIM model: application

GUT		Sorted Corrective Actions						Feedback	
Index	Classification	What?	Why?	Where?	How?	Who?	When?	Approval	Decision
100	1	Digital means to receive vehicle status information (e.g. app, excel, etc.)							
100	2								
100	3	Dynamizing washer: washer puts the keys in the right place, informing (computer-aided with a beep) the receptionist of the vehicle ready	Avoid wasting receptionists' time and energy	Reception and Car Repair's computers	Create a Digital <i>Kanban</i> board to control vehicles by repair status	Vasco Soares	Second half July	Overall added value	😊
64	6	Create a virtual vehicle control board from reception to delivery (e.g., app, excel, etc.)							
100	4	Hire/allocate someone to be always available to mechanics (Dynamic Parts clerk)	Avoid wasting mechanics' time	Car Repair	Place Parts Clerk with an active role in the Car Repair ( <i>Mizusumashi</i> )	Top management	September/October?	Need to test with extra parts clerk	😞
80	5	Planning the Head of Car Repair's agenda in accordance with the receptionists' planning	Avoid wasting both time and try avoiding unforeseen events	Head of car Repair and receptionists	Situation points between the Head of Car Repair and receptionists for vehicles in the Car Repair	Head of Car Repair dep.	End of July	This type of control prevents future problems	😊
64	7	Washing paper sheet used in the service distribution board (Car Repair) instead of being thrown away to the trash	Taking advantage of the reuse of paper that would be thrown away and that helps the receptionist to know which vehicles are in the Washing department	Car Repair's service distribution board	Motivate workers to put the RO's sheets (which they usually put in the trash) in the "Washing" tab of the service distribution board after having taken vehicles to Washing; works as a contingency plan for the Kanban virtual board	Vasco Soares	Second half July	Necessary practice which follows procedure	😊
60	8	Change procedure (work approval by receptionist)	Avoid wasting time in processing works that come from the Car Repair	Reception	Put works directly from the Car Repair to Reception	Head of Car Repair dep.	End of July	More practical	😊
60	9	Construction of a task board with tasks for each receptionist	Help managing and planning receptionists' tasks and priorities	Reception	Create visual management sheet regarding ROs by RO status and by priority of receptionists	Vasco Soares	Second half July	Monitoring ROs becomes more efficient	😊
48	10	Plan in a shared way with other receptionists (teamwork)	Plan for the next day	Reception	Promote sharing moments at the end of each day related with next day customers	Vasco Soares	Second half July	Appointments for the following day are always very incomplete	😞
48	11	Training receptionists on some topics	Decrease dependence on the Head of Car Repair and increase the autonomy of receptionists	Reception	Have weekly training sessions on topics such as warranties, fleet management, continuous improvement, etc.	After-Sales responsible	September/October?	There are many other trainings pending.	😞
48	12	Perform Disinfection soon after inspection in the presence of the customer	Increase focus on newcomers	Receptionists' inspection	Motivate workers about improving service efficiency with this procedure by alerting them to the waiting time that newcomers have with disinfection at the end	Vasco Soares	Second half July	Logical procedure	😊
48	13	Make the use of PPE's recurring in the Car Repair	Increase the safety of mechanics' operations	Car Repair	Buy PPE's and make workers aware of the importance of their use	After-Sales responsible	Without defined date	Need for gloves of different material	😞
45	14	Conduct weekly 5S internal audits at the Reception	Increase the efficiency of the receptionists' work	Reception	Create 5S program	Vasco Soares	July and August	Organization and pride are valued, so 5S should be encouraged	😊

1º PARK	OFICINA (EM SERVIÇO)	OFICINA (AGUARDA AUTORIZAÇÃO)	OFICINA (AGUARDA PEÇA)	LAVAGEM	PRONTO	ENTREGUE
<b>82ZU81</b>	<b>35B020</b>	<b>82SG87</b>	<b>51ZP32</b>	<b>80qx31</b>	<b>AH02TT</b>	<b>50zv40</b>
Local de Estacionamento: C	ultima plicagem 1/8 10:46		Val para Loures	Na Lavagem	Local de Estacionamento: C	Concluído
Responsível: Susana Poças	Responsível: Rafael Castanho	Responsível: António Rodrigues	Responsível: Alexandre Luis	Responsível: Rafael Castanho	Responsível: António Rodrigues	Responsível: Lavador
Data: 10:00	Hora: 10:45	Hora: 14:15	Hora: 16:40	Hora: 11:30	Hora: 08:58	Hora: 17:17
<b>980M82</b>	<b>af81pd</b>	<b>48UG59</b>	<b>28BP46</b>	<b>84SA44</b>	<b>67SF17</b>	<b>59ZA79</b>
Local de Estacionamento: M	concluído amanhã	3 dias	2 d	Na Lavagem	Local de Estacionamento: N.D.	Concluído
Responsível: Alexandre Luis	Responsível: Roberto Lourenço	Responsível: Alex Geronymo	Responsível: Roberto Lourenço	Responsível: Rafael Castanho	Responsível: Lavador	Responsível: Ricardo Rafael
Data: 08:57	Hora: 12:12	Hora: 12:00	Hora: 14:19	Hora: 17:26	Hora: 15:37	Hora: 11:00

**Figure 7.** Digital Kanban Board in Excel

From the 13 problems, 30 solutions for improvement emerged. Here only 14 went on to Part III (about 50% of the total solutions defined). And here the “Plan” step ends.

Steps 9 and 10 mark the 3rd part of the application of CI-IMIM to this case study (Part III). These steps relate to the “Do” step of the PDCA cycle in play. As can be seen in Table 11, the improvement solutions (ordered by priority) are detailed using the 5W1H tool. In this context, the action plan of the improvement solution is detailed according to the questions “What?”, “Why?”, “Where?”, “How?”, “Who?” and “When?”.

As can be seen, of the 11 solutions, 7 of them had positive approval from the company's workers. Feedback from employees was obtained from 2 independent brainstorming moments: a formal brainstorming with a receptionist, a mechanic, the Head of the Car Repair department, the person in charge of Quality (intermediate management), the head of the After-Sales (intermediate management) and the General Director (top management) and a series of informal brainstorming sessions with the company's employees for whom the solutions had the greatest impact. In this way, the satisfaction of the company's internal customers is more easily achieved since each of the company's workers got involved, giving their

endorsement to the solutions that were effectively decided to implement.

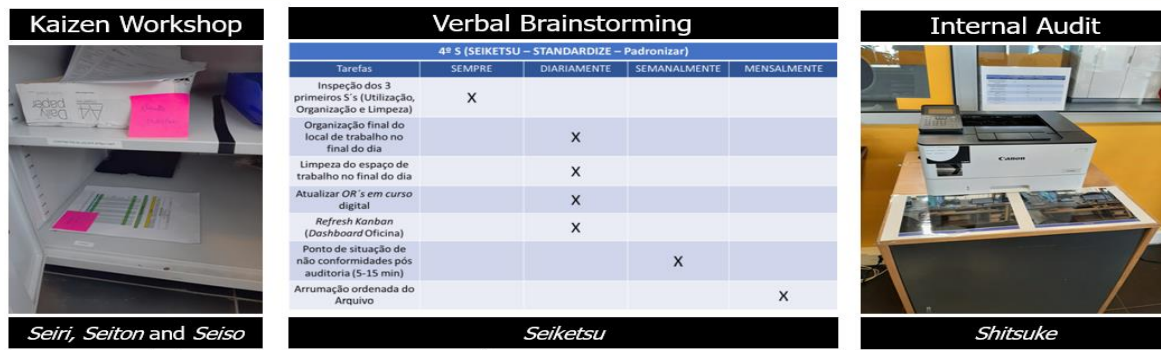
In this sense, the following solutions are considered in order of priority:

1. Digital Kanban board to control vehicles by repair status (7 different repair status in Figure 7);
2. Regular meetings between the Head of Car Repair department and receptionists for vehicles in the Car Repair department;
3. Motivate workers to put the RO's sheets (which they usually put in the trash) in the Washing tab of the service distribution board after the service is finished and before taking vehicles to the Washing (contingency plan for the Digital Kanban board);
4. Orientate customers' files directly from the Car Repair department to the reception department;
5. Create Visual Management sheet (task and priority planning table) regarding RO's by RO status and by priority of receptionists (Figure 8);
6. Motivate workers in improving the efficiency of the service with this procedure, alerting them about the waiting time that new arrivals have with the disinfection being carried out at the end;
7. Create 5S program to create new continuous improvement routines in Reception (Figure 9).

Aparar FECHADAS		Até 5 dias	De 6 a 15 dias	De 16 a 30 dias	De 31 a 60 dias	De 61 a 90 dias	De 91 a 180 dias	TOTAL			
Alexandre Luis		3	17	28	9	1	0	58			
Margarida Ferro		0	12	4	4	1	0	21			
Susana Poças		2	4	5	8	1	0	20			
Hélio Pita		0	7	2	2	1	0	12			
Nádia Evaristo		0	1	1	0	0	0	2			
							TOTAL	52047,41			
							MÉDIA	531,10			
							TOTAL	22349,12			
							MÉDIA	228,05			
Nº	Matrícula	Intervenção	Comentário (Motivo obra em curs)	Planeada	Contacto	Responsível	Estado	Venda Bruto	Venda Líquido	Antiguidade	Data Inicial
3716	535298	Verificar viatura não pega//Entrou em				Alexandre Luis	Preparar Fatura	241,40	217,26	De 6 a 15 dias	15/jul
3737	83R006					Alexandre Luis	Preparar Fatura	8,38	8,38	De 6 a 15 dias	16/jul
3738	15Z237	Viatura em Realnty e em frio as rotações sobem e descem				Margarida Ferro		41,92	41,92	De 6 a 15 dias	16/jul
3744	AH20HX	CLT queixa-se por vezes ouve um estalo à frente quando acelera mais				Alexandre Luis		66,20	66,20	De 6 a 15 dias	16/jul

**Figure 8.** Visual Management sheet in Excel




**Figure 9.** 5S Program (Kaizen Workshop + Verbal Brainstorming + Internal Audit)

### 5.3 CI-IMIM Validation

After having implemented all 7 improvement solutions together, several types of results were obtained:

results from Part IV of CI-IMIM, data from the New Current VSM, and other indicators. This step corresponds to the “Check” of the PDCA cycle.

Regarding the results from Part IV of CI-IMIM, these are present in the last 2 columns of Part IV. The 7 solutions are here associated with the respective mode, effect, and cause of failure (also present in Part I of the

new model) to facilitate the calculation of the new NPR indices (NPR'). Part IV is shown in Table 12. It should also be noted that these 7 improvement actions aim to combat 9 failure modes.

According to Part IV of CI-IMIM, all the proposed solutions had a positive effect on the problem they were intended to solve. The calculation of the NPR' followed a subjective criterion. The solutions “Digital Kanban board to control vehicles by repair status” and “Motivate workers to put the RO's sheets in the Wash tab of the service distribution board after the service is finished” had the greatest effect in reducing RPNs associated with the lack of knowledge of the repair/wash status and location of the vehicle

**Table 12.** Part IV from CI-IMIM model: application

Feedback		Implemented Solution	Respective Failure Mode	Respective Failure Effect	Respective Failure Cause	Updated Controls Detection	Result of Actions Taken				
Approval	Decision						R'	P'	N'	RPN'	Comparison with previous NPR
Overall added value	😊	Create a Digital Kanban board to control vehicles by repair status	Wrong timing in customer phone call for vehicle delivery	Long waiting time to pick up the vehicle after indication of the Reception that vehicle is ready	Lack of communication mechanism between Reception, Car Repair and Washing	Yes (if customer is waiting too long in the queue, there is evidence)	6	5	4	120	↓ 258 (120 < 378)
			Lack of knowledge of the repair/wash status and location of the vehicle	Long distances traveled by workers when delivering the vehicle		Yes (tabs “State” e “Parking place” in the Excel Virtual Kanban board)	6	5	3	90	↓ 342 (90 < 432)
			Misplacement of keys on the keychain	Failure to follow procedure	Yes (tabs “State” e “Parking place” in the Excel Virtual Kanban board)	5	4	3	60	↓ 210 (60 < 270)	
This type of control prevents future problems	😊	Situation points related with vehicles in Car Repair	Head of Car Repair depends on the receptionists	Frequent interruptions from Head of Car Repair to the receptionists	Wrong management of priorities and tasks by the Head of Car Repair	Yes (tab “Situation point with Head of Car Repair” in the Excel Virtual Kanban board)	5	5	3	75	↓ 240 (75 < 315)
Necessary practice which follows procedure	😊	Motivate workers to use the “Washing” tab in the service distribution board	Lack of knowledge of the repair/wash status and location of the vehicle	Long distances traveled by workers when delivering the vehicle	Lack of communication mechanism between Reception, Car Repair and Washing	Yes (check by visual observation if mechanics use the “Washing” tab)	6	5	3	90	↓ 342 (90 < 432)
More practical	😊	Put works directly from the Car Repair to Reception	Unnecessary approval of OR by the Head of Car Repair	Delays in processing works	Inappropriate procedure	No	6	5	9	270	↓ 216 (270 < 486)
Monitoring ROs becomes more efficient	😊	Create visual management sheet related with RO's in progress	Lack of agile allocation of vehicle delivery time	Misallocation of vehicle return hours	Failure to follow procedure	Yes (visual Management in Excel of the RO's in progress allows to control the information of all vehicles in progress)	5	4	3	60	↓ 210 (60 < 270)
Logical procedure	😊	Disinfection right after inspection in customer service	Disinfection performed at the end of customer reception instead of occurring immediately after inspection	Lack of focus on newcomers		Yes (visual control)	4	2	2	16	↓ 272 (16 < 288)
Organization and pride are valued, so 5S should be encouraged	😊	5S program	Disorganization and clutter of the work environment	Loss of information and/or lack of pride	Resistance to change	Yes (periodical audits)	5	6	3	90	↓ 150 (90 < 240)



(reduction of 342 values compared to the previous RPN).

Some of the new current state VSM results are presented in Table 13. The results indicate 23,8% of

Non-Value-Added Time (NVA)		Value-Added-Time (VA)
External NVA	Internal NVA	
03:08:44	02:34:34	18:16:41
54,98 %	45,02 %	

**Table 13.** NVA and VA time in Reception (New Current State VSM)

muda if it is considered the total time of analysis of 24 daily hours (8 hours per receptionist). Here, 7% is a necessary waste.



**Figure 10.** 5S Program Results for 1 Receptionist

The 5S results from one of the receptionists' workplaces can also be seen in Figure 10.

After having obtained the results present in the previous step (stage 11), the same results will be discussed and analysed in depth (stage 12). At this

stage, the "Act" of the PDCA cycle is distinguished. The evolution of the indicators under study is shown in Table 12.

The application of CI-IMIM in this case study proved to have been very fruitful in terms of finding innovative solutions suitable for solving the main problems observed in the MECHANICS department. As mentioned above, the solutions "Digital Kanban Board to control vehicles by repair status" and "Motivate workers to use the Wash tab" were the most successful ones. The first solution mentioned involves workers from all micro departments of MECHANICS (Reception, Car Repair and Car Wash departments). Although the fulfilling of the post-Inspection Disinfection procedure did not have the greatest impact, this solution was the one that obtained the lowest NPR (NPR went from 288 to 16). All indicators had a positive evolution. In fact, the average of the differential presented in the last column of Part IV of the new model had the value of 249. So, this evolution was significant. One of the main reasons involved here was the fact that the control measures have been improved for cases with no detection mechanism. Additionally, it was found that 5 out of 13 root causes previously investigated were fought: "Lack of communication mechanism between Reception, Car Repair and Car Wash departments"; "Lack of compliance with procedure"; "Wrong management of priorities and tasks by the Head of Car Repair department"; "Inappropriate procedure" and "Resistance to change". The first 2 were the most valued according to the criteria of the new model.

**Table 14.** Summary table of the evolution of KPI's

Indicators	Description	Perspective	Before	After	Target	Evolution
RPN	CI-IMIM RPN's	All	...	...	0	POSITIVE
KPI6	Net Promotor Score (NPS)	Customer	87,9%	73,9%	100%	NEGATIVE
KPI7	Customer Satisfaction Index (CSI)	Customer	9,3	9,1	10,0	NEGATIVE
KPI8	Initial Waiting Time	Customer	10,3 min	7,3 min	0 min	POSITIVE
KPI9	Customer waiting time for the vehicle after arrival at delivery	Customer	20 min	7,8 min	0 min	POSITIVE
KPI10	Distance traveled by the receptionist in customer service	Internal Processes	658 m	614 m	334 m	POSITIVE
KPI11	Time of interruptions in the processing of works	Internal Processes	7,4 min	6,0 min	0 min	POSITIVE
KPI12	Process Lead Time (LT)	Internal Processes	3h e 21 min	2h e 36 min	1h e 34 min	POSITIVE
KPI13	Process Cycle Efficiency (PCE)	Internal Processes	20,2%	22,4%	100,0%	POSITIVE
KPI14	Mudas in Reception	Internal Processes	37,9%	23,8%	7%	POSITIVE
KPI16	Voluntary adhesion to 5S by receptionists	Learning and Growth	0%	33%	100%	POSITIVE

Thanks to the solutions proposed via CI-IMIM, most of the KPIs relevant to this case study (KPI6 to KPI14 + KPI16) evolved positively comparing March (initial month) with July (final month). These are presented in Table 14.

In the future, the follow-up of actions takes place, from now on, in the project to establish actions aimed at the maintenance, analysis, and improvement of the applied improvement solutions (PDCA cycle Act). In this sense, these actions serve as a starting point for new action plans to catapult a new PDCA cycle. In this way, the identification of new problems, improvement opportunities, and solutions take place again. Some of the suggested actions are as follows:

1. Follow-up meetings – serve as a form of formal monitoring of the referred KPIs;
2. Procedure manuals/good practice guides for new solutions – allows the creation of standards that feed the learning culture of the company's employees;
3. Existence of an extra parts clerk who is dynamic in the Car Repair department (*Mizusumashi*) – this operator can simultaneously do his/her tasks and be close to the mechanics with pre-planned material.
4. Implementation of an *Andon* system in the event of certain signs – sound/light signals that allow workers to be aware of a problem and/or some outstanding action intervening in their tasks;
5. Level scheduling of appointments through a *Heijunka* system – unscheduled appointments (scheduled with the Head of Car Repair) should occur less and less gradually; while they occur, they must be dealt with by the receptionists and not with the Head of Car Repair to avoid as much as possible unforeseen events or other interruptions;
6. Computerization of tasks which involve communication between departments through an app or web app that can be accessed on tablets by workers – makes the Digital Kanban mechanism more intuitive and faster.

## 8. Conclusions

This article is involved in the design and application of a new model specialized in the creation and prioritization of differentiated and innovative solutions through continuous improvement and innovation methodologies and with the company's involvement, too. This model is called CI-IMIM. The practical application of the new model occurred in a Portuguese car dealer company. It was intended to achieve strategic goal 5 (O5) which corresponds to the

improvement of interdepartmental communication in After-Sales in MECHANICS.

Theoretically, CI-IMIM seemed to be very useful in any company/organization that would like to create and prioritize unique solutions. Besides, it proved to fit well with tools like the Ishikawa diagram, the 5Why's technique, and the Eisenhower matrix.

In the present case study, Phase II is the study methodology phase which is effectively concerned with the application of the new model. This model is divided into 4 parts. In summary, 22 problems were identified in Part I. Here, 13 problems passed to Part II where 30 improvement solutions arose. 14 (3 repeated) out of 30 solutions passed to Part III where 7 solutions were selected to be implemented in the company. The “Digital Kanban board to control vehicles by repair status.” and “Motivate workers to use the Wash tab” were the most successful solutions as they were responsible for a greater reducing effect in the RPN index. The Visual Management sheet and the 5S program were other Lean solutions that stand out in MECHANICS, namely in the reception.

Between March and July, 8 of the 10 KPIs were significantly positive, which proves the practical validation of CI-IMIM in car dealers like the one here. The KPIs of Internal Processes were the ones that stood out the most in a positive way, such as the PCE which improved by 2.2%, and the mudas in Reception which decreased by 14.1%. However, the NPS and CSI indicators evolved negatively because of the pandemic context and the need for an adaptation period that allows the new solutions to be well assimilated by the company.

For future work, it is recommended that new PDCA cycles should be based on the maintenance, analysis, and improvement of the improvement solutions applied in this case study. Follow-up meetings, procedures associated with the new solutions, more computerized tasks, and systems with *Mizusumashi*, *Andon* and *Heijunka* are suggested.

## Acknowledgements

The authors acknowledge Fundação para a Ciência e a Tecnologia (FCT.IP) for its financial support through the grant UIDB/00667/2020 (UNIDEMI).

## 9. References

- Andrés-López, E., González-Requena, I., & Sanz-Lobera, A. (2015). Lean Service: Reassessment of Lean Manufacturing for Service Activities. *Procedia Engineering*, 132, 23–30.
- Bariani, P., Berti, G., & Lucchetta, G. (2004). A Combined DFMA and TRIZ approach to the simplification of product structure. *Proceedings of the Institution of Mechanical Engineers, Part B: Journal of Engineering Manufacture*, 218(8), 1023–1027.
- Clarke, C. (Ed.). (2005). The history of production systems in the automotive industry. In *Automotive Production Systems and Standardisation: From Ford to the Case of Mercedes-Benz*. 71–125. Physica-Verlag HD.
- Costa, A. (2018). *Mapping risks and controls in the third sector through risk management tools*. MBA dissertation, Universidade Federal do Paraná, Brasil (in Portuguese)
- Dias, A., Navas, H., & Abreu, A. (2020). Design of a Continuous Improvement Model in a Portuguese Food Industry Company – A Case Study. *KnE Engineering*, 195–208.
- Geum, Y., Cho, Y., & Park, Y. (2011). A systematic approach for diagnosing service failure: Service-specific FMEA and grey relational analysis approach. *Mathematical and Computer Modelling*, 54(11), 3126–3142.
- Holweg, M. (2008). The Evolution of Competition in the Automotive Industry. Em G. Parry & A. Graves (Eds.), *Build To Order: The Road to the 5-Day Car* (pp. 13–34). Springer. London, UK.
- Kaitwade, N. (2020). COVID-19 shatters global automotive industry; sales of metal powder take a nosedive amid wavering demand. *Metal Powder Report*, 76, 137-139
- Liker, J. (2004). *The Toyota Way: 14 Management Principles from the World's Greatest Manufacturer* (1st edition). McGraw-Hill Education, New York, US.
- Liker, J. K., & Morgan, J. M. (2006). The Toyota Way in Services: The Case of Lean Product Development. *Academy of Management Perspectives*, 20(2), 5–20.
- Livotov, P., Chandra Sekaran, A. P., Law, R., & Reay, D. (2019). Systematic Innovation in Process Engineering: Linking TRIZ and Process Intensification. *Advances in Systematic Creativity: Creating and Managing Innovations*. 27–44). Springer. Cham, Switzerland.
- Ng, W., Teh, S., Low, H., & Teoh, P. (2017). *The integration of FMEA with other problem solving tools: A review of enhancement opportunities*, 890, Article 012139.
- Raisinghani, M. S., Ette, H., Pierce, R., Cannon, G., & Daripaly, P. (2005). Six Sigma: Concepts, tools, and applications. *Industrial Management & Data Systems*, 105(4), 491–505.
- Shang, G., & Low, S. (2014). *Lean construction management: The Toyota way*, 390.
- Sutrisno, A., & Lee, T. (2011). Service reliability assessment using failure mode and effect analysis (FMEA): Survey and opportunity roadmap. *International Journal of Engineering, Science and Technology*, 3(7), 25–38.
- Tohidi, H., & Jabbari, M. M. (2012). The important of Innovation and its Crucial Role in Growth, Survival and Success of Organizations. *Procedia Technology*, 1, 535–538.
- Toivonen, T. (2015). Continuous Innovation – Combining Toyota Kata and TRIZ for Sustained Innovation. *Procedia Engineering*, 131, 963–974.
- Vysotskaya, M. V., & Dmitriev, A. Y. (2021). *Improve the integrity testing process based on QFD, FMEA and TRIZ*. IOP Conference Series: Materials Science and Engineering, Workshop on Materials and Engineering in Aeronautics (MEA 2020), 1027, Article 012029, Moscow, Russia
- Wang, F., & Chen, K. (2010). Applying Lean Six Sigma and TRIZ methodology in banking services. *Total Quality Management & Business Excellence*, 21(3), 301–315.
- Womack, J., & Jones, D. (1996). *Lean Thinking: Banish Waste and Create Wealth in Your Corporation*. *Journal of the Operational Research Society*, 48 (11). 1148-1148.
- Yen, S., & Chen, J. (2005). An Eco-Innovative Tool by Integrating FMEA and TRIZ Methods. *2005 4th International Symposium on Environmentally Conscious Design and Inverse Manufacturing*, 678–683.

## AUTHOR BIOGRAPHIES



Vasco Ventura Soares holds a master's degree in industrial engineering and management from NOVA School of Science and Technology, Portugal. He took the Third Place Award in the International Conference on Systematic Innovation (ICSI, China,

October 15-17) in the category "Technical Aspects of Innovation Methods" and the Silver Medal Award in the Global Competition on Systematic Innovation (GCSI, China, October 15-17). He has also published a paper in a Portuguese Journal called Techniques, Methodologies and Quality (TMQ). Besides, he has more than 1 year of experience in the field of continuous improvement (e.g. Lean Thinking) due to work in a consumer electronics company and in a car maintenance company. Additionally, he has 3 more months of experience in a small consulting company in areas like Quality, Safety and Energy. His areas of interest include Systematic Innovation including TRIZ, Design & Manufacturing Management, Lean Six Sigma Management and Continuous Improvement.



Helena Navas has received a PhD from the NOVA School of Science and Technology – Universidade NOVA de Lisboa – Portugal. She is currently Assistant Professor at the Department of Mechanical and Industrial Engineering of the NOVA

School of Science and Technology - Portugal and researcher at UNIDEMI. She represents the Portuguese Association for Quality (APQ) in the Portuguese Technical Standardization Committee on Research Activities, Development, and Innovation (RDI) and represents FCT NOVA in the Portuguese Technical Standardization Committee on Project Management. Her research interests are in the areas of innovation, continuous improvement, quality, and process management.



## Feature selection using binary particle swarm optimization algorithm to predict repurchase intention from customer reviews

Dimas Adrianto, Dedy Suryadi\*

Industrial Engineering Department, Parahyangan Catholic University, Bandung, Indonesia

\* Corresponding author E-mail: dedy@unpar.ac.id

(Received 31 March 2022; Final version received 12 August 2022; Accepted 15 January 2023)

### Abstract

Indonesia has the most prominent digital economy in Southeast Asia and has a promising market for e-commerce companies to compete and dominate the online market share. This also gave rise to an increment in the number of customer reviews of a product or service provided. Online customer reviews can be utilized to analyze the repurchase intention of e-commerce customers. However, many features appearing in customer reviews increased the repurchase intention predictive model complexity. A process to choose a subset of features and reduces the number of features in data is called feature selection. This paper proposed a method of feature selection to pre-process the inputs to the predictive model. The selection is performed using a metaheuristic called Binary Particle Swarm Optimization (BPSO) combined with Sentiment Orientation-Pointwise Mutual Information to sort the features. The sorting corresponds to the particle dimension, which is a part of the particle encodings that affect the metaheuristic's performance in solving the problem. The results show that the proposed method reduces and selects the best features to construct a predictive model of repurchase intention from online customer reviews on two datasets that are written both in Indonesian and English. Compared to the baseline model before performing feature selection, the accuracy of the predictive models evaluated using k-Nearest Neighbors on both datasets increased by 5.40% (75.91% to 81.31%) and 8.50% (71.37% to 79.87%), respectively.

*Keywords: Binary Particle Swarm Optimization, Feature Selection, Online Customer Reviews, Repurchase Intention.*

## 1. Introduction

Indonesia is the tenth largest economy in the world and is expected to have the largest digital economy in Southeast Asia and will likely reach \$330 billion in value by 2030 (Chandra 2021). An article published by ISEAS in August 2021 (Negara and Soesilowati 2021) stated that Indonesia's digital economy has approximately 40% of the total regional market share. Rising internet penetration and the usage of smartphones in Indonesia are among the main factors contributing to the remarkable growth of e-commerce. The online market or e-commerce in Indonesia has continued its expansion during the COVID-19 pandemic. In addition, We Are Social in Digital 2021: Indonesia (DataReportal 2021) reported that, in January 2021, 87.1% of internet users in Indonesia aged 16 to 64 purchased a product online from any device. Many e-commerce companies have been competing to dominate the online market share in Indonesia. Note that, to dominate the market, companies must develop and improve the quality of products and services they offer and provide.

Repurchase intention is a customer's judgment on repurchasing the same products or services from the same company based on their experience (Hellier et al., 2003). Based on Social Learning Theory (SLT), self-efficacy influences people's decisions about what actions to take. People tend to learn from performance accomplishments (i.e., past experiences) to influence their actions. The role of self-efficacy is vital in the repeat buying behavior of e-commerce customers (Chen 2012).

Quality is an essential factor related to repurchase intention (Suryadi 2020). Customer satisfaction is crucial in mediating the effect of quality variables positively on customers' repurchase intention, such as product quality (Vashti and Antonio 2021). This argument is also supported by the research stating that product quality positively affects repurchase intention and is associated with customer satisfaction. In addition, price perception, through customer satisfaction, affects repurchase intention (Suhaily and Soelasih 2017). The previous research implies that quality and price perception are associated with customer satisfaction. When customers are satisfied, customers tend to make repeat purchases.

Most research and papers on repurchase intention or customer satisfaction rely on common traditional data collection methods, such as interviews (Mendoza

2020), surveys (Nguyen et al., 2021) (Trivedi and Yadav 2018), questionnaires (Tsai et al., 2016). These methods, however, are time-consuming and costly to carry out. This paper offers a relatively new method of predicting repurchase intention using online customer reviews. Customer review contains textual content on customers' opinion of products or services they had experienced that is related to their repurchase intention.

Nowadays, machine learning is an important technique in a broad area, including e-commerce and marketing. In marketing, machine learning offers advantages and is usually compared to traditional methods such as econometric methods. The role of machine learning techniques allows e-commerce to make strategic and crucial decisions on time. Moreover, machine learning provides valuable insights for e-commerce marketers and product manufacturers to improve their products or services (Suryadi 2020). Furthermore, feature selection and optimization may be applied to machine learning to achieve higher efficiency and performance for many problems, including marketing problems (Brei 2020).

Customer reviews are documents that may be represented as a collection of words or bag-of-words (features). These documents are high-dimension vectors where each dimension corresponds to the number of features (Nedjah et al., 2009). To solve this complexity, feature selection may be proposed to reduce the number of irrelevant and redundant features in the feature space while improving the accuracy of the constructed prediction model. In machine learning and statistics, feature selection is a technique to reduce the dimensionality of data by choosing subsets that consist only of relevant features by removing irrelevant, redundant, or noisy features from the original feature set (Miao and Niu 2016). Feature selection has been applied in broad practical applications, such as image processing (Bins and Drapper 2001), bioinformatics (Saeys et al., 2007), text mining (Forman 2003), and (Du et al., 2019).

Text mining has become more popular as it tries to gather valuable information from textual data (Dang and Ahmad 2014). Forman (2003) performed an experimental study on 229 text classification problem instances of twelve feature selection methods. Du et al. also performed feature selection on textual problems, specifically on online customer reviews. However, with the emergence of textual analysis, especially on online customer reviews,

the research on analyzing Indonesian textual content, such as customer reviews, is still limited. This paper uses customer reviews (written both in Indonesian and English) collected from the Indonesian cosmetics e-commerce, sociolla.com.

This research aims to apply feature selection to textual features (i.e., customer reviews) using Binary Particle Swarm Optimization (BPSO). Feature selection on textual features is used to help to construct a better and more efficient predictive model. The Binary Particle Swarm Optimization (BPSO) algorithm is a version of Particle Swarm Optimization that has been used in binary problems (Xue et al., 2014). Examples of binary problems are those involving labels such as "yes" or "no", "included" or "not included" (Khanesar et al., 2017). The subsequent section will further detail the arguments for selecting BPSO.

Similar previous works on selecting textual features using BPSO have the purposes of sentiment classification and text summarization (Shang et al., 2016; Suganya & Priya, 2017; Suganya et al., 2019). Those purposes are different from this paper's purpose, i.e., classifying customer repurchase intention. Customers who intend to repurchase may express

## 2. Literature reviews

### 2.1 Metaheuristics for feature selection

Feature selection is a technique for choosing a subset of features in data used during the pre-processing step (Cherrington, et al. 2019). Feature selection has two objectives: first, it is to reduce the number of irrelevant and redundant features (Bing et al., 2013) and to improve model significance and performance (Cherrington, et al. 2019). Feature selection is arduous as there can be a complex interaction between features and a large search space (Bing et al., 2013).

Feature selection may be performed by a number of methods. There are three main categories of feature selection algorithms: filter approaches, wrapper approaches, and hybrid approaches. The filter approach uses independent criteria. The wrapper approach, such as the metaheuristics algorithm, is bound to the predetermined classifying algorithms. In other words, this approach considers the interaction between the classifying algorithm and the metaheuristic algorithm (Shroff and Maheta 2015). The hybrid approach

positive and negative sentiments in their reviews. That is also the case for customers who do not intend to repurchase. Therefore, this paper's purpose is different from sentiment classification in the previous works.

Moreover, considering the essential role of encoding in metaheuristic algorithms, this paper proposes a novel way to sort the features (which correspond to the dimensions of a particle in BPSO) according to a measure called Sentiment Orientation-Pointwise Mutual Information (SO-PMI). Encoding the particle in BPSO is important because encoding influences the algorithm's performance in solving the problems (Zavala et al., 2014). More specifically, the importance is due to the dependence of the movement or variation operators in BPSO on the encoding (Osaba et al., 2022).

The remainder of this paper is divided into four sections as follows. In Section 2, literature reviews are presented. Section 3 presents the methodology used in this research. Section 4 presents the results of a case study and a discussion of the experiments in this research, followed by conclusions and potential future works in Section 5.

combined the filter approach and wrapper approach. Research done by Kohavi and John (Kohavi and John 1997) concludes that wrapper approaches are superior to filter approaches, such as tf-idf, even though filter approaches are arguably less expensive computationally.

Metaheuristic algorithms, such as Particle Swarm Optimization (PSO), Ant Colony Optimization (ACO), and Genetic Algorithm (GA), are methods that may obtain a near-optimal solution to a complex problem effectively. The binary version of Particle Swarm Optimization (BPSO) is chosen in this paper, considering that PSO computing time is the fastest among the three algorithms, even though GA generates a better performance (Gunantara and Putra 2019). Also, compared to GA, PSO is easier to implement to its few parameters and can converge faster (Cervante et al., 2012).

Compared to ACO, research that compared ACO, PSO, and a proposed hybrid method of ACO-PSO on a feature selection problem by Menghour and Souci-Meslati in 2016 showed that the simple PSO approach is the second best method in general, after the hybrid ACO-PSO (Menghour and Souci-Meslati 2016). The effectiveness of PSO on feature selection is also supported by survey research on feature selection using

five different binary metaheuristic algorithms (Binary Particle Swarm Optimization, Binary Differential Evolution, Binary Antlion Optimizer, Binary Grey Wolf Optimizer, and Binary Gaining Sharing Knowledge Based Algorithm). Even though, in general, the Binary Gaining Sharing Knowledge Based Algorithm performs better than the other four, Binary Particle Swarm Optimization (BPSO) takes relatively less computational time (depending on the size of the datasets) (Agrawal et al., 2021). Due to these results and advantages, PSO is considered a promising method for feature selection.

Other algorithms, such as random forest, AdaBoost, and gradient boosting, also can be utilized for feature selection. However, to the best of our knowledge, AdaBoost and other boosting algorithms select features based on their importance. The top-ranked features (based on their individual feature importance) are selected (Sun et al., 2011). These algorithms above do not generally consider the interaction between features. Wrapper method such as metaheuristic algorithms (Binary PSO in this case) has the ability to consider and capture the interaction between features.

Table 1 summarizes the examples of literature reviews on feature selection. Suryadi (Suryadi 2020) proposed a machine learning-based method to predict repurchase intention using online customer reviews on the cosmetics dataset collected online. The filter approach (tf-IDF and Fisher score) was used to represent a review's textual content and reduce the number of features. Then, three classification models were performed, and the results were significantly higher in the accuracy of three categories of products compared to the baseline model.

Particle Swarm Optimization was used on the feature selection problem. Bing et al. applied PSO to perform feature selection on fourteen datasets collected

from the UCI machine learning repository, plus six additional datasets. Kristiyanti and Wahyudi (Kristiyanti and Wahyudi 2017) proposed a method of using the wrapper approach for feature selection on opinion mining cosmetic product reviews. They compared the classification performance of Principal Component Analysis (PCA), Particle Swarm Optimization (PSO), and also Genetic Algorithm (GA). The method proposed in their paper is to use those three algorithms mentioned before so the accuracy of the Support Vector Machine (SVM) as a machine learning classifier algorithm in text classification can be increased. The results showed that the PSO-based SVM algorithm outperformed the other algorithms mentioned above.

Shang et al. (Shang, Zhou and Liu 2016) proposed a method using Binary PSO on feature selection. They modified it to overcome feature selection problems in sentiment classification, including unreasonable velocity update formula and lack of evaluation of a single feature. The modification is called Fitness-sum Proportionate Selection Binary Particle Swarm Optimization (FS-BPSO). The results indicated that FS-BPSO performed better than BPSO.

Suganya et al. (Suganya, Lavanya, & Gowrisankari, 2019) focused on using Fitness Based Particle Swarm Optimization (FBPSO) to select subsets of features for sentiment classification and summarization problems on a hotel review dataset. According to their experiments, FBPSO improved the performance using ROUGE-N metric as a performance evaluation of summary compared to the probabilistic ranking approach. Suganya and Priya (2017) also proposed a similar method on a hotel review dataset using a binary version of PSO (Binary PSO). The results also indicated that ROUGE-N metrics were improved.

**Table 1 Summary of the literature reviews on feature selection.**

Reference	Dataset	Method
(Bing et al., 2013)	Fourteen datasets from UCI machine learning repository and six additional datasets	Applying Particle Swarm Optimization on feature selection increases classification performance and reduces features and computational time.
(Shang et al., 2016)	Two UCI benchmark datasets	Modify PSO to become Fitness-sum Proportionate Selection Binary Particle Swarm Optimization (FS-BPSO) for feature selection in a sentiment classification problem.
(Kristiyanti and Wahyudi 2017)	Amazon's Cosmetic Product Review	Comparing the performance (accuracy) of PSO, GA, and PCA on feature selection. These algorithms are combined with SVM.
(Suganya & Priya, 2017)	Hotel Review Dataset (collected from TripAdvisor)	Applying PSO on feature selection to the sentiment classification and text summarization problems.
(Suganya et al., 2019)	Hotel Review Dataset of cities such as Beijing and London (collected from TripAdvisor)	Applying a proposed Fitness Based BPSO feature selection to the sentiment classification and text summarization problem.
(Suryadi 2020)	Online Customer Reviews (collected from sociolla.com)	Applying Fischer Score to reduce the dimensionality of textual features



## 2.2 Binary particle swarm optimization

Particle Swarm Optimization, or PSO, as one of the Evolutionary computation techniques, has been used in broad optimization fields, including feature selection (Bing et al., 2013). In 1995, Particle Swarm Optimization (PSO) was first introduced by Dr. Eberhart and Dr. Kennedy. Particle Swarm Optimization (PSO) is a metaheuristic swarm-intelligence-based algorithm that simulates the social behavior of birds. Each particle is a vector in a multidimensional search space, and each particle within a swarm has its position and velocity. Each particle moves according to the current particle velocity, the best position the particle has explored, and the global best position the swarm has explored

(Khanesar et al., 2017). Each particle updated its velocity using formula (1), and the position of each particle is updated using formula (2).  $r_1$  and  $r_2$  are random numbers.

$$v_{pd}^{new} = w \times v_{pd}^{old} + c_1 \times r_1 \times (pbest_{pd} - x_{pd}^{old}) + c_2 \times r_2 \times (gbest_d - x_{pd}^{old}) \quad (1)$$

$$x_{pd}^{new} = x_{pd}^{old} + v_{pd}^{new} \quad (2)$$

As described in the pseudocode below, the algorithm starts with initializing random particles and their position. The process stops when the process has reached the stopping criterion (Bing et al., 2013).

```

Initialize p-particle, particle position;
repeat
evaluate fitness value of each particle f(x);
update pbest of each particle and gbest;
for (each particle p = 1,2, ..., p) do
update velocity using formula 1;
update particles position using formula 2;
if f(x) > pbest, then pbest = f(x);
if pbest > gbest, then gbest = pbest;
end for
until ("stopping criterion is true")
    
```

Cherrington et al. (2019) summarized several traditional feature selection methods and reviewed the advantage of the Particle Swarm Optimization (PSO) filter based on feature selection. They also highlighted feature selection's limitations, opportunities, and improvement methods. They argued that initialization techniques could affect performance. Particle Swarm Optimization (PSO) must be tuned before performing feature selection to improve efficiency, performance, and analysis.

The main difference between BPSO and PSO is that the position of each particle is limited to [0,1] using formula (4), and to update the particle position.

The velocity is first transformed into a sigmoid value using formula (3).

$$S(v_{pd}) = \frac{1}{(1 + e^{-v_{pd}})} \quad (3)$$

$$x_{pd} = \begin{cases} 1, & \text{if } r_2 \leq S(v_{pd}) \\ 0, & \text{otherwise} \end{cases} \quad (4)$$

As described in the pseudocode of BPSO below, the main process of the BPSO algorithm is almost identical to PSO. The main difference is that before updating the particle positions, the velocity of each particle is transformed into a sigmoid value using a sigmoid function (Xue et al., 2014).

```

Initialize p-particle, particle position;
repeat
evaluate fitness value of each particle f(x);
update pbest of each particle and gbest;
for (each particle p = 1,2, ..., p) do
update velocity using formula 1;
calculate the sigmoid value using formula 3;
update particles position using formula 4;
if f(x) > pbest, then pbest = f(x);
if pbest > gbest, then gbest = pbest;
end for
    
```

**until ("stopping criterion is true")**

### 2.3 Sentiment orientation-pointwise mutual information

Pointwise Mutual Information (PMI) is a measure of association between two items based on information theory. In natural language processing, Pointwise Mutual Information compares the probability of two words dependently with the probability of those two words independently. Sentiment Orientation (or Semantic Orientation) is a numerical rating that indicates the direction of the sentiment of a word (i.e., positive or negative). Sentiment Orientation-Pointwise Mutual Information (SO-PMI) is a measure of the relevance of words to a reference sentiment word (i.e., positive or negative) using the information of words' co-occurrence in a corpus (Turney 2002).

The sentiment orientation of a word is calculated by comparing its correlation to a positive reference word with its correlation to a negative reference word (in this case, the words are "yes" or "no" repurchase intention) (Turney 2002). Mathematically, a word is assigned a numerical rating by the correlation to the positive reference word and subtracted to a numerical rating by the correlation to the negative reference word (Turney 2002). If the SO-PMI value of two words' co-occurrence is high, they have a strong correlation (Zhao, Zhang and Chai 2015). SO-PMI is calculated using formulas (6) and (7).

$$PMI(x; y) = \log_2 \frac{p(x, y)}{p(x)p(y)} \quad (6)$$

$$SO\ PMI(x) = PMI(x, positive) - PMI(x, negative) \quad (7)$$

$PMI(x; y)$  is a PMI value of  $x$  and  $y$ .  $p(x, y)$  is a value of co-occurrence numbers of  $x$  and  $y$  in the document. *The notation*  $p(x)$  is a value of occurrence numbers of  $x$  in the document. *The notation*  $p(y)$  is a value of occurrence numbers of  $y$  (repurchase intention "yes" or "no" in this case) in the document (Zhao, Zhang and Chai 2015).

In this paper, SO-PMI provides a meaningful encoding for particle dimensions by sorting features (words) based on their polarity towards repurchase intention reference words. If the SO-PMI value of a word is high (in other words, positive), that word has a

strong correlation with "yes" repurchase intention and vice versa.

### 2.4 k-nearest neighbors

The k-Nearest Neighbors (k-NN) classifier is a supervised machine learning algorithm. The k-NN is proposed to classify labels (repurchase intention "yes" or "no") by ranking the training data based on the Euclidean distance from their neighbors and comparing the label with k-most similar neighbors (Mandong and Munir 2018). Mathematically, formula (5) measures the Euclidean distance within the data.

$$d_i = \sqrt{\sum_{i=1}^p (x_{2i} - x_{1i})^2} \quad (5)$$

Govindarajan and Chandrasekaran (Govindarajan and Chandrasekaran 2010) evaluated the k-Nearest Neighbors (k-NN) classifier. They demonstrated their approach on an existing direct marketing dataset that classifies customers based on their characteristics. Their experiments showed that the proposed k-Nearest Neighbors (k-NN) performed better in accuracy. They concluded that k-Nearest Neighbors (k-NN) is not a problem-dependent algorithm and can be used for other problems or datasets.

Generally, k-NN is the simplest machine learning algorithm compared to Naïve Bayes and Support Vector Machine (SVM). Although Naïve Bayes tends to be much faster than k-NN when it applies to big data, Naïve Bayes could suffer from the zero probability problem and would result in a biased prediction. Compared to Support Vector Machine (SVM), k-NN is better if the training data is larger than the number of features. SVM is better with a dataset with a low sample size but a very high number of inputs (features) (Bzdok et al., 2018). Since this paper aims to see how a metaheuristic algorithm (Binary Particle Swarm Optimization) works on a textual feature selection problem, particularly on the problem of predicting repurchase intention, only k-NN is selected to lower the computational time due to its simplicity. Also, k-NN is the most preferred and most used classifier among all (Agrawal et al., 2021).

Therefore, this paper proposed combining a metaheuristic algorithm and a measure of associations

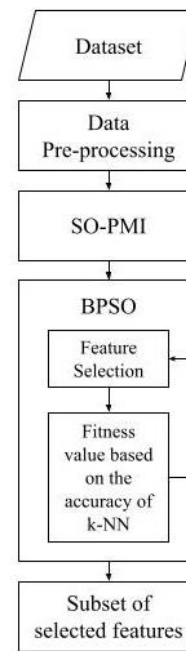
between textual features. Binary Particle Swarm Optimization (BPSO) as a metaheuristic algorithm is proposed to select the best features subset according to the highest accuracy of the k-NN prediction model.

### 3. Research methodology

This paper aims to implement Binary Particle Swarm Optimization (BPSO) combined with Sentiment Orientation-Pointwise Mutual Information (SO-PMI) to select the best features subset to construct a predictive model of repurchase intention from customer reviews. The research methodology in this paper is presented in five main steps, as shown in Fig. 1. This research methodology provides a systematic way of feature selection using a metaheuristic algorithm.

The first step is to gather the data. In this paper, the dataset is a set of customer reviews as well as the repurchase intention of each review. The dataset is gathered from sociolla.com, a beauty and cosmetic product e-commerce website in Indonesia. The product category used in this customer reviews dataset is limited only to the Moisturizer category from March 18, 2019, to August 26, 2019. The dataset is divided into two parts, i.e., Dataset 1: 2,614 reviews and Dataset 2: 6,439 reviews. These datasets are the reduced datasets from more than 120,000 reviews collected.

An imbalanced dataset is a common issue in a machine learning problem, especially in a real-world problem (such as feature selection in this research). Imbalanced datasets would impact the correlation between features, and the accuracy would be biased and inaccurate. The imbalanced issue on both datasets (Dataset 1 and Dataset 2) is addressed by using an undersampling method by randomly deleting data in the dataset from the majority class (in this case, the “no” label is the majority class) to balance it with the minority class. The result of under sampling is displayed in Table 2.



**Fig. 1. Research methodology**

**Table 2 Treating imbalanced dataset by undersampling**

Dataset		Label (Repurchase)	
		“Yes”	“No”
1	Before	611	2,003
	After	611	611
	Difference	0	-1,394
2	Before	2,539	3,900
	After	2,539	2,539
	Difference	0	-1,361

Both datasets are then randomly split using the sklearn package (Pedregosa, et al. 2011) in Python language programming into two samples, i.e., 80% training sample and 20% test sample. For dataset 1, the training sample has 978 reviews, and the rest, 244 reviews, are in the test sample. For dataset 2, the training sample has 4,062 reviews; the rest, 1,016 reviews, are in the test sample.

The cross-validation method was not used in this research, even though cross-validation would help to reduce the chance of overfitting. Cross-validation would increase training time and is computationally very expensive, as it needs a high and powerful processing system (hardware and software) (Joulani et al., 2015). In this research, there are two relatively large datasets (Dataset 1: 5,698 tokens (or words) with a total frequency of 54,981 and Dataset 2: 15,090 tokens (or words) with a total frequency of 297,422). Therefore, cross-validation was arguably not required

since these datasets were considered sufficiently large. Also, since the datasets are randomly split into the training set and the test set, even though cross-validation was not used, it is expected that the test set already represents the population of data. Therefore, the performance metrics (accuracy, precision, recall, and F-score) may also represent the expected performance.

Most customer reviews in the datasets are written in Indonesian and English. The examples of customer reviews collected are in Table 3. The first review in the examples written in Indonesian has a repurchase

intention of "yes," and the second one is written in English with a repurchase intention of "no."

Subsequently, the reviews in both datasets are transformed (during the data pre-processing step) into feature vectors. The contents of the feature vector are the frequency per feature per review. Based on Table 3, the value of  $X_{N,n}$  is the frequency of the N-th feature that appears in the n-th review. For example, if the value of  $X_{N,n}$  is 3, then the N-th Feature appears three times in the n-th review.

**Table 3 Examples of customer reviews collected**

Customer Review	Repurchase Intention (Yes/No)
"Pelayanannya bagus, harganya juga lebih murah dibandingkan sama toko sebelah" ("The service was great, also the price was cheaper than other stores")	Yes
"This product is so bad; I don't want to buy it anymore."	No

Next, the SO-PMI values of the features are calculated. Then, the features are sorted from the lowest SO-PMI value (strongly correlated to "no" repurchase intention) to the highest SO-PMI value (strongly correlated to "yes" repurchase intention). For

example, "sedih" ("sad") has the lowest SO-PMI value of -5.0348, and "suka" ("like") has the highest SO-PMI value of 4.8457. As an illustration, in Table 4, the first feature is the word with the lowest SO-PMI value, and the last feature is the word with the highest SO-PMI.

**Table 4 Illustration of feature vectors as the encodings of particles**

Review	Feature			Repurchase Intention
	1 <sup>st</sup> Feature	...	N <sup>th</sup> Feature	
1 <sup>st</sup> Review	1	...	$X_{N,1}$	Yes
...	...	...	...	...
n <sup>th</sup> Review	0	...	$X_{N,n}$	No

Subsequently, to classify the repurchase intention of each review, the k-NN algorithm is applied using the selected features. The accuracy is then used on Binary Particle Swarm Optimization (BPSO) as the fitness value of particles. BPSO algorithm guides the search for the best set of features, although not guaranteed to be optimal. The k-NN algorithm was used in this research since k-NN is generally an easy and simple machine learning algorithm.

The accuracy represents the fitness of a particle after implementing the selected features into the k-NN algorithm. The accuracy is the ratio of the number of correct predictions to the total number of predictions.

## 4. Experiments and results

After the datasets are collected, the data are then split into tokens (tokenization), transformed into standard forms (stemming), and stop word removal is performed. Most of the reviews in the datasets are written in Indonesian and English. Both dictionaries (Indonesian and English) are used in this paper using packages such as NLTK (Bird et al., 2009) and PySastrawi (Robbani 2018). The examples of data pre-processing are shown in Table 5.



**Table 5 Data pre-processing**

Before	After
“Pelayanannya bagus, harganya juga lebih murah dibandingkan sama toko sebelah” (“The service was great, also the price was cheaper than the other stores”)	[‘layan’, ‘bagus’, ‘harga’, ‘lebih’, ‘murah’, ‘banding’, ‘toko’, ‘sebelah’]
“This product is so bad; I don’t want to buy it anymore”	[‘product’, ‘bad’, ‘want’, ‘buy’]

Tokens are features that will be selected. The total tokens appear in both datasets (5,698 tokens and 15,090 tokens). The frequencies of each token that appears in the documents are calculated and sorted from the highest frequency to the lowest. Features are now reduced to 341 features and 605 features using Pareto Principle (80/20). Pareto Principle (80/20) removes the features (words) that do not belong to the set of the most frequent words that constitute 80% of the total frequency of all words. The remaining features are then sorted using SO-PMI from the lowest SO-PMI value (strongly correlated to “no” repurchase intention) to the highest SO-PMI value (strongly correlated to “yes” repurchase intention). The idea of implementing SO-PMI was to sort the feature index on particles and

to capture the effects of sentiment orientation of features towards the repurchase intention (“yes” or “no”) on model accuracy.

Some of the results are presented in Table 6. Using the SO-PMI formula, the SO-PMI value of the word “Sedih” (equivalent to “sad” in English) is -5.0348 and is negative, so the word “Sedih” is strongly correlated to “no” repurchase intention. The word “Sedih” is equivalent to “sad” in English, and it is reasonable that the word “Sedih” is correlated with “no” repurchase intention. The word “Suka” (equivalent to “like” in English) has a SO-PMI value of 4.8457 and is positive, so it is strongly correlated with “yes” repurchase intention.

**Table 6 Example of SO-PMI values**

Features	SO-PMI
Sedih (Sad)	-5.0348
Lacock (Not suitable)	-4.7192
...	...
Wajib (Compulsory)	4.8457
Cantik (Beautiful)	4.8457
Suka (Like)	4.8457

The sorted features are then transformed into a specific model (particle encoding for BPSO example shown in Table 7), and accuracy-based fitness value feature selection is performed. Based on Table 7, if the value of  $X_{N,n}$  feature is 1,  $X_{N,n}$  feature is selected in the particle. Conversely,  $X_{N,n}$  feature is not selected in the particle. The value [0,1] is a particle position for each dimension.

**Table 7 Particle encoding in BPSO**

Feature No.	1	2	...	N-1	N
Feature	“Sedih” (Sad)	“Gacocok” (Not suitable)	...	“Cantik” (Beautiful)	“Suka” (Like)
1 <sup>st</sup> Particle	1	0	...	1	0
...	...	...	...	...	...
n <sup>th</sup> Particle	0	1	...	1	0

The experiment was performed using PyCharm 2020 with a Python 3.7 64-bit version. The sklearn (Pedregosa, et al. 2011) and PySwarms (Miranda 2018) packages are used in this experiment. The hardware used in this experiment is a laptop with Intel Core i7-7700HQ with 16 GB DDR-4 of RAM. Both datasets are split into a training set (80%) and a test set (20%). Twelve replications of the experiment are performed using Dataset 1 with an average running time of around 6 hours, and five replications using Dataset 2 with an average running time of around 12 hours. Parameters are set by conducting preliminary experiments using Dataset 1 before performing the entire experiment.

The BPSO parameters used in this study are obtained from several experiments, which are conducted using the training dataset. The parameters are selected based on the model's accuracy using the training data considered optimal. The parameter of maximum iteration is set to 55, and the number of particles is 25, which is determined based on the literature (Yassin et al., 2012). First,  $c_1$  and  $c_2$  are determined using five replications on a training dataset

in order to see if there is a difference in accuracy between  $c_1 > c_2$  and  $c_1 < c_2$ .

Based on Table 8, the accuracy is higher when  $c_1 > c_2$  with the value of  $c_1 = 1$  and  $c_2 = 0.75$ . This result is then tested using a two-sample t-test to see if there is a statistical difference in accuracy between  $c_1 > c_2$  and  $c_1 < c_2$ . Using  $\alpha = 0.05$ , there is no statistical difference in accuracy. Then,  $c_1$  and  $c_2$  are determined using an identical value of 1. The selection of identical values is also supported by research on the parameter selection of PSO, which stated that if the value of  $c_1$  and  $c_2$  is identical, the fitness value generated by the model would be better (He et al., 2016).

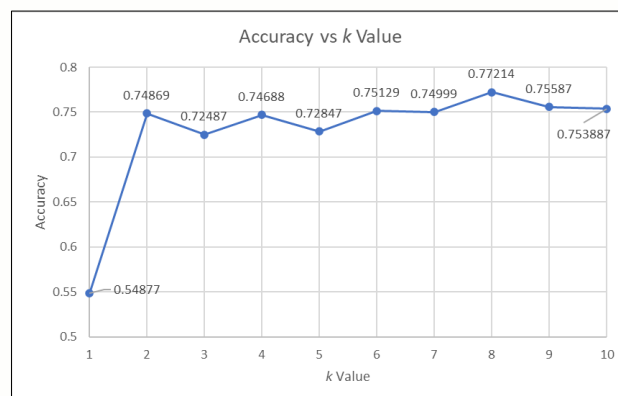
Next,  $w$  is determined using the same method.  $w = 1.5$  and  $w = 0.75$  are tested using the same training dataset used above. The value of  $w = 0.75$  does not converge at 55 iterations of BPSO. The value of  $w = 1.5$  converged early. Based on the experiment above,  $w = 1$  is determined to overcome premature convergence or slow convergence at 55 iterations of BPSO.

**Table 8 Results of testing the value of  $c_1$  and  $c_2$**

Run	Accuracy	
	$c_1 = 1 > c_2 = 0.75$	$c_1 = 0.75 < c_2 = 1$
1.	0.785851	0.776291
2.	0.804971	0.778203
3.	0.820868	0.791587
4.	0.743786	0.799235
5.	0.797323	0.789675

Parameters on k-Nearest Neighbors, the value of k is determined using the same method above. k = 1 to k = 10 is tested using the training dataset. The results are shown in Fig. 2. The parameter k = 8 generates the

highest accuracy of 0.77214 on the training dataset. Then k = 8 is used based on the experiments above.



**Fig. 2 Accuracy based on the value of k.**

Table 9 is a summary of the final BPSO parameters used in this paper. The results are tested using a two-sample t-test to see if there is a statistical difference in accuracy before feature selection and accuracy after feature selection. Using  $\alpha = 0.05$ , there is a statistical difference in accuracy for both datasets (Dataset 1: from 341 features reduced to 229 features; Dataset 2: from 605 features reduced to 389 features). Table 8 is a summary of the results.

Table 10 indicates that despite the number of features decreasing, the accuracy of repurchase intention predictive models constructed are increased on both datasets. For each dataset, the first row corresponds to no feature selection. The second row shows the result of selecting features according to the Pareto (80/20) rule. Finally, the last row displays the result of applying BPSO. The respective number of features and the accuracy are shown for each row. The other performance metrics

**Table 9 Parameters used in this research**

Parameters	Value
$c_1$	1
$c_2$	1
Max. Iteration	55
No. of Particles	25
$k$	8
$w$	1

**Table 10 Model Performance (Accuracy)**

Dataset	Pareto (80/20)	SO-PMI	BPSO	Feature	Accuracy
1	-	-	-	618	0.7591
	✓	-	-	341	0.7744
	✓	✓	✓	<b>229</b>	<b>0.8131</b>
2	-	-	-	2100	0.7137
	✓	-	-	605	0.7788
	✓	✓	✓	<b>389</b>	<b>0.7987</b>

**Table 11 Results Summary**

Dataset	Feature	Accuracy	Precision	Recall	F-Score
1	618	0.7591	0.9548	0.7996	0.8701
	229	0.8131	0.9649	0.8249	0.8892
%	-62.9%	<b>5.40%</b>	<b>1.01%</b>	<b>2.53%</b>	<b>1.91%</b>
2	2100	0.7137	0.9832	0.7881	0.8748
	389	0.7987	0.9706	0.8067	0.8806
%	-81.5%	<b>8.50%</b>	-1.26%	<b>1.86%</b>	<b>0.58%</b>

As a comparison between no feature selection and applying BPSO, Table 11 shows the performance metrics other than accuracy. Based on Table 11, there are improvements in accuracy and almost all other performance metrics. According to the result, the proposed method of BPSO on feature selection constructs a better-performing predictive model of repurchase intention from customer reviews.

Table 12 shows 20 (out of 150) selected features on both datasets. It is worth noting that in this particular research case, features are tokens, and most of the tokens are “emotional” and “preference” words

within a sentence in a review rather than a physical feature of the product itself, such as packaging. These emotional terms could describe customers’ personal preferences and mediate customers’ repurchase intention.

Since this research is based on sociolla.com, Indonesia-based e-commerce, most of the features (tokens) are in Bahasa Indonesia, and many are in informal or standardized forms. For example, the word “Pudar” (dull) and “Mudarin” (to dull). “Mudarin” is an informal form of the word “Memudarkan” (to dull).

**Table 12 Examples of Selected Features**

No.	Dataset 1	Dataset 2
1	<i>efektif</i> (effective)	<i>maksimal</i> (maximum)
2	<i>kental</i> (thick)	<i>ngefek</i> (effective)
3	<i>diskon</i> (discount)	<i>kental</i> (thick)
4	<i>segar</i> (fresh)	<i>diskon</i> (discount)
5	<i>terjangkau</i> (affordable)	<i>segar</i> (fresh)
6	<i>halus</i> (gentle)	<i>murah</i> (cheap)
7	<i>nyaman</i> (mild)	soothing
8	<i>dingin</i> (cool)	<i>lembut</i> (gentle)
9	<i>cerah</i> (glowing)	<i>nyaman</i> (mild)
10	<i>lembab</i> (moisturize)	<i>dingin</i> (cool)
11	<i>aman</i> (safe)	<i>terang</i> (glowing)
12	<i>awet</i> (long lasting)	<i>cantik</i> (beautiful)
13	<i>wangi</i> (fragrant)	<i>lembab</i> (moisturizing)
14	<i>wajib</i> (compulsory)	<i>aman</i> (safe)
15	hydrating	<i>awet</i> (long lasting)
16	<i>suka</i> (like)	<i>baik</i> (good)
17	Ngilangin (Removing)	Instant (Instant)
18	Mulus (Smooth)	Bersih (Clean)
19	Pudar (Dull)	Mudarin (Dull)
20	Manfaat (Benefit)	Simple (Simple)

Based on Table 12 above, the 20 (out of 150) selected features show similarities between the two datasets. These selected features (including “emotional” terms) may provide managerial insights in marketing the products, especially regarding the aspects related to customer repurchase intention.

For example, based on dataset 1, customers tend to repurchase a moisturizer product that is affordable, hydrating, long-lasting, gentle, and mild. These emotional terms describe what customers personally want or need in their moisturizing products.

Although this paper only uses moisturizer products as the main case study, this proposed method is not bound only to moisturizer products. Different product categories could also be analyzed using this method. Different product categories might result in different "emotional" or "preference" words as a feature in the model. Nevertheless, in general, using "emotional" and "preference" words as features in the model could help practitioners and companies gather valuable information regarding their products and predict customers' behavior, especially the repurchase intention behavior. Emotional factors could mediate the influence of shopping characteristics and customers' behavior. Keeping the customer happy by satisfying their personal needs increases the customer's intention

to repurchase (Pappas et al., 2014). Companies may use these emotional appeals in their marketing campaigns or objectives (Ali et al., 2020).

## 5. Conclusions

The result in this paper shows that the proposed Binary Particle Swarm Optimization (BPSO) succeeds in selecting features that generate predictive models with high accuracy. This paper proposes the SO-PMI to sort the features and subsequently encode the particle dimensions according to the sorted values.

The accuracy of a repurchase intention prediction model is improved by selecting only the relevant features and also reducing the computational time (k-NN training time).

This research provides a relatively new idea of a feature selection method for the repurchase intention predictive model. To the best of our knowledge, this is the first attempt to utilize BPSO on feature selection to construct a repurchase intention prediction model that uses SO-PMI to create a meaningful particle encoding, as illustrated in Table 4. Furthermore, this is also the first attempt at utilizing data from an Indonesian e-commerce website with reviews that are mainly written



in Bahasa Indonesia. In 2022, Bahasa Indonesia had approximately 300 million speakers worldwide (Grehenson 2022).

There are ideas that may be realized for future work due to the limitations of this paper. The ideas that could be considered for future work are designing other encodings of the BPSO particles and comparing the performances of various encodings, applying other types of supervised learning algorithms as the classification method (e.g., Naïve Bayes, Decision Tree, or even Deep Learning method), and exploring approaches to finding the more suitable model hyperparameters to construct a better model in BPSO also for the chosen supervised learning algorithm.

## References

- Agrawal, Prachi, Hattan F. Abutarboush, Talari Ganesh, & Ali Wagdy Mohamed. (2009-2019). "Metaheuristic Algorithm on Feature Selection: A Survey of One Decade of Research (2009-2019)." *IEEE Access*, 2021: 26766 - 26791.
- Ali, Amjad, Nazia Abdul Rehman, Khurram Shakir, & Ibrahim Noorani. (2020). "Impact of Emotional Advertisement on Consumer Decision Making for Ice Cream Brands in Pakistan." *International Journal of Management (IJM)*, 2020: 320-338.
- Bing, Xue, Mengjie Zhang, & Will N. Browne. (2013). "Particle Swarm Optimization for Feature Selection in Classification: A Multi-Objective Approach." *IEEE Transactions on Cybernetics* 43, no. 6 (2013): 1656-1671.
- Bins, Jose, & Bruce A. Drapper. (2001). "Feature selection from huge feature sets." *Proceedings Eighth IEEE International Conference on Computer Vision. ICCV 2001. Vancouver, 2001.*
- Bird, S., E. Klein, & E. Loper. (2009). *Natural Language Processing with Python*. O'Reilly Media Inc.
- Brei, Vinicius Andrade. (2020). "Machine Learning in Marketing: Overview, Learning Strategies, Applications, and Future Developments." *Foundations and Trends in Marketing* 14, no. 3 (2020): 173-236.
- Bzdok, Danilo, Martin Krzywinski, & Naomi Altman. (2018). "Machine learning: Supervised methods, SVM and kNN." *Nature Methods* (Nature Publishing Group), 2018: 1-6.
- Cervante, Liam, Bing Xue, Mengjie Zhang, & Lin Shang. (2012). "Binary Particle Swarm Optimisation for Feature Selection: A Filter Based Approach." *IEEE World Congress on Computational Intelligence. Brisbane: IEEE, 2012.*
- Chandra, Grace Nadia. (2021). *Indonesia to Double the Size of Current SE Asia's Digital Economy 2030*. Jakarta: Jakarta Globe, 2021.
- Chen, Yue-Yang. (2012). "Why Do Consumers Go Internet Shopping Again? Understanding the Antecedents of Repurchase Intention." *Journal of Organizational Computing and Electronic Commerce* 22, no. 1 (2012): 38-63.
- Cherrington, Marianne, David Airehrour, Joan Lu, Fadi Thabtah, Qiang Xu, and Samaneh Madanian. (2019). "Particle Swarm Optimization for Feature Selection: A Review of Filter-based Classification to Identify Challenges and Opportunities." *2019 IEEE 10th Annual Information Technology, Electronics and Mobile Communication Conference (IEMCON)*. Vancouver: IEEE, 2019.
- Dang, Shilpa, & Peerzada Hamid Ahmad. (2014). "Text Mining: Techniques and its Application." *International Journal of Engineering & Technology Innovations* 1, no. 4 (2014): 22-25.
- DataReportal. Digital (2021): Indonesia. February 11, 2021. <https://datareportal.com/reports/digital-2021-indonesia> (accessed October 3, 2021).
- Du, Jiahua, Jia Rong, Sandra Michalska, Hua Wang, & Yanchun Zhang. (2019). "Feature selection for helpfulness prediction of online product reviews: An empirical study." *PLoS ONE* 14, no. 12 (2019).
- Forman, George. (2003). "An Extensive Empirical Study of Feature Selection Metrics for Text Classification." *Journal of Machine Learning Research* 3 (2003): 1289-1305.
- Govindarajan, M., & R. Chandrasekaran. (2010). "Evaluation of k-Nearest Neighbor classifier performance for direct marketing." *Expert Systems with Applications* 37, no. 1 (2010): 253-258.

- Grehenson, Gusti. (2022). Universitas Gadjah Mada. May 23, 2022. <https://ugm.ac.id/id/berita/22527-penutur-bahasa-indonesia-capai-300-juta-jiwa>.
- Gunantara, Nyoman, & I Dewa Nyoman Nurweda Putra. (2019). "The Characteristics of Metaheuristic Method in Selection of Path Pairs on Multicriteria Ad Hoc Networks." *Journal of Computer Networks and Communications*, 2019.
- He, Y., W. J. Ma, & J. P. Zhang. (2016). "The Parameters Selection of PSO Algorithm Influencing on Performance of Fault Diagnostics." *MATEC Web of Conference*. Hongkong, 2016.
- Hellier, Phillip K., Gus M. Geursen, Rodney A. Carr, & John A. Rickard. (2003). "Customer repurchase intention: A general structural equation model." *European Journal of Marketing* 37, no. 11/12 (2003): 1762-1800.
- Joulani, Pooria, András György, & Csaba Szepesvári. (2015). "Fast Cross-Validation for Incremental Learning." *International Joint Conference on Artificial Intelligence (IJCAI-2015)*. Buenos Aires, 2015.
- Khanesar, Mojtaba Ahmadi, Mohammad Teshnehlab, & Mahdi Aliyari Shoorehdeli. (2017). "A Novel Binary Particle Swarm Optimization." *15th Mediterranean Conference on Control & Automation*. Athens, 2017.
- Kohavi, Ron, & George H. John. (1997). "Wrapper for feature subset selection." *Artificial Intelligence* 97, no. 1-2 (1997): 273-324.
- Kristiyanti, Dinar Ajeng, & Mochamad Wahyudi. (2017). "Feature selection based on Genetic Algorithm, particle swarm optimization and principal component analysis for opinion mining cosmetic product review." *2017 5th International Conference on Cyber and IT Service Management (CITSM)*. Denpasar: IEEE, 2017.
- Mandong, A., & U. Munir. (2018). "Smartphone Based Activity Recognition using K-Nearest Neighbor Algorithm." *International Conference on Engineering Technologies (INCENTE'18)*. Konya, 2018.
- Mendoza, Enrico C. (2020). "A Study of Online Customers Repurchase Intention Using the 4Rs of Marketing Framework." *International Review of Management and Marketing* 11, no. 2 (2020): 1-10.
- Menghour, Kamilia, & Labiba Souici-Meslati. (2016). "Hybrid ACO-PSO Based Approaches for Feature Selection." *International Journal of Intelligent Engineering & Systems* 9, no. 3 (2016): 65-79.
- Miao, Jianyu, & Lingfeng Niu. (2016). "A Survey on Feature Selection." *Procedia Computer Science* 91 (2016): 919-926.
- Miranda, L. J. V. (2018). "PySwarms: a research toolkit for Particle Swarm Optimization in Python." *Journal of Open Source Software* 3, no. 21 (2018).
- Nedjah, Nadia, Luiza de Maced Mourelle, Janusz Kacprzk, Felipe M.G. Franca, & Alberto Ferreira de Souza. (2009). *Intelligent Text Categorization and Warsaw*: Springer-Verlag Berlin Heidelberg, 2009.
- Negara, Siwage Dharma, & Endang Sri Soesilowati. (2021). "ISEAS - Yusof Ishak Institute." August 2, 2021. [https://www.iseas.edu.sg/wp-content/uploads/2021/07/ISEAS\\_Perspective\\_2021\\_102.pdf](https://www.iseas.edu.sg/wp-content/uploads/2021/07/ISEAS_Perspective_2021_102.pdf) (accessed October 1, 2021).
- Nguyen, Lan, Thu Ha Nguyen, & Thi Khanh Phuong Tan. (2021). "An Empirical Study of Customers' Satisfaction and Repurchase Intention on Online Shopping in Vietnam." *Journal of Asian Finance, Economics and Business* 8, no. 1 (2021): 971-983.
- Pappas, Ilias O., Panos E. Kourouthanassis, Michail N. Giannakos, & Vassilios Chrissikopoulos. (2014). "Shiny happy people buying: the role of emotions on personalized e-shopping." *Electron Markets*, 2014: 193-206.
- Pedregosa, F., et al. "Scikit-learn: Machine Learning in Python." (2011). *Journal of Machine Learning Research* 12, no. 85 (2011): 2825-2830.
- Robbani, H. A. (2018). "PySastrawi 1.2.0." 2018. <https://pypi.org/project/PySastrawi/> (accessed 2020).
- Saeyns, Yvan, Iñaki Inza, & Pedro Larrañaga. (2007). "A review of feature selection techniques in bioinformatics." *Bioinformatics* 23, no. 19 (2007): 2507-2517.

- Shang, Lin, Zhe Zhou, & Xing Liu. (2016). "Particle swarm optimization-based feature selection in sentiment classification." *Soft Computing - A Fusion of Foundations, Methodologies, and Applications* 20, no. 10 (2016): 3821-3834.
- Shroff, Kandarp P., & Hardik H. Maheta.(2015). "A Comparative Study of Various Feature Selection Techniques in High-Dimensional data set to Improve Classification Accuracy." 2015 International Conference on Computer Communication and Informatics (ICCCI -2015). Coimbatore, 2015.
- Suhaily, Lily, & Yasintha Soelasih.(2017). "What Effects Repurchase Intention of Online Shopping." *International Business Research* 10, no. 12 (2017): 113-122.
- Sun, Chensheng, Jiwei Hu, & Kin-Man Lam. (2011). "Feature subset selection for efficient AdaBoost training." 2011 IEEE International Conference on Multimedia and Expo. Barcelona: IEEE, 2011.
- Suryadi, Dedy.(2020). "Predicting Repurchase Intention Using Textual Features of Online Customer Reviews." 2020 International Conference on Data Analytics for Business and Industry: Way Towards a Sustainable Economy. Sakheer: IEEE, 2020.
- Trivedi, Shrawan Kumar, & Mohit Yadav. (2018). "Predicting online repurchase intentions with e-Satisfaction as mediator: a study on Gen Y." *Journal of Information and Knowledge Management Systems* 48, no. 3 (2018): 427-447.
- Tsai, Huei-Ting, Hsin-Cheng Chang, & Ming-Tien Tsai.(2016). "Predicting repurchase intention for online clothing brands in Taiwan: quality disconfirmation, satisfaction, and corporate social responsibility." *Electron Commer Research* 16 (2016): 375-399.
- Turney, Peter D.(2002). "Thumbs Up or Thumbs Down? Semantic Orientation Applied to Unsupervised Classification of Reviews." *Proceeding of the 40th Annual Meeting of the Association for Computational Linguistics (ACL)*. Philadelphia, 2002.
- Vashti, Hannah, & Tony Antonio. (2021). "The Role of Price Promotion and Product Quality in Influencing the Intention to Repurchase Cok-Kis Cookies." *International Conference on Entrepreneurship (ICOEN)*. KnE Social Sciences, 2021. 441-459.
- Xue, Bing, Mengjie Zhang, & Will N. Browne. (2014). "Particle swarm optimisation for feature selection in classification: Novel initialisation and updating mechanisms." *Applied Soft Computing* 18, no. 1 (2014): 261-276.
- Yassin, I. M., M. N. Taib, R. Adnan, M. K. M. Salleh, & M. K. Hamzah. (2012). "Effect of swarm size parameter on Binary Particle Swarm Optimization-based NARX structure selection." 2012 IEEE Symposium on Industrial Electronics and Applications. Bandung, Indonesia, 2012.
- Zhao, M., T. Zhang, & J. Chai. (2015). "Based on SO-PMI Algorithm to Discriminate Sentimental Words 'Polarity in TV Programs' Subjective Evaluation." 8th International Symposium on Computational Intelligence and Design. Hangzhou, 2015.

## AUTHOR BIOGRAPHIES



Dimas Adrianto, is a Master's student at the Industrial Engineering Department, Parahyangan Catholic University, Bandung, Indonesia



Dedy Suryadi, is a faculty member at the Industrial Engineering Department, Parahyangan Catholic University, Bandung, Indonesia. His current research interests include machine learning, Natural Language Processing, and metaheuristics.



# Robustified Principal Component Analysis for Feature Selection in EEG Signal Classification

R. John Martin

Faculty of Computer Science & Information Technology, Jazan University, KSA

\* Corresponding author e-mail : [jmartin@jazanu.edu.sa](mailto:jmartin@jazanu.edu.sa)

(Received 5 January 2022; Final version received 19 December 2022; Accepted 4 February 2023)

## Abstract

Feature engineering is an important step in data analysis, especially for machine learning applications. A wide range of feature selection methods are being used in Electroencephalography (EEG) signal processing applications. Principal Component Analysis (PCA) is considered an ideal method for feature selection whenever high dimensional data is obtained, especially in signal processing applications. Following an examination of various EEG signal processing frameworks, PCA emerged as the winner in the battle to reduce dimensionality. Despite its widespread use, it has been found to be ineffective for EEG signal processing problems like epileptic seizure detection due to the nonlinear nature of the signal properties. Traditional methods for solving PCA are insufficient in this case, so suggest a novel technique. In this paper, PCA is explored with an EEG classification model. The proposed work demonstrates how PCA is robustified for an EEG signal processing scenario by applying kernel functions. Statistical features are extracted from EEG data after preprocessing by the Desecrate Wavelet Transform (DWT). Initially, the classical PCA algorithm is applied for feature selection by reducing the dimensionality. Later, the algorithm is robustified by applying a Gaussian kernel in a nonlinear, high-dimensional feature space. In an EEG classification of epileptic seizure detection, the adoption of robustified PCA outperforms conventional PCA in terms of accuracy.

*Keywords:* PCA, Dimensionality Reduction, Electroencephalography (EEG), Feature Engineering, Signal Processing

## 1. Introduction

Feature dimension is one of the challenging factors in machine learning based signal processing frameworks. Larger the number of features makes harder to process and visualize the data sets and working on it (F. Heydarpour et al., 2020 and Rahiminasab et al., 2020). As most of the features are correlated one another, they may appear redundantly. This is the significance in adopting dimensionality reduction techniques in machine learning frameworks as stated in (R. John Martin, 2018). Wide ranges of feature reduction algorithms are being used for biomedical signal processing applications especially in Electroencephalography (EEG). Principal Component Analysis (PCA) is the commonly used dimension reduction algorithm in EEG based frameworks of epileptic seizure detection. The main characteristic of PCA is to express the data by reducing number of dimensions without much loss in the required data. It is a process of reducing the number of variables under consideration by setting a set of principal variables.

The purpose of PCA is to identify the subset of features in our dataset that best capture information on the entire dataset, allowing us to minimize dimensions with minimal information loss. For example, one can reduce the dimension of training data before feeding it to a ML model for classification to reduce computation time as in (R. John Martin, 2022). High correlation filters, random forests, and backward feature elimination are some of the strategies for dimensionality reduction. PCA effectively handles this problem by determining principal components, which are linear combinations of the original features. These components are extracted in such a way that the first captures the most variance in the dataset; the second collects the remaining variance while staying uncorrelated to the first, and so on.

Using the PCA can result in some information loss if we do not choose the right number of principal components for our data set and its variance. When we apply Principal Component Analysis to our data set, the original features are transformed into principal components: linear combinations of original data features. But which features, variables, or characteristics in the data set are the most significant? After performing the PCA, answering this question can be difficult. The loss of information is caused by nonlinear relationship between the features, which is also supported by a wide range of studies on EEG classification frameworks as stated in table 1. It is essential to keep the significant feature components in the dataset that will play a crucial

role in the classification frameworks. This is the prime motive of this research.

The core objective is to minimize the data loss in the EEG classification models by enhancing the classical PCA. After extensive research, it was discovered that the PCA technique may be utilized to extract the necessary single or multiple EEG feature frequencies from an EEG input. Each signal's characteristic frequency yields just two valid eigenvalues. The number of effective eigenvalues is proportional to the number of raw signal frequencies and has no bearing on the size of signal sub - bands. Hence, the wavelet method of signal sub-band is obtained using DWT and applied to PCA. Now, the major challenging factor in this process is nonlinearity. Abnormalities in a multi-channel EEG must have nonlinear properties, so its signal sub bands must be processed and the significant features retained using nonlinear kernel-based analysis using robustified PCA (Cao, H *et al.*, 2022).

In this attempt, the classical PCA is robustified by using nonlinear kernel (Katayama H *et al.*, 2022) so as to avoid the loss in cumulative EEG signal feature dimension, which will lead to accurate disease diagnosis. The contributions of this research include:

- Study the existing EEG based epileptic seizure detection frameworks using classical PCA as feature selection method
- Propose an enhanced PCA with nonlinear kernel
- Experiment Classical PCA and Nonlinear PCA, and compare their performance

The following sections of this paper reviews and exhibit the use of PCA and its variants in EEG signal processing with the application of epileptic seizure detection frameworks. Section two of this paper provides a comprehensive analysis of dimensionality reduction techniques used with EEG signal processing applications. Section three presents the concepts of conventional PCA and robustified PCA with an experimental framework of epileptic seizure detection. An EEG signal classification is used for validating how best the robustified PCA responds. The outcomes of the experiments are given in section four and the conclusion in section five.

## 2. Related Works

Different approaches are adopted for feature selection processes in EEG signal data analysis. Many works used non-linear statistical methods for reducing the feature dimension. Gajic et al. (2014) adopted scatter matrix method of feature reduction in epileptic seizure detection problem. In an Alzheimer's disease detection problem, Trambaiolli et al. (2017) used eight different algorithms for reducing features. Ozan Kocadagli et al. (2017) reported that they have employed fuzzy relations for reducing the features for epileptic seizure classification.

Ming-ai Li et al. (2016) extracted features using DWT and used an approach called parametric t-Distributed Stochastic Neighbor Embedding (P. t-SNE) for extracting reduced nonlinear features from MI-EEG. Edras Pacola et al. (2017) used Linear Discriminant Analysis (LDA) for obtaining distinctive features for binary classification by reducing the extracted features using wavelets. In a comparative study by Bugli C et al. (2007), Independent Component Analysis (ICA) and PCA were analyzed for efficient event detection. Similarly, Kavita Mahajan et al. (2011) also employed PCA and ICA for dimensionality reduction for their EEG classification problem. For evaluating the performance of various dimensionality reduction techniques, Harikumar et al. (2015) applied PCA, ICA and SVD to a epileptic seizure detection problem. Sharmila et al. (2017) used PCA and LDA to dimension reduction of extracted features using DWT for classification of epileptic EEG. Paulo Amorim et al. (2017) adopted PCA, LDA and ICA to reduce the feature space for an EEG classification problem. Xiao-Wei Wang et al. (2014) used PCA, LDA, and correlation-based feature selector (CFS) for dimensionality reduction in an emotional state classification problem using EEG.

Hadi et al. (2016) inducted Sequential Forward Feature Selection (SFS) algorithm for selection of features and to reduce the dimensionality for classification of epileptic EEG. Elahi et al. (2013) employed two methods such as SFS and LDA for feature reduction in order to maximize classification accuracy. According to Ahmad M. Sarhan (2017), statistical moments are applied in an epileptic seizure detection problem to reduce the dimensionality of input and to choose the features. Wavelet coefficients are used manually to reduce feature dimension after wavelet analysis in the works reported in (Satchidanada Dehuri et al., 2013) and (Benzy V.K. and Jasmin E.A., 2015).

In a multi-channel EEG data analysis by Gopika Gopan et al. (2015), feature reduction is achieved by limiting channel dimension. The extracted features from different domains are reduced by using PCA and Analysis of Variance (ANOVA) methods as reported in (Lina Wang et al., 2017). Similarly, Rajendra Acharya et al. (2012) used PCA for feature dimension reduction and ANOVA for feature selection in a wavelet framework of seizure detection problem.

A recent seizure detection framework of John Martin et al. (2021) used kernel PCA for feature optimization to enhance the classification accuracy and it is claimed that the kernel PCA is working well with SVM for EEG classification. To attain maximum separability extracted features are reduced in dimension using PCA as reported by M Aminion et al. (2010). Similarly, Xie et al. (2014) attain the dimensionality reduction by removing insignificant components using PCA for epileptic EEG classification. In another work by Xie et al. (2011) used multi-scale PCA by combining WT and PCA to obtain reduced features. Noertjahjani et al. (2016) used PCA as an effective feature extraction method for the epileptic EEG classification using SVM.

Roозbeh Z et al. (2017) applied robust feature extraction method by combining PCA and cross-covariance technique (CCOV) in order to reduce the feature dimension of EEG. In an EEG based vigilance estimation problem proposed by Li-Chen Shi et al. (2013), tried three other PCA variants for feature dimension reduction such as L1 norm PCA, sparse PCA and robust PCA along with standard PCA. Williamson et al. (2012) stated that principal components are obtained by reducing extracted features for their SVM classifier. "Table.1" shows the diversified approaches used for feature selection in the recent EEG classification frameworks of seizure detection.

Table 1 Summary of feature selection methods used in EEG classification problems

Reference	Feature Selection Method Adopted
Gajic et al. (2014, 2015)	Scatter Matrix
Ozan Kocadagli et al. (2017)	Fuzzy Relations
Ming-ai Li et al. (2016)	P. t-SNE
Hadi et al. (2016)	SFS
Elahi et al. (2013)	SFS & LDA
Ahmad M. Sarhan (2017)	Statistical moments
Satchidanada et al. (2013)	Wavelet coefficients
Benzy V.K. et al. (2015)	Channel reduction
Gopika Gopan et al. (2015)	LDA
Edras Pacola et al. (2017)	LDA
Kavita Mahajan et al. (2011)	ICA & PCA
Bugli C et al. (2007)	ICA & PCA

Harikumar et al. (2015)	PCA, ICA and SVD
Sharmila et al.(2017)	PCA & LDA
Paulo Amorim et al. (2017)	PCA, LDA & ICA
Xiao-Wei Wang et al. (2014)	PCA, LDA & CFS
Lina Wang et al. (2017) Rajendra Acharya et al. (2012)	PCA & ANOVA
John Martin R et al. (2021) Aminion et al. (2010) Xie et al.(2011, 2014) Noertjahjani et al. (2016) Chunchu R et al.(2014) Manisha Chandani et al..(2017) Sabeti M. et al..(2011) Hashem et al. (2017) Esma Sezer, et al (2012) Harikumar R.et al. (2015) Williamson JR et al. (2012)	PCA
Li-Chen Shi et al. (2017)	L1 norm PCA, sparse PCA and robust PCA

PCA is frequently employed for feature selection in the EEG signal classification frameworks of epileptic seizure detection, according to the referenced literatures. When comparing PCA to one or more alternative feature selection methods such as ICA, SVD, LDA, CFS, and ANOVA [ Harikumar et al. (2015), Kavita Mahajan et al.(2011), Bugli C et al. (2007), Sharmila et al.(2017), Paulo Amorim et al. (2017), Xiao-Wei Wang et al. (2014), Lina Wang et al. (2017) and Rajendra Acharya et al. (2012) ], it is clear that PCA is the best method for reducing feature dimension. According to John Martin R et al. (2021), Hashem et al. (2017), and Xie et al. (2011, 2014), the PCA significantly improves classification performance over other feature selection approaches.

Though PCA would identify the highly significant features in an EEG classification problem, it is critical to maintain every required feature to avoid misclassification, especially in epileptic seizure detection applications. This is the driving force for using kernel approaches to improve the PCA's robustness.

### 3. Methods

Principal Component Analysis (PCA) is a feature reduction method which transforms a high dimensional dataset into a low-dimensional orthogonal feature space while retaining the maximum variance of the original high dimensional dataset. In the framework of EEG classification, PCA is inducted for feature dimension reduction, which will consolidate the most significant feature vectors into one or more principal components.

Initially, wavelet domain feature extraction is materialized using multiscale approximation principle of DWT as stated in (R. John Martin, 2018). Extracted high-dimensional features in DWT are further subjected to analysis in order to obtain compact dimension in size to enable the classification process effective and efficient. The proposed research is carried out in two stages: first, traditional PCA is implemented for EEG classification (Sec. 3.1) , and then, using the kernel function, robustified PCA is developed (Sec. 3.2).

The assumptions that are used to approach PCA for optimum productivity include: *Linearity*: The principal components (PCs) are a linear combination of the original features. PCA may not provide expected results, if this is not true. *Large variance implies more structure*: Variance is an important measure in PCA which indicates how significant a particular dimension is. Hence high variance vectors will be emerged as principal components. *Orthogonality*: In PCA, principal components are considered as orthogonal.

#### 3.1. Feature Dimension Reduction using Classical

##### PCA

Each orthogonal feature vector is referred to as a Principal Component (PC). Eigen values are scalar factors of the degree of variance within the particular PCs. Principal components are graded by their corresponding Eigen values, and accordingly, the first PC captures the most significant variance in the dataset. The second one is perpendicular to the first and gets the next significant variance. The two major steps in PCA include: i) Perform mean normalization and finding the covariance matrix: The mean of the original signal data in all dimensions is first subtracted to produce a data set with a zero mean. Consequently, the covariance matrix is calculated. And ii) Compute eigenvalues and eigenvectors: The covariance matrix decomposition to obtain a matrix of eigenvectors in a n-dimensional space (n PCs) and their corresponding eigenvalues. This will be done with the help of the following algorithm:

Reducing data from n-dimensional to k-dimensional space. Computing the covariance matrix S:

$$S = \sum_{i=1}^n (x_i - m)(x_i - m)^T \quad (1)$$

S is an [n x n] matrix.

Compute eigenvectors and eigenvalues of matrix S

[U, V] = eigs (S), where eigs provides eigenvector. U and V are matrices, where U matrix is an [n x n] matrix, turns out the columns of U are the u



vectors, so to reduce a system from n-dimensions to k-dimensions to take the first k vectors from U (first k columns).

$$U = [u^{(1)} \ u^{(2)} \ \dots \dots u^{(n)}] \in \mathbb{R}$$

It needs to find the way to change 'x' (which is n dimensional) to z (which is k dimensional). Thus reduces the dimensionality.

Take first 'k' columns of the 'u' matrix and stack in columns, where n x k matrix - call this Ureduce

(a) Calculate 'z' as follows,

$$z = (U_{\text{reduce}})^T \times x \quad (2)$$

so,  $[k \times n] * [n \times 1]$  Generates a matrix which is  $k * 1$ .

Features are extracted through DWT based multiresolution analysis (MRA). Conventional method of PCA is materialised by applying features. The feature matrix X is in dimension 300 x 54 (100 samples each from F, N & S segments and each with 54 features). The input feature space is normalized by de-mean the feature matrix. As the first step the covariance matrix of the feature matrix is obtained. The eigenvalues and eigenvectors are then calculated by using covariance matrix. This has been achieved by using the following Matlab code:

$$[\text{coeff}, \text{score}, \text{latent}, \sim, \text{explained}] = \text{pca}(X);$$

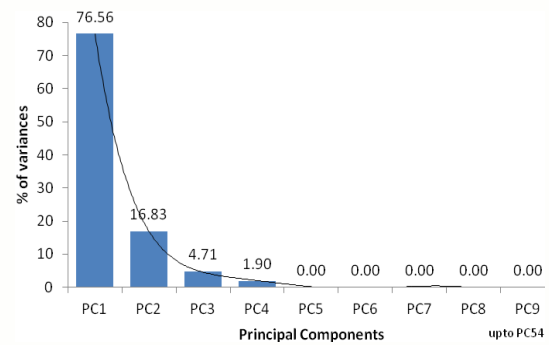
Where, "coeff" are the eigenvectors of the covariance matrix called principal component vectors, "latent" is the output and are the eigenvalues of the covariance matrix. Multiply the original data by the principal component vectors to get the projections of the original data on the principal component vector space. This is also the output "score".

The features now in principal component space with variations specified in a vector "explained" is in "Table 2". The feature variations obtained after conventional PCA is represented by using a scree plot in "Fig. 1". From the scree plot it is noticed that the first 3 principal components (PC1, PC2, and PC3) together explain 98.1% of the variation. Thus the feature dimension is reduced to three and the remaining is considered insignificant.

**Table 2** Vectors in principal component space during classical PCA

Principal Components	Variation
PC1	76.56131087
PC2	16.82612710
PC3	4.71003453
PC4	1.90019686

PC5	0.00232034
PC6	0.00000449
PC7	0.00000420
PC8	0.00000101
PC9	0.00000028
PC10	0.00000019
PC11	0.00000010
PC12	0.00000003
PC13	0.00000001
..... Up to PC54	



**Fig. 1** Scree plot showing percentage of variances among PCs in Classical PCA

### 3.2 Feature Dimension Reduction using Robustified PCA

In high dimensional biomedical data like EEG, PCA is best used for expressing linear variability. But the characteristic of the high dimensional EEG data set is that it has a non-linear nature. In those circumstances PCA cannot determine the variability of data accurately. In order to address this issue of non-linear dimensionality reduction, kernel-based PCA can be recommended. Some improvisations are recommended with the usage of kernel functions for nonlinear mapping so that the principal components are computed efficiently in high dimensional feature spaces.

In general, non-linear methods (Harikumar R et al., 2015) are being applied to robustify the classical PCA. The extended form of a classical PCA is called Kernel Principal Component Analysis (Chenouri S et al., 2015; Schölkopf Bernhard et al., 1998) by adopting kernel methods. Some innovative approaches applied towards classical PCA which may enhance dimension reduction process are termed as robust PCA.

The linear transformation of PCA functionalities are carried out in a reproducing kernel Hilbert space with a nonlinear mapping. In kernel-based method, the mapping carried out by Kernel PCA depends on the choice of the kernel function K, probably it may include the linear kernel; and the nonlinear kernel functions such

as the polynomial kernel and the Gaussian kernel. In this method, principal components are computed efficiently in a high-dimensional feature spaces that are related to the input space by some nonlinear mapping.

Kernel PCA chooses the principal components which are nonlinearly related to the input space by performing PCA in the high dimensional input space obtained through nonlinear mapping, where the low-dimensional latent structure is, expected to be found easily.

Consider a feature space  $\Phi$  such that:

$$x \rightarrow \Phi(x) \quad (3)$$

Let's suppose  $\sum_i^t \Phi(x_i) = 0$ ; it will formulate the kernel PCA objective function as follows:

$$\min \sum_i^t \|\Phi(x_i) - U_q U_q^t \Phi(x_i)\| \quad (4)$$

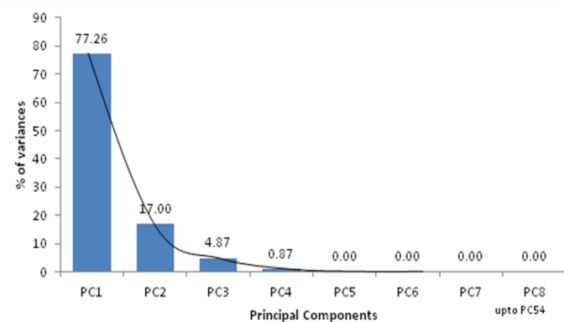
Where  $U$  represents the eigenvectors of  $\Phi(X)\Phi(X)^T$ . Note that if  $\Phi(X)$  is  $n \times t$  and the dimensionality of the feature space  $n$  is large, then  $U$  is  $n \times n$  which will make PCA impractical.

In order to reduce the dependence on  $n$ , it is assumed that a kernel  $K(\cdot, \cdot)$  will compute  $K(x,y) = \Phi(x)^T \Phi(y)$ . Given such a function, compute the matrix  $\Phi(X)^T \Phi(X) = K$  efficiently, without computing  $\Phi(X)$  explicitly. Significantly,  $K$  is  $t \times t$  here and does not depend on  $n$ . Thus, it can be computed in a run time that depends only on  $t$ . And also, it is observed that the PCA can be formulated fully in terms of dot products between data points. Replacing dot products by kernel function  $K$ , which is in fact equivalent to the inner product of a Hilbert space yields to the Kernel PCA algorithm. In order to attain optimum classifier performance in this proposed model, Gaussian kernel is inducted to robustify the conventional PCA.

On implementation of robustified PCA using Gaussian kernel function, the features of the input data is mapped into the principal components space. The variations of the principal components expressed in a vector "explained" are given in "Table 3". Observing the concentrated principal components in PC1, PC2, and PC3 obtained from robustified PCA, it is clear that the three principal components mentioned above can identify 99.13 percent of the variations in the input data. This is 1.03% ahead of the classical PCA. The scree plot in "Fig.2" illustrates the concentrations in the principal components of robustified PCA.

**Table 3:** Vectors in principal component space during robustified PCA

Principal Component	Variation
PC1	77.25931087
PC2	16.99712710
PC3	4.86903453
PC4	0.87319686
PC5	0.00132034
PC6	0.00000449
PC7	0.00000420
PC8	0.00000101
PC9	0.00000028
PC10	0.00000019
PC11	0.00000010
PC12	0.00000003
PC13	0.00000001
.....	Upto PC54



**Fig 2:** Scree plot showing percentage of variances among PCs in Robustified PCA

#### 4. Results and Discussions

The major objective of this research was to improve the classical PCA to avoid the loss of important feature dimensions which are contributing towards accurate classification. The proposed method, two variants namely Classic PCA and robustified PCA are implemented towards the EEG classification problem.

In the first phase of the research, classical PCA is adopted to identify most significant features which are concentrated in the principal components. The percentage of variances in the data set concentrated in the principal components PC1, PC2 and PC3 together as 98.1% as represented in table 1 and figure 1, which means 1.9 % of the feature properties are remain with the rest of the principal components. Eliminating remaining 1.9% of the feature properties may lead to misclassification.

As a second experiment, robustified PCA is implemented with the dataset. While looking at the percentage of variances in scree plots, it's evident that robustified PCA can explain 99.13 percent of the

features, which is 1.03 percent ahead of classical PCA in the first phase. This is a clear indication of the enhanced performance of the robustified PCA over classical PCA. Thus, it is presumed that this enhancement will lead to perfect classification of EEG signals.

In order to verify this enhancement in an EEG classification framework, the reduced feature matrix is prepared as training and test sets from the pool of 300 signal feature inputs representing ictal and interictal EEG samples of different subjects named Z, O, N, F and S (R. John Martin et al, 2022). The most significant features concentrated in the first three principal components (PC1, PC2 and PC3) are applied to the classifier for epileptic EEG detection on both scenarios. The SVM nonlinear Polynomial kernel-based classifier is used to classify the signal inputs on two subjects namely seizure (ictal) and seizure-free (interictal). In order to perform a 5-fold cross validation, 5 sets of training and corresponding test samples are prepared from the reduced feature matrix.

"Table.4a" demonstrates the performance of the classifier on the selected feature dimensions using classical PCA. It should be noted that the SVM-based classifier demonstrated 98.9% accuracy, implying that the classifier may exhibit 1.1% error in EEG signal classification, which is a cause for concern in disease diagnostics.

Subsequently, the classification model is used with robustified PCA using its three principal components PC1, PC2 and PC3. It is observed that the robustified PCA using Gaussian kernel is doing better in the EEG classification framework of epileptic seizure detection which is 0.7% ahead of classical PCA as stated in "Table 4b". This clearly shows that by identifying the most important EEG signal feature properties using the first three principal components, the robustified PCA significantly improves classification performance, resulting in accurate disease diagnosis.

**Table 4: a)** Classifier Performance with Features selected using Classical PCA

Classifier	Kernel Parameters	5-fold Cross Validation		
		SEN	SPE	ACC
SVM-Polynomial Kernel	d=2	0.937	0.967	0.938
	d=5	0.965	0.996	0.989

**Table 4: b)** Classifier Performance with Features selected using Robustified PCA

Classifier	Kernel Parameters	5-fold Cross Validation		
		SEN	SPE	ACC
SVM-Polynomial Kernel	d=2	0.937	0.967	0.938
	d=5	0.965	0.996	0.989

SVM-Polynomial Kernel	d=2	0.927	0.989	0.975
	d=5	0.989	0.994	0.996

## 5. Conclusion

In this paper, a methodology has been proposed to enhance the classical PCA in EEG classification frameworks. The article began by reviewing the literature on EEG signal classifications for epileptic seizure detection utilizing classical PCA as a feature selector. As it has been mentioned, the majority of authors employed traditional PCA in their frameworks, who established that the EEG machine learning frameworks responded well to PCA combinations with only mediocre accuracy. Though PCA is a popular approach for reducing feature dimensions when there are a large number of features in classification problems, its performance is questionable when there is a nonlinear relationship between the data variables. This was the inspiration for the proposed research to enhance the classic PCA by incorporating a nonlinear kernel. Initially the classical PCA is experimented and tested with SVM based nonlinear classifier. After that, the classical PCA is enhanced with a Gaussian kernel, implemented and tested with EEG signal classification. On comparing the feature variations with selected principal components, it is noted that the kernelized PCA performed better. Thus, the classical PCA is enhanced. The EEG classification model performed better than classical PCA when the reduced features from robustified PCA were applied. As a result, the proposed PCA enhancement significantly improves disease diagnosis by eliminating misclassification of EEG signals. Furthermore, this research experiment yields significant outcomes that will be beneficial for future signal processing researchers.

## References

- Ahmad, Mohammad, Sarhan. (2017). A Low Complexity Algorithm for Epileptic Seizure Detection using Statistical Moments and Support Vector Machines. *Biomedical Letters*, 3(2), 79-86.
- Aminian, Masoud., Aminian, F., Schetino, L., and Ameli, A. (2010). Electroencephalogram (EEG) signal classification using neural networks with wavelet packet analysis, principal component analysis and data normalization as preprocessors. *Proceedings of the 21st Midwest Artificial Intelligence and Cognitive Science Conference, MAICS 2010*, 55-62.
- Benzy, V, K., and Jasmin, E, A. (2015). A Combined Wavelet and Neural Network Based Model for

- Classifying Depth of Anaesthesia. *Procedia Computer Science*, 46, 1610-1617.
- Bugli, C., and Lambert, P. (2007). Comparison between Principal Component Analysis and Independent Component Analysis in Electroencephalograms Modeling. *Biometrical Journal*, 49(2), 312-327.
- Cao, H., Wang, G., Sun, J., Deng, F., & Chen, J. (2022). Deep Contrastive Principal Component Analysis Adaptive to Nonlinear Data. *IEEE Transactions on Signal Processing*, 70, 5738–5750. <https://doi.org/10.1109/tsp.2022.3224647>
- Chenouri, S., Liang, J., and Small, C.G. (2015). Robust dimension reduction. *WIREs Comput Stat.*, 7, 63-69.
- Chunchu, Rambabu., and B, Rama, Murthy. (2014). EEG Signal with Feature Extraction using SVM and ICA Classifiers. *International Journal of Computer Applications*, 85(3), 0975 – 8887.
- Edras., Pacola., Veronica Quandt., Paulo Liberalesso., Sérgio Pichorim., Fábio, Schneider., and Humberto Gamba. (2017). A versatile EEG spike detector with multivariate matrix of features based on the linear discriminant analysis, combined wavelets, and descriptors. *Pattern Recognition Letters*, 86, 31-37.
- Elahi, Z., Boostani, R., and Motie, Nasrabadi, A. (2013). Estimation of hypnosis susceptibility based on electroencephalogram signal features. *Scientia Iranica*, 20(3), 730–737.
- Esma, Sezer., Isik, Hakan., and Saracoğlu, Esra. (2012). Employment and Comparison of Different Artificial Neural Networks for Epilepsy Diagnosis from EEG Signals. *Journal of Medical Systems*, 36,(1). 347–362.
- Gajic, D., Djurovic, Z., Di, Gennaro, S., and Fredrik, Gustafsson. (2014). Classification of EEG signals for detection of epileptic seizures based on wavelets and statistical pattern recognition. *Biomedical Engineering: Applications, Basis and Communications*, 26(2).
- Gajic, D., Djurovic, Z., Gligonjevic, J., Gennaro, S, D., Gajic, I, S. (2015). Detection of Epileptiform Activity in EEG Signals Based on Time-Frequency And Non-Linear Analysis. *Front. Comput. Neurosci.*, 9, 38.
- Ghayab, HRA., Li, Y., Abdulla, S., Diyk, M., and Wan, X. (2016). Classification Of Epileptic EEG Signals Based On Simple Random Sampling And Sequential Feature Selection. *Brain Informatics*, 3(2), 85-91.
- Gopika, Gopan, K., Neelam, Sinha., and Dinesh, Babu, J. (2015). EEG Signal Classification In Non-Linear Framework With Filtered Training Data. *IEEE 23rd European Signal Processing Conference (EUSIPCO)*, 624 – 628.
- Harikumar, R., and Sunil, Kumar, P. (2015). Dimensionality Reduction Techniques for Processing Epileptic Encephalographic Signals. *Biomedical & Pharmacology Journal*, 8(1), 103-106.
- Harikumar, R., and Sunil, Kumar, P. (2015). Principal Component Analysis as a Dimensionality Reduction Technique and Sparse Representation Classifier as a Post Classifier for the Classification of Epilepsy Risk Levels from EEG Signals. *J. Pharm. Sci. & Res.*, 7(6), 282-284.
- Hashem, Kalbkhani., and Mahrokh, G, Shayesteh (2017). Stockwell transform for epileptic seizure detection from EEG signals. *Biomedical Signal Processing and Control.*, 38, 108–118.
- Heydarpour, F., Abbasi, E., Ebadi, M. J., & Karbassi, S. M. (2020). Solving an Optimal Control Problem of Cancer Treatment by Artificial Neural Networks. *International Journal of Interactive Multimedia and Artificial Intelligence*, 6(4), 18. <https://doi.org/10.9781/ijimai.2020.11.011>
- Jihun, Ham., Daniel, D, Lee., Sebastian, Mika., and Bernhard, Schölkopf. (2004). A kernel view of the dimensionality reduction of manifolds. *In Proceedings of the twenty-first international conference on Machine learning (ICML '04), ACM, NY, USA*, 47.
- John, Martin, R., and Swapna, S, L. (2022). A Machine Learning Framework for Epileptic Seizure Detection by Analyzing EEG Signals. *Int. J. Com. Dig. Sys.*, 11(1), 1383-1391.
- John, Martin, R., Sujatha, S., and Swapna, S,L. (2018). Multiresolution Analysis in EEG Signal Feature Engineering for Epileptic Seizure Detection. *International Journal of Computer Applications*, 180(17), 14-20.
- John, Martin, R., Swapna, S,L., and Sujatha, S. (2018). Adopting Machine Learning Models for Data Analytics-A Technical Note. *International Journal of Computer Sciences and Engineering*, 6(10), 360-365.
- John, Martin, R., Uttam, Sharma., Kiranjeet, Kaur., Noor Mohammed, Kadhim., Madonna, Lamin., Collins Sam, Ayipeh. (2022). Multi-Dimensional CNN Based Deep Segmentation Method for Tumor Identification. *BioMed Research International*. Hindawi. Article ID 5061112. Vol. 2022.
- Katayama, H., Mori, Y., & Kuroda, M. (2022). Variable Selection in Nonlinear Principal Component Analysis. *Advances in Principal Component Analysis*. <https://doi.org/10.5772/intechopen.103758>.
- Kavita, Mahajan., M, R, Vargantwar., and Sangita, M, Rajput.. (2011). Classification of EEG using PCA, ICA and Neural Network. *International Conference*



- in *Computational Intelligence (ICCIA). Proceedings published in International Journal of Computer Applications® (IJCA)*.
- Li-Chen, Shi., Ruo-Nan, Duan., and Bao-Liang, Lu. (2013). A Robust Principal Component Analysis Algorithm for EEG-Based Vigilance Estimation. *Conf. Proc IEEE Eng Med Biol Soc.*, 6623-6626.
- Manisha, Chandani., and Arun, Kumar. (2017). Classification of EEG Physiological Signal for the Detection of Epileptic Seizure by Using DWT Feature Extraction and Neural Network. *American Journal of Information Management*, 2(3), 37-42.
- Ming-ai, Li., Xin-yong, Luo., and Jin-fu, Yang. (2016). Extracting the nonlinear features of motor imagery EEG using parametric t-SNE. *Neurocomputing*, 218, C, 371-381.
- Ozan, Kocadagl.,i and Reza, Langari.. (2017). Classification of EEG signals for epileptic seizures using hybrid artificial neural networks based wavelet transforms and fuzzy relations. *Expert Systems with Applications*, 88, 419-434.
- Paulo, Amorim., Thiago, Moraes., Dalton, Fazanaro., Jorge Silva., and Helio, Pedrini. (2017). Electroencephalogram signal classification based on shearlet and contourlet transforms. *Expert Systems with Applications*, 67, 140-147.
- Rahiminasab, Atefeh, Peyman Tirandazi, M. J. Ebadi, Ali Ahmadian, and Mehdi Salimi. 2020. An Energy-Aware Method for Selecting Cluster Heads in Wireless Sensor Networks. *Applied Sciences*. 10(21), 7886.
- Rajendra, Acharya, U., Vinitha, Sree, S., Ang, Peng, Chuan, Alvin., and Jasjit, S, Suri. (2012). Use of principal Component Analysis for automatic classification of epileptic EEG activities in wavelet framework. *Expert Systems with Applications*, 39, 9072-9078.
- Sabeti, M., Katebi, S, D., R, Boostani., and G, W, Price. (2011). A new approach for EEG signal classification of schizophrenic and control participants. *Expert Systems with Applications*, 38, 2063-2071.
- Satchidanada, Dehuri., Alok, Kumar, Jagadev., and Sung-Bae, Cho. (2013). Epileptic Seizure Identification from Electroencephalography Signal Using DE-RBFNs Ensemble. *Procedia Computer Science*, 23, 84-95.
- Schölkopf, Bernhard., Smola, Alex., and Müller, Klaus-Robert. (1998). Nonlinear component analysis as a kernel eigenvalue problem. *Neural Comput.*, 10, 1299-1319.
- Sharmila., and Mahalakshmi, P. (2017). Wavelet-based feature extraction for classification of epileptic seizure EEG signal. *Journal of Medical Engineering & Technology*, 41(8).
- Siswandari, Noertjahjani., Adhi, Susanto., Risanuri, Hidayat., and Samekto, Wibowo. (2016). Ictal Epilepsy And Normal EEG Feature Extraction Based On PCA, KNN And SVM Classification. *Journal of Theoretical and Applied Information Technology*, 83(1).
- Trambaiolli, L, R., N. Spolaôr., A, C, Lorena., R. Anghinah., J, R, Sato. (2017). Feature selection before EEG classification supports the diagnosis of Alzheimer's disease. *Clinical Neurophysiology*, 128(10), 2058-2067.
- Wang, L., Xue, W., Li, Y., Luo, M., Huang, J., Cui, W., and Huang, C. (2017). Automatic Epileptic Seizure Detection in EEG Signals Using Multi-Domain Feature Extraction and Nonlinear Analysis. *Entropy*, 19(6).
- Williamson, JR., Bliss, DW., Browne, DW., and Narayanan, JT. (2012). Seizure prediction using EEG spatiotemporal correlation structure. *Epilepsy Behav.*, 25(2), 230-238.
- Xiao-Wei Wang., Dan Nie., and Bao-Liang Lu. (2014). Emotional state classification from EEG data using machine learning approach. *Neurocomputing*, 129, 94-106.
- Xie, S., and Krishnan, S. (2011). Signal decomposition by multi-scale PCA and its applications to long-term EEG signal classification. *The 2011 IEEE/ICME International Conference on Complex Medical Engineering, Harbin Heilongjiang*, 532-537.
- Xie, S., and Krishnan, S. (2014). Dynamic Principal Component Analysis with Nonoverlapping Moving Window and Its Applications to Epileptic EEG Classification. *Hindawi Scientific World Journal*, 2014(6), 419308.
- Zarei, R., He J., Siuly, S., and Zhang, Y. (2017). A PCA Aided Cross-Covariance Scheme For Discriminative Feature Extraction From EEG Signals. *Comput Methods Programs Biomed*, 146, 47-57.



Dr. John Martin has over 25 years of experience as an academic in the field of Computer Science. Dr. John Martin earned his Ph.D. in Computer Science from Bharathiar University in India and specialized in Machine Learning with an application to biomedical data analytics. He has

vast experience in higher education as an educator and administrator in India and the Middle East. Currently, he is working in the School of Computer Science and Information Technology at *Jazan University* (Ministry of Education), KSA. He has published extensively in the fields of machine intelligence and biomedical data analytics and has served as editor and reviewer for refereed journals. His research was patented both nationally and internationally. His accomplishments as an educator, mentor, author, researcher, adjudicator and consultant are acknowledged by the global community. His research interests include Machine Intelligence, Signal Processing and Healthcare Data Analytics.

## A Novel Underwater Packet Scheduling based on Modified Priority Backpressure and Peak Age of Information approach

A Caroline Mary<sup>1</sup>, A V Senthil Kumar<sup>2</sup>, H R Chennamma<sup>3</sup>

<sup>1</sup> Research Scholar, Department of Computer Science, Hindusthan College of Arts and Science, Coimbatore, India, [marypsgtech@gmail.com](mailto:marypsgtech@gmail.com)

<sup>2</sup> Professor and Director, Department of MCA, Hindusthan College of Arts and Science, Coimbatore India, [avsenthilkumar@yahoo.com](mailto:avsenthilkumar@yahoo.com)

<sup>3</sup> Department of Computer Applications, Sri Jayachamarajendra College of Engineering, India, [anuamruthesh@gmail.com](mailto:anuamruthesh@gmail.com)

\* Corresponding author E-mail: [marypsgtech@gmail.com](mailto:marypsgtech@gmail.com)

(Received 11 November 2022; Final version received 24 January 2023; Accepted 9 March 2023)

### Abstract

Objective: To develop an effective underwater packet scheduling algorithm considering the freshness of information. Underwater Communication has gained interest in recent days and Energy consumption, Freshness of information is considered to be an important metric. Applications include defense-based applications, environmental monitoring, pollution detection, and ocean study. This paper works on the Age of Information-based concepts to ensure that information's freshness is sustained. Peak AoI is used as an important metric to assign priority to the packets resulting in less packet delay and loss. Congestion is reduced as it works on the principle of backlog. And normalized class-based AoI helps in giving importance according to the importance of packets. As the value of information reduces over time, the proposed technique helps to maintain freshness. By varying the number of nodes and speed of nodes simulation results are shown, and delay, throughput, packet delivery ratio, and energy consumed are calculated. Energy consumed is almost reduced to two third and as Information freshness is given importance, packet loss and delay are reduced.

*Keywords:* Backpressure, Peak AoI, Priority, Underwater

## 1. Introduction

For a better understanding of the sea world, underwater communication is essential. Increasing demands for real-time status update has led to Age of Information (AOI) -based scheduling. Applications like Tsunami based data, underwater monitoring, etc. demand the Age of Information to be less with less packet delay and loss. Finding out the minimum achievable AoI is also an important question (Jaya et al. (2021)). The time elapsed since the generation of the latest successfully received information is considered an AoI measure (Roy D. Yates et al.(2021)). The difference between the current time and the time when the last received information was

generated is  $t - \tau$ . To be a minimum AOI, packets must be transmitted often, and delays in packets have to be reduced (Xingran Chen et al. (2022)).

Estimation error might increase if there is a delay in packets. By minimizing the Age of Information, errors can be reduced to a great extent (X. Zhang et al. (2021)) (X. Chen, X. Liao & S. Saedi-Bidokhti (2021)). Pre-processed data is preferred to make sure that the transmission is easier. But it might affect the freshness of information (C. Xu et al. (2019)).

Peak AOI is a factor considered in the Age of Information Concept. Concepts are based on replacing the old packet when a new packet arrives. This can be helpful to minimize the age of the information to some extent so that the freshness of information can be increased.

Several scheduling algorithms are available that help to decrease packet loss and increase throughput etc. One such algorithm is Backpressure Algorithm (BPA), which is apt for underwater-based communication as it helps to reduce congestion to a greater extent. BPA works on the principle of backlog. Hence, packet loss can be reduced and the timeliness of information can be taken care of. The problem with the BPA is that the smaller queue might be made to suffer, so special consideration to the priority queue is a must. The backpressure algorithm takes care of traffic and congestion control. The priority queue concept will help manage real-time packets so that important packets will not suffer packet delay.

To avoid smaller queues from getting affected, the Queue Length Stabilizer technique is used. After a timespan, the number of packets serviced is noted, and the average is found. If a queue has a smaller number of packets scheduled, the queue size is increased virtually. This might help in the smaller queue to be scheduled too.

Active priority is helpful, as it's not static like most of the priority algorithms. The priority value is based on its priority, TTL, and delay. As the information freshness parameter is important, the AoI-based metric is also used to find the packet's priority. The Metric used here is Peak AoI. Peak AoI represents the worst-case AoI. It is the maximum time elapsed since the preceding piece of information was generated.

Stale information is generally not needed to be transmitted. Here, the source node manages to discard the packets that are not needed. A peak age metric is suggested, which can help know the max-age value before an update.

Underwater Pragmatic Routing Approach through Packet Reverberation mechanism (UPRA-PR) (SHAHZAD ASHRAF et al. (2020)). Considers the thorp propagation mechanism for rummaging and rejection of unavoidable noises and allows only litigate packets to minimize the route failure probability.

In our proposed method, packet loss, delay, and Age of Information are controlled, thereby minimizing energy usage.

The adaptive traffic control algorithm works based on the BPA. This helps in reducing congestion. It is based on accurate real-time traffic information and global traffic information. Results show that it decreases the average vehicle traveling time (Arnan Maipradit et al. (2021)).

In our proposed method, a modified Backpressure technique is suggested with QLS, so even shorter queues are also not affected and congestion is also in control.

A low-complexity algorithm, which helps in minimizing AoI is suggested by (Igor Kadota et al. (2019)). A randomized policy, MaxWeight Policy, Drift-plus-Penalty, and Whittle Index policy are suggested and algorithms are simulated. It is based on reducing undesirable states. Lyapunov optimization.

In our proposed method, with the help of the undesirable effect of Peak AoI, a threshold is created and active priority is assigned based on it.

The query age of information (QAoI) metric, an adaptation of the AoI concept for pull-based scenarios, is considered in (Tahir Kerem et al. (2022)).

To avoid failures, separating loads dynamically and allocating them to be scheduled is proposed by (Reshma Sultana S et al. (2020)). Priorities based on deadlines are suggested. Consideration of the value of information is not done.

Our proposed method works on Active priority with consideration to TTL, its Priority of it, Freshness of Information.

Priority-based Edge scheduling algorithms are proposed in (Arkadiusz Madej, Nan Wang, et al. (2020)). Three levels of priority 50%, 35%, and 15% are assigned, thereby low-priority process doesn't starve.

Our proposed method works on the principle of Age gain and QLS thereby avoiding starvation and packet loss.

In priority-based applications, measures are not taken to increase the freshness of information. Here, in the proposed method, a novel way of Peak AoI is used as a threshold to gain momentum.

The organization of the paper is as follows,

To avoid traffic congestion, a backlog-based technique is used to find the optimal commodity,

PeakAoI is calculated which can be used to deal with the freshness of Information.

PeakAoI value is normalized as different packets have different importance with respect to information freshness.

To the Active priority scheduling algorithm, PeakAoI is added as one of the metrics to denote the priority of the packet to be scheduled and packets are scheduled accordingly.

The proposed method helps in keeping the time-critical information fresh using a normalized PeakAoI approach and the active priority scheduling algorithm helps

in less packet loss, energy, and delay which makes it suitable for real-time underwater-based applications.

## 2. Methodology

### 2.1. Age of information (AoI)

The freshness of information is an important metric to be addressed. AoI is a new metric used to denote information's freshness. In real-time-based applications, the latest state information only matters. In health-based applications, natural calamities-based applications, etc. timeliness of the information plays a major role. The status updates packet comes with a timestamp to denote the time at which the packet was generated. The old information seems to have no value, once the new information comes in. When the receiver is interested in the timeliness of information, the Age of Information has its importance.

$$\Delta(t) = t - U(t) \quad (1)$$

Age of Information is the time elapsed since the last received packet was generated.

Packets are selected such that, minimization of age happens. Usually, fixed threshold or adaptive threshold measures are used.

Peak AoI

AoI measures are generally based on the frequency of generation of packets too, and not just the delay of the packet. Peak AoI metric represents the worst-case AoI. It denotes the maximum time elapsed since the previous information was generated.

Delay time can be denoted as  $T_i = r_i - t_i$ . (Basel Barakat et al. (2019))

$$(2)$$

$t_i$  => Time at which the status was updated or the packet was generated at the source

$r_i$  => Denotes the time at which the packet reached the destination.

$X_i$  => Denotes the time between the generation of updates.

$$X_i = t_i - t_{i-1} \quad (3)$$

Peak Age of Information value of the particular update =>  $PT = (1/n - 1) \sum_{i=1}^{n-1} X_i + T_i$  (Basel Barakat et al. (2019))

$$(4)$$

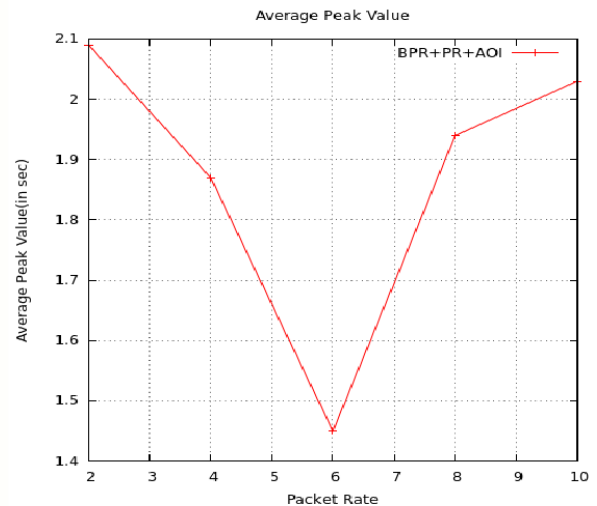


Fig 1. Denotes the average peak AOI value. AoI metrics are based on Packet rates too. So, the packet rate is varied and the Average PeakAoI is calculated as shown in Fig. 1.

### 2.2. Congestion control with backpressure algorithm

The backpressure algorithm is effective in scheduling packets. Especially it helps in avoiding traffic in networks. The backpressure algorithm works on the principle of Backlog. The optimal commodity is found based on the backlog (Zhuo Lu, Member, IEEE, Yalin E. Sagduyu, Senior Member, IEEE & Jason H. Li, Member, IEEE (2016)).  $Q_a^{(c)}(t) - Q_b^{(c)}(t)$ .

Backpressure algorithms can be used to Multi commodity Networks. Based on the optimal commodity, a flow is selected for transmission. It provides high throughput. But there might be delays, especially in the case of smaller queues.

$W_{ab}(t) = \max [ Q_a(t) - Q_b(t) ]$  ( Zhuo Lu et al.(2016)) => Weight calculation (5)

Transmission matrices are created and the link is selected.

The problem with the Backpressure Technique is that the smaller queue might suffer. It might be difficult for the smaller queues to be scheduled before the deadlines or before facing issues like packet delay.

### 2.3. AoI with priority

Our ultimate goal is to ensure that real-time and non-real-time packets are scheduled effectively. To achieve this queue length and stabilizer queues can be used. Where, if a particular queue is unattended based on a threshold value, the queue size can be virtually increased, so that the queue might get a chance to be scheduled.



Dynamic priority is considered to be useful in the real-time congestion of the network (Bo Hu et al. (2020)).

Real-time packets or high-priority packets must be made sure to reach the destination without packet loss or delay. So, it's better to include a priority queue, where real-time packets are given first preference to be scheduled. When added to the Backpressure technique, this solution is beneficial as important packets never suffer. Based on TTL and delay, the priority can be updated to make sure that Packet loss and delay are less. This active priority technique adds strength to the proposed algorithm.

AoI measures help to make sure that the freshness of information is sustained. In addition to priority technique parameters like TTL, and Delay, the AoI metric namely Peak AoI is also considered to determine the priority of packets.

When the rate of communication varies, the Age of Information varies too. When the Rate increases, AoI decreases based on the resources available too. A threshold value is set based on the value obtained from the PeakAoI. If the threshold is met, the priority of the packet is increased from the current level, as the packet might turn out to be less valuable because of the timeliness.

When the new packet arrives, and if the old packet is still in transmission, the old packet is dropped as it does not have any value compared to the new packet.

Simulation results prove that packet loss, delay, energy, routing overhead, etc are reduced and the packet delivery ratio, throughput, etc. have been improved using the above method suggested.

---

**Algorithm 1** PeakAoI

1. Using the priority of the packet, TTL, and Delay constraint assign the priority of Packets.
  2. Calculate PeakAoI.
  3. Calculate the threshold value based on PeakAoI obtained.
  4. Using the Threshold value, make changes to the priority of packets if needed.
- 

## 2.4 Class-based normalized PAoI

Based on the priority of packets, TTL, and delay allocate classes for the packet. Let us consider that there are 3 classes of Priority, High, Medium, and Low. Assign High priority packets to 10 and 20,30 for medium and low-priority packets.

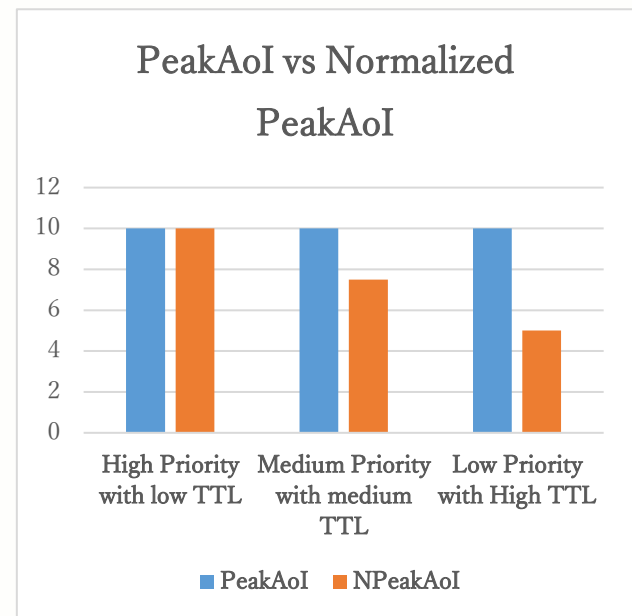
By considering priority and TTL, have 3 classes. 1. High Priority with less TTL, 2. Medium Priority with medium TTL, 3. Low priority with high TTL.

AoI metric is used to measure the freshness of information. But the freshness of information will have its importance according to the information, so this class-based PAoI will help find the exact or needed PAoI.

**Table 1.** Normalized PAoI

	PAoI value	Normalized PAoI
High Priority with less TTL	10	10
Medium priority with medium TTL	10	7.5
Low Priority with high TTL	10	5

When a scheduling algorithm has to schedule the packets based on AoI, this normalized PAoI will better solve the purpose of the freshness of information according to the packet. As every packet are different, their values over time will be different too. Important packets will lose their value more, over time when compared to low-priority packets. Keeping this in mind the PAoI value is normalized accordingly. This method can help keep emergency and most critical application data fresh.



**Fig 2.** Normalized PeakAoI

Normalized PeakAoI helps in giving importance to time-critical packets where Information freshness is vital.

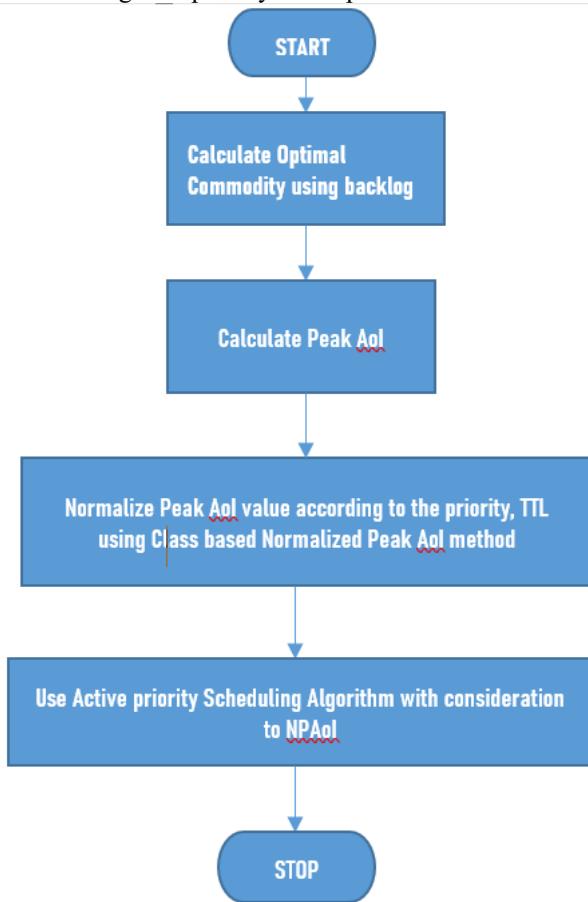
## 2.5. Modified priority backpressure algorithm with consideration to real-time packets and AOI

**Algorithm 2** Modified Priority

Assign priority of packet based on the packet's importance

If TTL < then

Upgrade the priority of the packet to the next level  
 End if  
 Use Backpressure Algorithm to schedule the packets by calculating the optimal commodity and weight  
 Make sure that smaller queues are not affected by larger queues by adjusting the queue size  
 Calculate PeakAoI  
 Normalize the obtained PeakAoI value according to the importance of the packets  
 Calculate the threshold value  
 If AoI  $\geq$  threshold value of Normalized Peak AoI  
 Reassign the priority of the packet End if



**Fig 3.** Flowchart representing the Proposed Algorithm

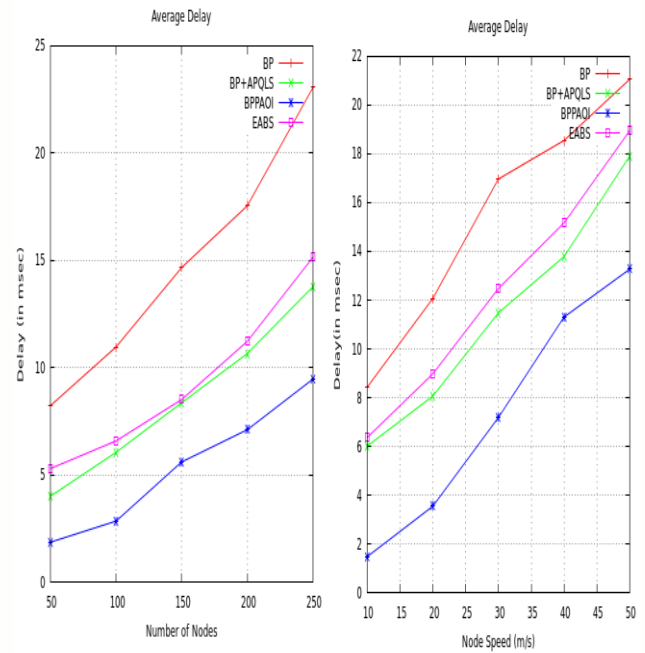
### 3 Results and discussions

#### 3.1 Simulation results

Simulation is carried out in NS3 and results are found. 2 Parameters, Number of nodes, and speed of node in the Network are varied to find out the working of the proposed algorithm under such circumstances.

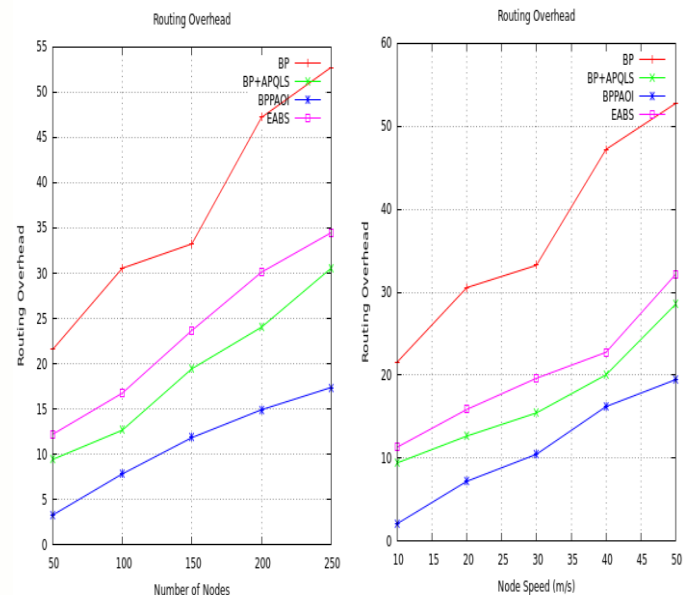
(a) Results by varying No. of Nodes in the Underwater N/W

(b) Results by varying the speed of Nodes in the Underwater N/W.



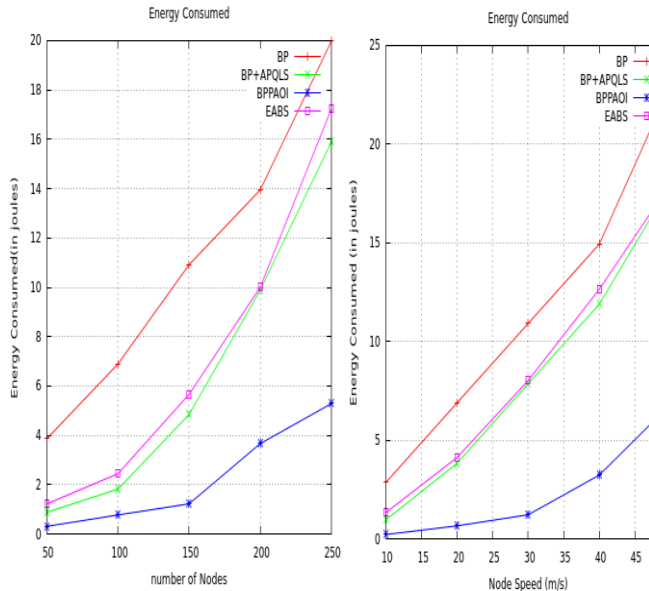
**Fig 4.** Delay in an underwater Network is calculated by (a) varying the number of nodes and (b) Varying the speed of nodes.

It can be seen from Fig. 4 that the Backpressure algorithm with priority and AOI-based technique [ BPPAOI] outperforms the rest of the algorithms. The delay is much reduced when compared to traditional backpressure, EABS (Tie Qiu et al.(2018)), and the BP + APQLS Technique (A Caroline Mary et al.).



**Fig 5.** Routing Overhead in an underwater Network is calculated by varying the number of nodes and speed of nodes.

It can be seen from Fig. 5 that the Backpressure algorithm with Active priority and AOI-based technique [BPPAOI] outperforms the rest of the algorithms in terms of Routing overhead.



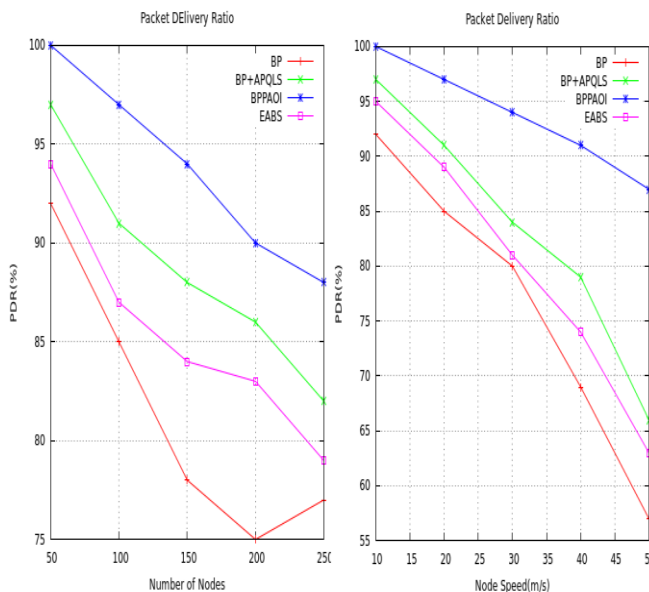
**Fig 6.** Energy consumption in an underwater Network is calculated by varying the number of nodes and speed of nodes.

Fig. 6 shows that the Backpressure algorithm with priority and AOI-based technique [BPPAOI] outperforms the rest of the algorithms in terms of Energy consumption. The initial energy of 50 Joules is assumed for all nodes and the final energy consumption is noted.

**Fig. 7.** The Packet Delivery Ratio in an underwater Network is calculated by varying the number of nodes and speed of nodes.

It is found that the BPA with Priority and AOI-based technique [BPPAOI] outperforms the rest of the algorithms in terms of packet delivery ratio as seen in Fig. 7 and in Table 2.

As smaller queues and high-priority real-time packets are given importance and as the Freshness of Information is maintained, the proposed algorithm results better. The use of peak AoI as a metric in considering the priority of the packet has improved the scheduling algorithm in terms of Packet delivery ratio, and energy consumption as shown in Figure 6. Normalized Peak AoI helps in treating time-critical applications with utmost importance. The use of active priority which adapts dynamically to the needs of the network has yielded better results. Energy plays a vital role in underwater Networks, as of the proposed algorithm energy consumption is less and Information freshness has yielded less packet loss and delay.



**Table 2. Comparison of Existing and Proposed Algorithms**

	BP	BP+ APQLS	EABS	Proposed with AOI BPPAOI
Congestion	Taken care of using Backlog based technique (Arnan Maipradit, Tomoya Kawakami, Ying Liu1 Juntao Ga & Minuro Ito (2021))	Taken care of using Backlog based technique and Active priority	Taken care of using Backlog based Technique	Taken care of using Backlog based technique and Active priority
Freshness of Information	Not considered	Not considered	Not considered	PeakAoI acts as an important metric as shown in Figure 1.
Priority	Not considered	Based on its priority, TTL and delay	Based on the importance of packets	Based on its priority, TTL, delay, Freshness of Information

#### 4. Conclusions

This study proposed a new method that is suitable for scheduling real-time packets in an underwater environment, which demands the freshness of Information, less delay, and less loss. The inclusion of Normalized class-based Peak AoI in assigning priority to the packets is a novel technique. Simulation results prove that almost one-third of Packet delay is reduced and the packet delivery ratio is more than around 90% which helps in reducing energy consumption. Other AoI-based metrics like the Freshness ratio of Information can be considered to improve the energy consumption and delay in future works. Underwater networks are vulnerable to security threats and malicious attacks. Added security can be considered in future works to avoid malicious attacks in underwater networks.

#### References

Arkadiusz Madej, Nan Wang, Nikolaos Athanopoulos, Rajiv Ranjan & Blesson Varghese (2020). Priority-based Fair Scheduling in Edge Computing. IEEE 4th International Conference on Fog and Edge Computing (ICFEC). DOI 10.1109/ICFEC50348.2020.00012.

Arnan Maipradit, Tomoya Kawakami, Ying Liu1 Juntao Ga & Minuro Ito (2021). An Adaptive Traffic Signal Control Scheme Based on Backpressure with Global Information. Journal of Information Processing Vol.29 124–131. Available from DOI: 10.2197/ipsjip.29.124.

Basel Barakat; Hachem Yassine; Simeon Keates; Ian Wassell & Kamran Arshad (2019). How to Measure the Average and Peak Age of Information in Real Networks?. European Wireless 25th European Wireless Conference. **Print ISBN:978-3-8007-4948-5**

Bo Hu, Xin Liu, Jinghong Zhao, Siya Xu, Zhenjiang Lei, Kun Xiao, Dong Liu & Zhao LiA (2020). Packet Scheduling Method Based on Dynamic Adjustment of Service Priority for Electric Power Wireless Communication Network. Wireless Communications and Mobile Computing. <https://doi.org/10.1155/2020/8869898>

Caroline Mary A, Dr.A V Senthil Kumar & Dr.Omar S. Saleh. (“in the press”). Packet Scheduling in the Underwater Network using Active Priority & QLS-based Energy Efficient Backpressure Technique. International Conference on Innovative Computing & Communication (ICICC-2023)

Chen.X, X. Liao & S. Saeedi-Bidokhti (2021). Real-time Sampling and Estimation on Random Access Channels: Age of Information and Beyond. IEEE International Conference on Computer Communications (INFOCOM). DOI: 10.1109/INFOCOM42981.2021.9488702

Igor Kadota, Abhishek Sinha & Eytan Modiano (2019). Scheduling Algorithms for Optimizing Age of Information in Wireless Networks with Throughput Constraints. O IEEE/ACM TRANSACTIONS ON NETWORKING. Volume 27, issue 4. Available from DOI: 10.1109/TNET.2019.2918736

Jaya Prakash Champati, Ramana R. Avula, Tobias J. Oechtering, & James Gross (2021). Minimum Achievable Peak Age of Information Under Service Preemptions and Request Delay. <http://dx.doi.org/10.1109/JSAC.2021.3065047>

Reshma Sultana S, Sindhu Meena K, Alaguvathana P & Abinaya K (2020). Priority-based Scheduling Algorithm using Divisible Load Theory in Cloud. International Conference on Inventive Systems and



Control. ISBN: 978-1-7281-2813-9, 2020.  
**DOI:** 10.1109/ICISC47916.2020.9171178

Roy D. Yates; Yin Sun; D. Richard Brown; Sanjit K. Kaul; Eytan Modiano & Sennur Uluk (2021). Age of Information: An Introduction and Survey. IEEE Journal on Selected Areas in Communications. Volume: 39, Issue: 5).  
**DOI:** 10.1109/JSAC.2021.3065072

SHAHZAD ASHRAF 1, MINGSHENG GAO 1, ZHENGMING CHEN 2, HAMAD NAEEM 3, ARSHAD AHMAD 4, & TAUQEER AHMED (2020). Underwater Pragmatic Routing Approach Through broad Reverberation Mechanism. VOLUME 8. IEEE Access, Digital Object Identifier 10.1109/ACCESS.2020.3022565

Tahir Kerem Oğuz , Elif Tuğçe Ceran , Elif Uysal & Tolga Girici (2022). Implementation and Evaluation of Age-Aware Downlink Scheduling Policies in Push-Based and Pull-Based Communication. Entropy. 24, 673. Available from <https://doi.org/10.3390/e24050673>

Tie Qiu, Senior Member, IEEE, Ruixuan Qiao, Student Member, IEEE, & Dapeng Oliver Wu, Fellow, IEEE (2018). EABS: An Event-Aware Backpressure Scheduling Scheme for Emergency Internet of Things. IEEE TRANSACTIONS ON MOBILE COMPUTING, VOL. 17, NO. 1.

**DOI:**10.1109/TMC.2017.2702670

Xingran Chen, Student Member, IEEE, Konstantinos Gatiss, Member, IEEE, Hamed Hassani, Member, IEEE & Shirin Saeedi Bidokhti, Member, IEEE (2022). Age of Information in Random Access Channels. IEEE TRANSACTIONS ON INFORMATION THEORY.  
**DOI:** 10.1109/TIT.2022.3180965

Xu. C, H. H. Yang, X. Wang, & T. Q. S. Quek (2019). On peak age of information in data preprocessing enabled IoT networks. IEEE Wireless Communications and Networking Conference (WCNC). **DOI:** 10.1109/WCNC.2019.8885690

Zhang. X, M. M. Vasconcelos, W. Cui & U. Mitra, (2021) Distributed remote estimation over the collision channel

with and without local communication. IEEE Transactions on Control of Network Systems, vol. Early Access.  
**DOI:** 10.1109/TCNS.2021.3100405

Zhuo Lu, Member, IEEE, Yalin E. Sagduyu, Senior Member, IEEE & Jason H. Li, Member, IEEE (2016). Securing the Backpressure Algorithm for Wireless Networks. IEEE Transactions on Mobile Computing. **DOI:** 10.1109/TMC.2016.2582161

## Application of text mining in PTT forum in analysis of consumer preference for online shopping platforms

Wen-ni Shih, Yu-sen Lin

Graduate Institute of Human Resource and Knowledge Management

National Kaohsiung Normal University

[wennie2468@gmail.com](mailto:wennie2468@gmail.com); [easonlin@nknu.edu.tw](mailto:easonlin@nknu.edu.tw)

(Received 26 June 2023; Final version 2 November 2022; Accepted 10 February 2023)

### Abstract

With the advent of economic development and Internet technology, offline retail stores have gradually shifted to virtual shopping networks, and consumers' online shopping has become increasingly prosperous. Moreover, since the COVID-19 pandemic, the public has taken the initiative to reduce the number of outdoor activities, which has increased consumers' willingness to shop online. This research takes Shopee, PChome, and MOMO online platforms as the research subjects. We obtained data from 2020 to 2021 on PTT e-shopping and lifeismoney boards. In addition, we used web text crawling analysis, R data text mining and positive/negative sentiment analysis, and word cloud to determine popular keywords related to online shopping issues, and consumers' preferences for online shopping platforms are studied. The results show that "seller", "problem", and "offer" are the most discussed keywords indicating that people care about the consumer experience to a certain extent. The next most frequent keywords are "coat", "dress", "shopee", "discount", "cheap", "Taobao", and "Taiwan", which will appear according to the needs of consumers in different seasons. Based on the sentiment analysis, the consumers posted more positive articles than negatives in PChome (2.33) and MOMO (2.34) compared to Shopee (1.11). Through term frequency analysis, we can understand the trends and suggestions brought by popular keywords of online shopping to consumers and online store sellers, and also allow online store sellers to analyze the key decision concerns and the possibility of customers' behavior.

*Keywords* : *Big data analysis, Online Shopping, Internet platform, Text mining*

# 應用 PTT 論壇文字探勘探討消費者對於 網購平台的偏好之研究

施文妮 林裕森\*

高雄師範人力與知識管理研究所

wennie2468@gmail.com

easonlin@nkn.edu.tw \*通訊作者

(Received 26 June 2023; final version 2 November 2022; Accepted 10 February 2023)

## 摘要

隨著經濟發展與網路科技時代的來臨，實體門市逐漸轉至虛擬網路，消費者的線上購物體驗也日漸蓬勃。自 2020 年新冠肺炎疫情爆發後，民眾為了避免群聚感染風險，主動減少外出次數，進而帶動了消費者到網路平台上購物的意願上升。本研究以蝦皮、PChome 以及 MOMO 此三大網路購物平台作為研究對象，透過 PTT 論壇在 2020 ~ 2021 於 e-shopping 與 lifeismoney 看板文章，運用網路爬文分析、R 語言的文字探勘及正/負向詞彙分析，再以視覺化文字雲圖表呈現網路購物議題之熱門關鍵字，探究消費者對網購平台的偏好性。本研究結果顯示，「賣家」、「問題」、「優惠」為每季討論度最高的熱門詞彙，顯現出民眾對於消費體驗的感受與用最划算的方式來購買商品的想法，都有一定的在意程度。其次常出現的關鍵字還有「外套」、「洋裝」、「蝦皮」、「折扣」、「便宜」、「淘寶」、「台灣」，會依照消費者在不同季別需求出現。情感分析也顯示論壇中的正負評文章的比值，PChome (2.33) 與 MOMO (2.34) 皆優於蝦皮 (1.11)。本研究藉由詞頻分析，了解網路購物熱門關鍵字所帶給消費者以及網購經營者的趨勢及建議。例如，可提供即將經營網路購物的潛在業者，了解網路購物的意義與優勢，增加新的行銷通路與顧客產生新的連結，以維持品牌延續及發展性。同時，近兩年疫情的發生，打亂了民眾購物的需求與管道，可以讓網路購物經營業者透過本研究分析，參考顧客對於在平台消費的關鍵決定性以及可能因素，也能避免供過於求而增加進貨成本。

**關鍵字:** 大數據分析，文字探勘，網路購物，網路平台。

## 1. 緒論

### 1.1 研究背景與動機

在網路時代的趨勢下，實體購物的行為逐漸轉至虛擬網路，再不受限於時間與地點的情況中，消費者的線上購物體驗也日漸蓬勃。Similarweb 數據分析平台顯示，統計 2020 年 4 月份的流量前三名網購電商平台使用分別為蝦皮、PChome 以及 MOMO (Similarweb, 2022)，現在人人只要利用網路拍賣平台，就能成為自己的老闆，不需要自架平台和社群媒體，或是投入大量成本來進行廣告宣傳效果，即可建立自己的賣場來累積流量。

然而，自 2020 年的新冠肺炎爆發後，疫情影響了大至全球經濟、小至每個人的生活模式，為了減少感染風險，人們主動減少外出的次數，相對減少實體門市店面的消費，進而帶動了消費者到網路平台上購物的選擇意願上升。台灣中央通訊社在 2020 年 08 月 05 日報導指出，疫情帶動網購業績成長，台灣在

2020 年上半年統計零售網路銷售額高達 1,587 億，相較於前一年同期增長了 17.5% (程倚華 (2021))；而比較於實體零售的營業額，2020 年相比前一年的同期卻衰減了 4.8% (中央通訊社, 2020)。自 2017 年台灣零售業網路銷售額是 2,283 億，到了 2019 年增長到了 2,873 億，在 2020 年更是突破了 3,000 億 (SHOPLINE TRENDS, 2021)。隨著科技日新月異、網際網路的蓬勃發展及行動支付的崛起，人們的消費方式越來越便利，疫情的爆發則是加速人們使用科技及便利功能的契機，於是促進更多在網路電商平台上進行購物。藉著此議題研究，我們可以反思的是網路消費偏好與購物動機類別，是否對數位電商購物模式所帶來的改變不只是正向也有負向結果。

要如何從大數據中蒐集到網路使用者在電商平台購物的偏好呢？疫情期間，消費者對於網購平台的喜好度為何？消費者因為疫情而被迫使用網購平台購物時，消費者的意見是否被重視？他們共同的意見有哪些？本研究採取目前台灣網路最大論壇 PTT 進行樣本蒐集，相較於其他社群平台，像 PTT 論

壇這樣的非營利組織，以匿名方式提供線上言論空間，藉由眾多不同種類的議題，都能在 PTT 上激盪出討論的聲浪，造成生活上的影響。

因此，本研究以蝦皮、PChome 以及 MOMO 三大網路電商平台為研究對象，利用網路爬文技術，在 PTT 相關購物版，包含 e-shopping、e-seller、Lifeismoney 各討論版，來分析不同網路電商平台的購物行為以及消費者購買搜尋關鍵字，以及文字雲呈現購物平台間的相關討論與詞頻分析，探討顧客對以上三大網路電商平台的偏好性。

## 1.2 研究目的

本研究探討網購平台之相關議題，透過 PTT 論壇文章摘錄後的資料分析與處理，再運用文字雲的視覺化呈現話題字詞篩選權重分析，並探討議題的情感分析狀況，進而了解消費者對於網購平台的使用偏好之差異。本研究的目的為：

- (1) 藉由 PTT 論壇相關文章，探討消費者對於三大網購平台之喜好度
- (2) 藉由 PTT 論壇相關文章，探討消費者對於網購正負向文章之評論

## 2. 文獻探討

### 2.1 網路購物

網路購物又可稱作電子商務 (Electronic Commerce)，是指顧客透過網際網路進行線上搜尋、受到媒體傳播後間接購買或直接購買之行為，也就是透過網際網路來從事各項商品的交易買賣方式，與作為契約訂定的管道。Zwass(1996)對電子商務研究指出「使用網際網路來維持商業的關係、進行相關商業交易活動及從事商業資訊的共享，各種使用網際網路來從事相關商業活動的行為，都可以稱之為電子商務。」

根據學者 Kalakota & Whinston(1997)認為，廣義的電子商務是一種現代化的經營模式，藉由電腦網路將購買與銷售、產品與服務等商業活動結合，可以滿足網路消費者「產品品牌」、「有試用期」、「可退換貨」、「售後服務」等需求，達到降低成本的要求，

並提高消費者購買意願。消費者網路購物環境，主要為電子商務模式中的 B2C(Business To Customer) 與 C2C(Customer To Customer)。總體而言，網路購物主要是能即時互動、低成本與無遠弗屆的新通路，消費者也透過網路商店更快速以及更容易的方式購買。而企業透過網路銷售產品，可以避免傳統商店人力、設備與店面等實體成本。

### 2.1.1 網路購物的發展

根據資策會(MIC)產業分析師陳冠文預期，發展線上通路的競爭局勢只會更激烈，不只實體零售業者，品牌業者、外送業者都是潛在競爭者。面臨後疫情時代的購物轉型，四大類型實體零售業者，包含百貨購物中心、連鎖便利商店、超市或量販店，皆須發展線上通路。合併店面優勢作為基礎，發展線上線下相互導流、並增加顧客與業者的互動來獲取網路購物忠誠度、以門市作快速出貨前置倉及取退換貨處等策略，提升全通路的銷售、行銷與顧客管理能力。除此之外，在 D2C(Direct-to-Consumer)趨勢之下，家居、3C 與服務等品牌業者也應積極布局數位轉型，建立自身的線上通路。

### 2.2 疫情下的購物模式

嚴重特殊傳染性肺炎(COVID-19)在 2019 年的 12 月開始大流行之後，全球陷入了經濟危機，導致多數產業需被迫停止實體營業，是近十年來所面臨到的最大經濟蕭條困境。聯合國貿易和發展會與 NetComm 瑞士電子商務協會做了一項研究(圖 1)，針對九個新興經濟體的國家中約 3,700 名消費者的進行問卷調查，發現到超過一半的調查受訪者，在疫情期間更頻繁地在線購物，更加的依賴互聯網獲取新聞、疫情相關信息和多媒體娛樂。

此調查統計顯示，每位購物者的平均每月在線支出已顯著下降，新興經濟體的消費者對於網上購物的轉變最大。他們將更多的支出放在化妝品及個人用品上。而旅遊業和觀光業的跌幅最大，每位網路購物者在此項的平均支出下降了 75%。





圖 1. 統計 2019 年九個經濟體國家的網購商品類形變化

資料來源: NetComm 瑞士電子商務協會(2019)

### 2.2.1 台灣在疫情下的購物影響

台灣在 2021 年中爆發嚴重疫情，從 5 月中至 7 月底，中央流行疫情指揮中心發布全國進入疫情第三級警戒，將近快三個月的嚴謹防疫規範，民眾避免前往鬧區，讓許多店家被迫停止營業，使得實體店面業績進入寒冬狀態，人潮與業績大幅驟降。眾信聯合會計師事務所消費產業負責人謝明忠觀察後指出「疫情改變人類的工作、學習和消費型態，而消費產業也隨著金流支付、線上線下整合及數據分析等創新工具與系統發展，帶出前所未見的商業模式，創造出了無接觸經濟的概念」。

為了避免染疫，消費者足不出戶，只有在家網購，實體人潮轉變為線上流量，帶動電商平台逆勢成長。而網路搜尋量較先前明顯成長 3 倍以上為防疫用品，例如：口罩、乾洗手和酒精等用品。許多網路購物平台皆因應疫情而開設防疫用品專欄；而民生日用品如衛生紙、尿布、洗衣精等銷量也大幅成長了 40%。電商平台產業指出：雖然實體店業績受到疫情波及的影響大，但同時經營電商官網實體門市的品牌店家，平均有雙位數的成長。原先會前往實體店的光顧的消費者也會轉移到線上購物，店家便可經營網路社群，把在家滑手機、不外出逛街的人導到官網選購。無論是賣生活用品、服飾鞋包、美妝，經營線上平台的品牌對於業績成長就能避免掉過多的經濟擔憂。

有別於實體門市會因為營業時間及地點而受到限制，利用線上與線下融合 OMO(Online merge Offline)虛實融合的品牌網路商店，就不會有這樣的困擾。電商平台對於消費者來說，就像便利商店般全時段服務的隨身店舖。根據財團法人台灣網路資

訊中心統計，受訪者中有 65.2% 會透過網路購物，平均一個月花費 2,661 元，表示在網路購物下所帶動的經濟潛力，有龐大的提升機會。關鍵評論網(THE NEWS LENS)也統計民眾對三級警戒後選擇網購的考慮因素，受訪者表示，自 2021 年 2 月至 8 月，選擇網購主要是因為價格合理，其次是有提供免運或是較低運費，第三名考量原因是促銷活動，網路消費決策前三名都與「錢」有直接關係。

### 2.3 網路行銷

根據 Janal(1995)研究認為，網路行銷是針對使用網際網路與商業網路服務之特定用戶，銷售產品與服務的系統，配合公司的整體行銷規劃，藉由網路系統促使用戶可利用網路工具與服務，獲取資訊及購買產品。而李宛穎(1999)以網際網路作為行銷通路與傳播之媒介，進行產品或服務的銷售或促銷，並促使顧客利用網路工具和服務，獲取其所需要的資訊與購買產品；近年來顧客導向理念之興起，使得網路購物研究也逐漸重視網路使用者購物動機對於新興購物模式接受度與行銷策略偏好之影響。再者，對於消費者進行市場區隔，乃是依據各個購買者特質或反應來區分不同消費群，再因應各種消費群對網路購物之特殊需求與期望，來擬定行銷策略組合(Armstrong & Kotler, 2000; 余強生、曾雍欽, 2003)。透過了解網際網路使用者購物動機不同，也促使面對不同網路行銷策略之偏好也會有所差異(Hoffman & Novak, 1996; Alba et al, 1997)。

### 2.3.1 網路行銷現況

根據經濟部統計處零售業網路銷售報告顯示(圖2),消費者轉移至線上通路購物與日俱增,各類零售產業皆於電子商務表現上呈持續成長,消費者在疫情期間無法前往實體通路,使得線上購物等關鍵字之搜尋量大增。不管是實體店面或是電商品牌,都可以善用網路開店、人人都能建立社群及利用直播通路銷售,快速拉近與消費者之間的距離,在疫情期間運用數位工具持續經營品牌和消費者產生連結互動,鞏固老顧客的關係,也創造與新客戶的連結。改變經營模式跳脫疫情帶來的危機,同時也能掌握最新的市場消費動向。

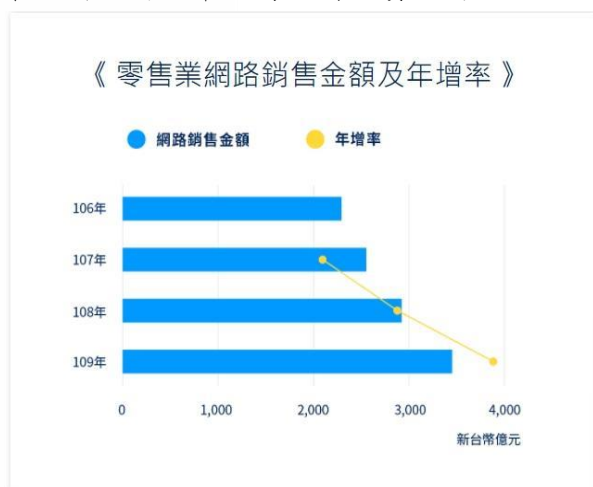


圖2 近三年零售業網路銷售報告  
資料來源:經濟部(年)

### 2.4 文字探勘技術

文字探勘(Text Mining)是利用資料探勘方法用於文字資料分析中,從文字中萃取出重要資訊之分析技術,而輿情分析(Public opinion analysis)則是文字探勘分析中的一種應用方式,利用斷詞和語句來判斷文章中隱藏之情感與正負向的情緒。隨著近年行動科技發達,輿情分析隨之快速興起,此分析方法不僅具有社群大數據之特性,更具高時效、資料易於蒐集、低成本以及可快速反應民意等優點。文字探勘可用於顧客關係管理,進一步探究顧客消費行為與偏好,使得企業能夠延續發展新的行銷通路,找到不同客群最適合之最佳行銷管道(朱瑀馨, 2007)。而另一項研究,學者則利用社交媒體

Facebook 和 Twitter 上的非結構化文字內容,分析三個最大的披薩供應商的社群資訊,進而協助披薩業者能夠瞭解競爭對手之情報(He et al., 2013)。

在數據探勘的過程,無論是為了甚麼目的,或是要如何應用,其共通點皆為先釐清定義問題、研究解決方法、資料選擇,在進入建立模式與資料驗證的階段,詳細說明如下(圖3)。

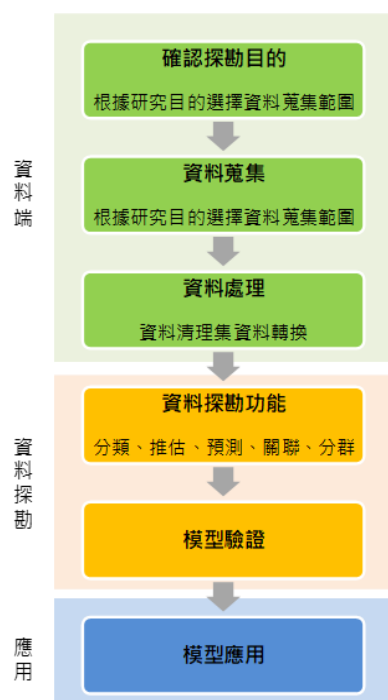


圖3 資料探勘流程圖

(1) 確認探勘目的: 訂定探勘問題、釐清研究目標、規劃研究流程。

(2) 資料蒐集: 針對研究目的,從不同來源及不同特質數據庫中,篩選符合研究目的及條件的資料。

(3) 資料處理: 對資料的正確性進行驗證,以便篩出錯誤或遺失資料,包含資料清理(異常值偵測、校正錯誤與清理不合理資料)及資料轉換(資料格式標準化)。

(4) 資料探勘: 依照研究目的與資料性質,對應其資料探勘功能之方法。資料探勘常見功能可分為5大類。

①分類(Classification): 按照分析對象的屬性分門別類,加以定義。

②推估(Estimation): 根據既有連續性數值之相關屬性資料,已獲得某一屬性未知值。

③預測(Prediction)：根據分析對象屬性之過去觀察值，來預測該屬性未來值。

④關聯(Association)：從所有物件決定那些物件應該放在一起。

⑤分群(Sequence)：將異質母體中區隔為較具同質性之群組(Clusters)。同質分組相當於行銷術語中的區隔化，假定事先未對於區隔加以定義，而資料中自然產生區隔。

(5) 建立模型：分析方法分為監督式學習(Supervised Learning)與非監督式學習(Unsupervised Learning)。

(6) 模型驗證：監督式學習方法所建立模型，經訓練集建立後，使用測試集進行模型驗證，若測試集結果與訓練集一致，才能進行應用；若測試集結果與訓練集不一致，則要回到資料探勘重新建立模型。

(7) 模型應用：經過嚴謹建立之模型，於訓練集與測試集比對後，訂定最終模型，此模型方能進入模型實務應用與模型之機器學習。

## 2.5 網路輿情分析技術

在輿情分析方面，目前最常使用之方法包含文字雲、詞頻分析與語意分析等，透過這些方法可將質化之文章內容轉換成量化之數值結果，有關上述方法之內容與方法說明如下：

### 2.5.1 網路輿情分析技術

網路輿情分析是以某項事件為中心的一種分析方法，如圖 4 所示，透過分析網路上消費者的意見發表，來進一步針對分析的結果對於民眾、媒體的情

感意見做說明，甚至影響後續的發展性。目標客群對於品牌、產品或是服務的想法，進行歸納與剖析。分析結果直接通過網路發佈，或點擊型式表現出來，網路民眾與媒體雖不具有如現實環境的實在性，虛擬的環境隱藏網民真實身分，因此發表民眾真實的想法，在相當程度上反映出及代表著民眾意識及立場，以及客觀存在著頗為厚實的民意基礎(公平交易委員會，2020)。

網路輿情分析所使用資料皆是從網路爬文而來，透過抓取網路上的文章、點擊、按讚等資訊，來得之民眾的想法。網路輿情分析的優點在於：

可知未經包裝的民意，民眾的想法直接的呈現。

過去因使用問卷蒐集民意，但在敏感性問題時，民眾普遍不願接受訪問，但網路輿情分析對於敏感性話題的資料收集較容易取得。

網路輿情資料來源是由廣大的網路空間，不會受到問卷設計及訪問者訪問方式影響，反映出來的民意亦較客觀。

資料蒐集不需耗費龐大的資金及人力成本，而且資料更新非常快速且便捷。

網路輿情之資料來源，是利用網路爬文快速蒐集網路媒體與民眾所產生的文章與紀錄，使用文字探勘方法進行網路聲量的長期追蹤、文章剖析，將大量質化資料轉為有價的量化數值，即時了解民眾對於該議題的長期討論熱度變化、評價變化、討論內容的意向及走向，以供決策者快速掌握民眾與媒體對於議題的想法，以及解媒體與民眾對於該議題反對或支持的內容，或於適時的時間點拋出民眾所需之計畫、方向、作法及建議等，以有效提升支援政策決策之時效性(台灣E化資安分析管理協會，2020)。



圖 4 網路輿情分析關係圖

資料來源：台灣E化資安分析管理協會







### 3.2.2 資料處理

為了提高資料分析結果的正確率，使用自行撰寫之程式來刪除文章內之回文，去除文章中存有多餘或非本意的資料，例如：發文作者、顏文字等。

### 3.2.3 資料分析

#### 1. 詞頻分析

詞頻分析是經過斷詞後，從大量非結構化的文字中找出關鍵字，進而檢視詞頻較高之詞彙分析，計算某字詞在語庫中或同一篇文章的出現頻率，來衡量字詞的重要性。簡單來說，字詞統計出現次數頻率越大，代表該詞彙被討論或使用的機率越高。

#### 2. 情感分析

情感分析(Semantic Analysis)是利用詞語庫定義情緒詞彙，情緒詞彙會隨著上下詞語呈現情緒的拉扯，計算該情緒詞彙於該段落之情感分數，會隨著上下詞語呈現情緒的拉扯，進而計算該篇文章之情感分數。本研究藉由台灣大學的中文情感極性辭典(NTUSD)，每篇文章情感傾向以 0 為基準，情感分數大於 1，則該篇文章屬於正向評價，若小於-1 則為負向評價，等於介於-1 到+1 之間則是中立，詞語庫中包含正向與負向詞語，正向詞語像是讚、很好、很優、超好的等字眼，負向詞語則是爛、差、不好等字眼，最後計算文章中的詞彙總數。

#### . 文字雲視覺化

資料分析視覺化常用於將關鍵字具體化描述，文字雲(WordCloud)的使用在於能讓閱讀者在不詳

再透過中文斷詞 jieba 的 R 套件，對文章裡的文字資料進行斷句，對於字元資料而言，每一個字詞都是代表語義的最小單位，因此有必要將文章中的文字做分割的動作，才能轉換為分析處理之格式。

讀所有文章的前提下，將文字檔轉為可視化的詞頻率統計權重表，快速聚焦在大量文章中的主要內容。本研究藉由繪製文字雲工具 WordArt，設定文字的大小與顏色區別，突顯出文章裡討論詞頻多寡與差異性。

## 4. 研究結果

本研究以網路購物為研究核心，擷取 PTT 論壇中的 e-shopping 及 lifeismoney 板裡 2020-2021 年度所有看板文章，透過 R 語言之套件爬文下載文章進行字詞分析，並進行文字資料預先處理，分析出視覺化可用資訊，分別產生詞頻、並透過整理相關關鍵字的詞頻分析及管理意涵，再進行情感分析之研究步驟，藉以探討網路購物對於消費者之議題分析說明結果。

### 4.1 樣本描述

本研究以 PTT 中的 e-shopping 及 lifeismoney 板為主要論壇分析看板，資料擷取時間為 2020 年 1 月至 2021 年 12 月之文章，以 1-3 月、4-6 月、7-9 月、10-12 月將 txt 檔案歸類依月季別，e-shopping 板有 10,779 篇、lifeismoney 板有 14,595 篇。

表 1 總計看板文章篇數彙整  
資料來源:本研究整理

看板 年/季	e-shopping	lifeismoney	總文章數
2020Q1	1,691	1,560	3,251
2020Q2	1,537	1,841	3,378
2020Q3	1,271	1,630	2,901
2020Q4	1,561	2,104	3,664
2021Q1	1,276	1,555	2,831
2021Q2	968	1,737	2,705
2021Q3	1,160	1,767	2,927
2021Q4	1,315	2,401	3,715
總計	10,779	14,595	

### 4.1.1 購物平台偏好性之情感傾向

本章節以 PTT 論壇 e-shopping 板及 lifeismoney 板，運用 R 語言作為研究工具，分析出蝦皮、PChome、MOMO 三大網購平台之文章情感傾向，深入了解消費者對於不同平台的喜好度。

統計 PTT 論壇 2020 年 1 月至 2021 年 12 月的區間時間裡，將三大網路購物平台做文章情感傾向分析，如圖 7 所示，蝦皮網購平台文章篇數共有 2,416 篇，其中 507 篇為正向文章，455 篇為負向文章，1,454 篇為中立文章，文章情感傾向結果(正評文章/負評文章)之比值為 1.11 (507/455)。

PChome 網購平台的文章篇數共有 1,050 篇，其中 255 篇為正向文章，109 篇為負向文章，686 篇為中立文章，文章情感傾向結果為 2.33 (255/109)。MOMO 網購平台的文章篇數共有 1,390 篇，其中 350 篇為正向文章，149 篇為負向文章，891 篇為中立文章，文章情感傾向結果為 2.34 (350/149)。

根據分析數據比值的顯示，透過 R 語言正負向分析蝦皮、PChome、MOMO 三大網購平台，發現到三個網購平台百分比皆大於 1，表示分析結果均為正向，其中 PChome 及 MOMO 網購平台的正向情感分析，比起蝦皮網購平台更加顯著。

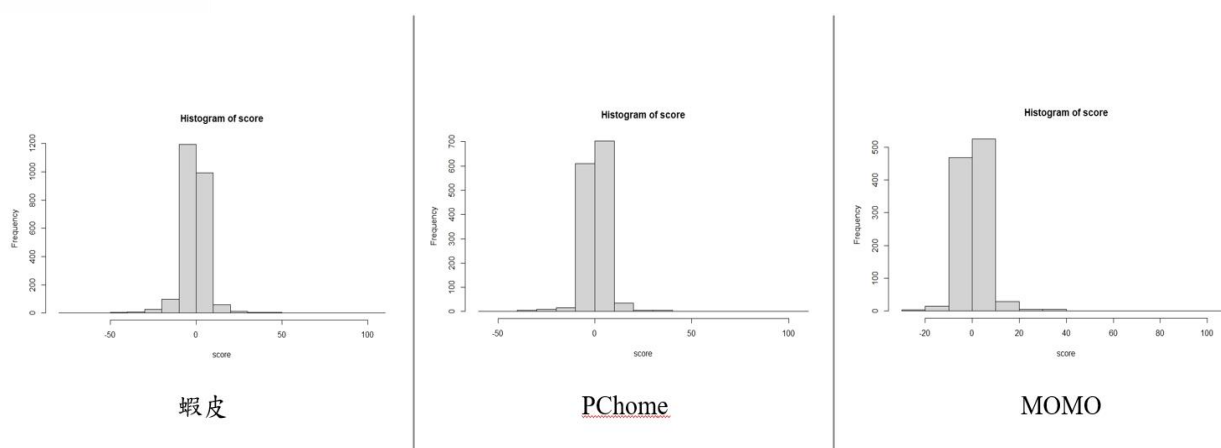


圖 7 三大網購平台文章情感傾向結果  
資料來源:本研究整理

## 4.2 熱門詞彙分析

### 1. 2020 與 2021 第一季詞彙分析

2020 與 2021 年第一季前 50 出現頻率最高關鍵字排名，可以發現兩年度重複出現的共有詞彙有賣家(2,426)、優惠(2,104)、洋裝(1,809)、問題(1,781)、外套(1,458)、福袋(1,243)、上衣(1,169)、大衣(1,169)、折扣(1,149)、選手(1,119)、款式(1,102)、價格(1,071)、蝦皮(1,065)、衣服(1,063)、顏色(1,026)、猶豫不決(1,005)、拍賣(988)、便宜(845)、淘寶(819)、客服(797)、尺寸(783)、免費(774)、針織(767)、可愛(758)、原價(749)、材質(745)、好看(706)、消費(697)、設計(679)、官網

(667)、點數(655)、品牌(652)、會員(622)、店家(585)、運費(549)、手機(520)。

扣除兩年度重複性的前 50 關鍵字詞頻率，屬於 2020 年出現的特有詞彙有台灣(562)、日本(497)、連線(384)、口罩(362)、板友(347)、日貨(336)、下標(330)、需求(311)、毛衣(294)、質感(287)，而屬於 2021 年出現的特有詞彙有咖啡(328)、買一送一(274)、捐血(271)、集運(269)、兌換(255)、今年(254)、過年(228)、方便(214)、白色(213)。

根據 2020 與 2021 年的第一季關鍵字統計比較，以下為本季推測分析說明：

(1) 本季時間接近農曆過年，因此會出現福袋、折扣、價格、優惠、免費、便宜等關鍵字詞。

(2) 因本季的天氣狀態，使得消費者對於保暖衣物需求較高，因此會出現針織、大衣、外套、材質關鍵字。

### 2. 2020 與 2021 第二季詞彙分析

2020 與 2021 年第二季前 50 出現頻率最高關鍵字排名，可以發現兩年度重複出現的共有詞彙有優惠(2,365)、賣家(1,856)、問題(1,768)、便宜(1,099)、折扣(995)、台灣(987)、特價(980)、價格(974)、洋裝(960)、選手(959)、客服(918)、淘寶(872)、猶豫不決(868)、衣服(778)、疫情(695)、點數(669)、拍賣(656)、品牌(630)、尺寸(628)、會員(619)、顏色(617)、買一送一(585)、運費(566)、集運(560)、官網(549)、材質(538)、設計(524)、日本(513)、折價券(466)、免運(464)、可愛(419)。

扣除兩年度重複性的前 50 關鍵字詞頻率，屬於 2020 年出現的詞彙有上衣(448)、原價(421)、適合(348)、好看(322)、訂單(290)、全家(286)、口罩(268)、夏天(235)、貼圖(228)、退貨(228)、信用卡(223)、襯衫(222)、白色(217)、外套(207)。而屬於 2021 年出現的詞彙有外帶(351)、防疫(310)、款式(308)、捐血(247)、水桶(229)、手機(222)、廣告(222)、板友(218)、門市(215)、方便(211)、咖啡(203)。

根據 2020 與 2021 年的第二季關鍵字統計比較，以下為本季推測分析說明：

(1) 2021 年度第二季因進入台灣疫情三級警戒風暴，消費者在限制外出的情況下開始搜尋更多的商品類別有外帶、防疫、貼圖、廣告、全家等關鍵字。

(2) 在這一季當中，有部分鄉民在看板發表購物抱怨文章，因此特別出現的關鍵字有水桶(是一種版主可動用的處罰之一)、退貨等關鍵字。

### 3. 2020 與 2021 第三季詞彙分析

2020 與 2021 年第三季前 50 出現頻率最高關鍵字排名，可以發現兩年度重複出現的共有詞彙有賣家(1,808)、優惠(1,800)、問題(1,505)、洋裝(1,160)、折扣(1,110)、便宜(1,033)、價格(1,023)、選手(987)、蝦皮(975)、猶豫不決(888)、客服(869)、款式(866)、特價(837)、上衣(837)、點數(815)、衣服(743)、淘寶(719)、拍賣(711)、尺寸(707)、運費(662)、日本(651)、品牌(643)、免費(601)、設計(597)、官網(595)、材質(586)、會員(571)、原價(568)、買一送一(549)、折價券(501)、照片(493)、官方(489)、好看

(474)、免運(471)、信用卡(470)、可愛(460)、全家(457)、集運(454)、限時(418)。

扣除兩年度重複性的前 50 關鍵字詞頻率，屬於 2020 年出現的詞彙有價格(478)、振興(289)、會員(261)、家樂福(259)、導購(257)、退貨(249)、賣場(230)、三倍(220)、評價(217)、加碼(212)、水桶(201)、韓貨(197)。而屬於 2021 年出現的詞彙有捐血(520)、訂單(365)、顏色(289)、疫情(287)、兌換(265)、全聯(229)、禮券(228)、限定(221)、白色(207)。

根據 2020 與 2021 年的第三季關鍵字統計比較，以下為本季推測分析說明：

(1) 2020 年度 7 月開始，政府為了振興經濟發放三倍卷，而隔年 9 月也因疫情的加劇發行了五倍卷，因此出現特有關鍵字為振興、三倍、疫情、禮卷、加碼、家樂福、全聯、限定等關鍵字。

### 4. 2020 與 2021 第四季詞彙分析

分別為 2020 與 2021 年第四季前 50 出現頻率最高關鍵字排名，可以發現兩年度重複出現的共有詞彙有優惠(2,763)、問題(2,111)、賣家(1,994)、蝦皮(1,405)、折扣(1,300)、便宜(1,274)、特價(1,159)、淘寶(1,129)、台灣(1,113)、客服(1,085)、洋裝(1,073)、點數(997)、選手(963)、猶豫不決(875)、外套(856)、款式(844)、官網(799)、折價券(742)、集運(736)、尺寸(731)、上衣(730)、衣服(729)、拍賣(728)、運費(727)、原價(694)、買一送一(677)、會員(676)、限定(670)、官方(668)、日本(662)、手機(659)、免費(652)、兌換(646)、品牌(635)、限量(632)、限時(622)、免運(622)、加碼(611)、照片(610)、顏色(351)。

扣除兩年度重複性的前 50 關鍵字詞頻率，屬於 2020 年出現的詞彙有信用卡(332)、退貨(322)、好看(310)、材質(306)、樂天(301)、襯衫(297)、可愛(287)、划算(288)、咖啡(284)、黑色(282)。而屬於 2021 年出現的詞彙有價格(607)、訂單(497)、導購(369)、紅包(334)、五倍(311)、振興(310)、全家(303)、禮卷(301)、疫苗(287)、捐血(285)。

根據 2020 與 2021 年的第四季關鍵字統計比較，以下為本季推測分析說明：

(1) 賣場在年末期間配合檔期，而推出一系列的購物促銷，包含雙 11、黑色星期五及 12 月的聖誕節活動，因此第四季會出現限時、折扣、優惠、便宜、加碼、免費、折價券等關鍵字。







**圖 8** 2020 與 2021 年度關鍵字文字雲(以季分類)

資料來源:本研究整理

信(164)、兄弟(158)、紅包(127)、街口(121)、樂天(101)。

### 4.3 正負向分析

#### 4.3.1 正向文章之分析

##### 1. MOMO

論壇中提到 MOMO 網購平台的文章,2020 年度有 608 篇,2021 年度有 782 篇。在正向詞頻分析結果計算後,可以看見兩年度常出現的正向詞彙(次數)前十名有回饋(532)、活動(519)、優惠(398)、便宜(314)、點數(302)、情報(291)、分享(367)、喜歡(225)、價格(224)、推薦(204)。而僅屬於 MOMO 購物平台在正向詞頻分析中會出現的特有詞彙有味道(140)、金額(120)、家樂福(120)、購物網(104)。

##### 2.PChome

論壇中提到 PChome 網購平台的文章,2020 年度有 480 篇,2021 年度有 570 篇。在正向詞頻分析結果計算後,可以看見兩年度常出現的正向詞彙(次數)前十名有回饋(710)、活動(534)、優惠(339)、情報(253)、點數(250)、連結(195)、價格(178)、便宜(175)、導購(172)、猶豫不決(171)。而僅屬於 PChome 購物平台在正向詞頻分析中會出現的特有詞彙有中

##### 3. 蝦皮

論壇中提到蝦皮網購平台的文章,2020 年度有 1193 篇,2021 年度有 1223 篇。在正向詞頻分析結果計算後,可以看見兩年度常出現的正向詞彙(次數)前十名有蝦皮(1,350)、賣家(608)、心得(544)、分享(515)、活動(463)、便宜(434)、價格(364)、推薦(356)、情報(340)、優惠(319)。而僅屬於蝦皮購物平台在正向詞頻分析中會出現的特有詞彙有衣服(223)、洋裝(221)、淘寶(220)、可愛(208)、外套(169)、顏色(159)、朋友(145)、照片(145)。

##### 4. 三大網路購物平台正向文章詞頻文字雲：

分別將 MOMO、PChome、蝦皮之正向文章詞頻,以文字雲視覺化呈現(圖 9),可以明顯觀察到最常出現的正向文章詞彙為回饋、活動、優惠、便宜、點數、情報、分享、心得。其次出現正向詞彙有喜歡、價格、連結、導購、猶豫不決、蝦皮、賣家、推薦。從文字雲中,可以了解到一般民眾對於用最經濟實惠的方式購買商品有一定的重視程度。


**圖 9** 三大網路購物平台正向文章詞頻文字雲

資料來源;本研究整理

### 4.3.2 負向文章之分析

#### 1. MOMO

在負向詞頻分析結果計算後，可以看見兩年度常出現的負向詞彙(次數)前十名有問題(191)、水桶(187)、退貨(177)、情報(168)、活動(153)、分享(148)、廣告(137)、回饋(136)、違規(133)、處分(124)。而僅屬於 MOMO 購物平台在負向詞頻分析中會出現的特有詞彙有違規(133)、處分(124)、包裹(118)、累犯(112)、抱怨(107)、退文(104)。

#### 2. PChome

在負向詞頻分析結果計算後，可以看見兩年度常出現的負向詞彙(次數)前十名有問題(163)、客服(145)、情報(143)、回饋(141)、活動(139)、退貨(138)、台灣(128)、蝦皮(126)、悠遊(122)、價格(97)。而僅屬於 PChome 購物平台在負向詞頻分析中會出現的特有詞彙有物流(42)、消費者(41)、平台(41)、螢幕(39)。

### 3. 蝦皮

在負向詞頻分析結果計算後，可以看見兩年度常出現的負向詞彙(次數)前十名有賣家(1,474)、問題(682)、訂單(423)、取消(371)、退貨(354)、退款(266)、心得(227)、抱怨(225)、運費(208)、評價(199)。而僅屬於蝦皮購物平台在負向詞頻分析中會出現的特有詞彙有截圖(150)、假貨(139)、水桶(133)。

#### 4. 三大網路購物平台負向文章詞頻文字雲：

分別將 MOMO、PChome、蝦皮之負向文章詞頻，以文字雲視覺化呈現(圖 10)，可以明顯觀察到最常出現的負向文章詞彙為問題、水桶、退貨、情報、活動、回饋。其次出現負向詞彙有分享、廣告、違規、處分、客服、台灣、蝦皮、悠遊、價格、賣家、訂單、取消、退款、心得、抱怨、運費、評價。從文字雲中，透過負向文章詞彙的篩選後發現，顧客重視的希望能夠透過其他購買者的經驗回饋，去決定是否要購買產品，以及賣家對於不遵守網路購物流程的消費者，有權力拒絕服務。



圖 10 三大網路購物平台負向文章詞頻文字雲

資料來源；本研究整理

## 5. 結論與建議

### 5.1 研究結論

1. 根據 2020 與 2021 年的每一季前後關鍵字熱門關鍵字結果統計相互比較，結合管理意涵的闡述研究之說明，做出以下的幾點分析：
2. 第一季的時間接近農曆過年，因此常出現與過年

3. 節慶相關等關鍵字詞，例如，如果未來 2023 年第一季發現“福袋”關鍵字出現的頻率多寡，與消費者的購物行為模式有很大的相關性，則取決於賣家應該多準備或少準備以“福袋”為名義的行銷活動。第一季的天氣狀況，使得關鍵字會常出現對於保暖衣物需求較高的詞彙。是否對於購買新的衣物為因應過年的過節氣氛，而非單純因為氣候因素，值得更多的數據以進行長期觀察。
3. 2021 年度第二季因進入台灣疫情三級警戒風暴，

消費者在限制外出的情況下開始搜尋更多的疫情相關類別關鍵字。疫情在同樣的時間是否皆達到疫情變化的高峰期，賣家可以留意在未來這一季是否消費者對於醫療及防疫相關用品的需求大增。

4. 在這兩年的第二季觀察中，可以發現有部分鄉民在看板發表購物抱怨文章，因此特別出現的關鍵字有水桶(是一種版主可動用的處罰之一)、退貨等關鍵字。可以看出當時疫情因素，工廠公司開始分流上班，甚至縮短營業時間與人力，間接影響到業者的供貨與物流，也讓消費者的購物品質下降。本土業者與海外廠商的作業，都因為疫情而影響至商品到達顧客手上的流程，是在接下來的年度中業者要隨時留意的。
5. 在第三季中，2020 年度 7 月開始，政府為了振興經濟發放三倍卷，而隔年 9 月也因疫情的加劇發行了五倍卷，因此會出現特有振興卷等關鍵字。依目前的情況來看，疫情影響了民眾的消費，看不見疫情局勢終止的情況下，沒有節慶活動為名義，也可以透過發行商品卷與加碼卷，促進民眾消費意願。
6. 第四季的時間裡，賣場在年末配合檔期，而推出一系列的購物促銷，包含雙 11、黑色星期五及 12 月的聖誕節活動。賣家可以透過這樣的歲末檔期把過季商品或是庫存用優惠、禮包的方式做出清活動，刺激消費者在購物體驗上有物超所值的購物感受。另外於這一季的觀察中，2021 年度的捐血意識提升，因此每一季提到捐血關鍵字的詞頻率上升，加上還有全家、禮卷、咖啡等關鍵字，也可以看出因應捐血，醫療單位會利用兌換便利商店的商品配合方式，鼓勵民眾來捐血，利用做好事可以得到其他回饋為出發點。
7. 針對正向文章與負向文章的詞頻，本研究做出以下的幾點分析：
8. 正向詞彙統計結果中，可以發現到消費者在 MOMO 平台上對於購買產品的嗅覺體驗上較在意，例如香氣、清潔劑等產品。其次 MOMO 也有與家樂福量販店合作的購物金優惠販售，因為才會有相關詞彙的產生。PChome 對於與其他企業互惠及合作有很大的興趣。例如 2021 年中信兄弟球隊在 12 月奪下總冠軍，因此 PChome 推出了特別購物優惠，以及在 Line 社群平台建立抽獎購物金紅包的活動，以及與樂天信用卡和街口支付方式，讓消費者能運用指定付款方式得到點數回饋金。最後，使用蝦皮瀏覽的消費者，對於衣著的商品類型有較大的關注度，甚至對於商品風格的需求比起其他兩個網路購物平台有更高的興趣。
9. 負向詞彙統計結果中，可以觀察出許多在 MOMO 購物平台消費的顧客，未在取貨時間內領貨，使得業者須要做出適當的申明及懲處來提醒消費者勿再犯，也保障雙方的權益。而 PChome 的特有負

向詞頻大多與購物感受有關。最後在蝦皮購物網站平台，可以觀察到網友在蝦皮購物碰到了品牌保障的問題。而若有網友發表抱怨文章，一旦發文超過十分鐘即不可自行刪除，違反無法在論壇上發言五年，因此有了水桶這項關鍵字。

## 5.2 研究限制與建議

根據本研究的結果與討論，從研究關鍵字分析可以提出後續研究方向與建議，以及對於網路購物平台經營業者與網路顧客，點出幾項實務面的參考，在進行往後的行銷策略、購買決策與品牌形象的定位上，也能有進一步的實質幫助。

### 5.2.1 對於網路購物平台使用者的建議

1. 提供即將經營網路購物的潛在業者，可以了解網路購物的意義與優勢，增加新的行銷通路與顧客產生新的連結，以維持品牌延續及發展性。
2. 近兩年新冠疫情的發生，重新打亂了民眾對於物質的需求，可以讓網路購物經營業者透過本研究分析，參考顧客對於在平台消費的關鍵決定性以及可能因素，也能避免供過於求而增加進貨成本。

### 5.2.2 對於研究者的建議

1. 本論文只針對近兩年的資料作蒐集，然而後續的研究者可以運用本篇論文在持續追縱疫情的狀況，做更新一步的資料更新，提供業者更即時的參考回饋。
2. 本論文在正向與負向詞頻分析中，相同的詞彙面臨到一體兩面的解釋情況，建議後續研究者可以進一步詳細了解文章內容，去解釋文字確切屬於正向或是負向的情緒表達。
3. 本研究在不影響文章正負評結果的情況下，已先在資料處理上做刪除回文動作，建議後續研究者可以另外整理文章回文之研究，探究網友評論的情感分析。

### 5.2.3 研究限制

1. 本研究並未針對疫情發生前消費者對於網購平台的偏好，原因是因為研究者認為，疫情期間的網購生態、銷售項目與物流服務等，有了根本上的變化，過去認為隨手可得的生活用品，因為疫情期間的相關交通與人流管制措施，都需要透過網購來取得。因此，研究者認為，更完整的消費者對於網購平台的意見，應該是針對疫情前、疫情中與疫情後的資料，做一個縱貫性研究(longitudinal study)。



2. 本研究只針對 PTT 論壇上 e-shopping 板及 lifeismoney 板文章進行研究分析,後續的研究者可以在拓展運用不同的社群,例如 Dcard 論壇或是臉書粉絲團等樣本,或是發現新的 PTT 相關討論板以增加新的資料文章來做議題研究。
3. 本研究僅選擇出熱門使用前三名的網路購物平台蝦皮、PChome、MOMO 進行資料分析,但未能探究出各個平台主攻的商品類型或市場,建議後續的研究者可以拉長資料蒐集範圍,或是詳細針對關鍵字追蹤做文章分類,進行網路購物平台的商品差異化分析。
4. 本論文的資料來源,MOMO 以及 PChome 的樣本數與蝦皮在同樣的蒐集範圍時間卻有文章篇數的差距,建議未來研究者可以增加不同購物平台的比較,得到更多研究結果意涵的可能。

## 文獻

### 中文

- MIKAKO (2020)。Top10 台灣十大網路購物電商平台排名。取自  
<https://www.top10.com.tw/life/938/top-10-online-shopping-website/>。檢索日期：2021 年 11 月 15 日。
- MIC 產業情報研究所 (2021)。【網購消費者調查】52.9%消費者購物頻率虛實各半,實體零售網購崛起。  
取自 <https://mic.iii.org.tw/news.aspx?id=597>。檢索日期：2021 年 11 月 12 日。(MIC, 2021)
- SHOP LINE TRENDS, 2021 台灣疫情消費趨勢報告。檢索日期：2021 年 11 月 18 日。取自  
<https://trends.shopline.tw/covid-19>
- Jasmine Huang (2020)。武漢肺炎：疫情如何衝擊零售業？網路電商危機變轉機！取自  
<https://www.91app.com/blog/coronavirus-retail-impact/>。檢索日期：2021 年 11 月 12 日。
- 台灣 E 化資安分析管理協會(2020)。網路輿情分析瞬息萬變,中中事件演變提早因應。取自  
<https://www.netadmin.com.tw/netadmin/zh-tw/technology/C6FD7FD04ECB8B6624E8E014A922>。  
檢索日期：2021 年 11 月 15 日。(CSAM, 2020)

- 公平交易委員會(2020)。網路銷售市場競爭評估之實證研究。109 年委託研究報告。中華民國 109 年 12 月 (FTC, 2020)
- 中央通訊社 (2020)。疫情帶動網購業績成長,上半年銷售額年增 17.5%。取自  
<https://www.cna.com.tw/news/afe/202008050107.aspx>。檢索日期：2021 年 11 月 12 日。(CNA, 2020)
- 朱瑀馨 (2007)。運用資料探勘技術於人壽保險業顧客關係管理之研究,淡江大學保險學系保險經營研究所。(Zhu, 2007)
- 余強生、曾雍欽 (2003)。網際網路購物者特性,購物動機,期望的網站服務與顧客滿意度之間的結構化方程式模型。企業管理學報,(57),37-64。(Yu and Zeng, 2003)
- 李宛穎 (1999)。線上銷售考量因素之研究,國立中山大學企業管理研究所碩士論文。(Li, 1999)
- 林金錫 (2015)。華語教學一點靈-兩款文字雲網站。檢索日期：2021 年 12 月 3 日。取自  
<http://huayu.chinhsilin.com/?p=70> (Lin,2015)
- 林韋伶 (2021)。消費習慣回不去了！警戒降級後,「這幾類商品」在電商通路持續暢銷。檢索日期：2021 年 11 月 15 日。取自  
<https://www.businesstoday.com.tw/article/category/183015/post/202108100028> (Lin, 2021)
- 陳映竹 (2013)。消費者行動消費現況分析。檢索日期：2021 年 11 月 15 日。取自  
<https://mic.iii.org.tw/industry.aspx?id=77> (Chen, 2013)
- 程倚華 (2021)。網購銷售額年增 17%！後疫情零售怎麼做？專家拆解數位轉型 4 階段 3 技巧。檢索日期：2021 年 11 月 15 日。取自  
<https://www.bnext.com.tw/article/61940/deloitte-retail-omo> (Cheng, 2021)
- 蕭乃沂等 (2015)。政府應用巨量資料精進公共服務與政策分析之可行性研究,國家發展委員會。(Xiao, 2015)

### English

- Alba, J., Lynch, J., Weitz, B., Janiszewski, C., Lutz, R., Sawyer, A., & Wood, S. (1997). Interactive home shopping: consumer, retailer, and manufacturer incentives to participate in electronic marketplaces. *Journal of Marketing*, 61(3),38-53.
- He, W., Zha, S., & Li, L. (2013). Social media competitive analysis and text mining: A case



- study in the pizza industry. *International Journal of Information Management*, 33, 464-472.
- Hoffman, D. L., & Novak, T. P. (1996). Marketing in hypermedia computer-mediated environments: Conceptual foundations. *Journal of Marketing*, 60(3), 50-68.
- Janal, D. S., (1995), "Online Marketing Handbook: How to Sell, Advertise, Publicize, & Promote Your Product & Services on Internet & Commercial Online Systems" Van Nostrand.
- Kalakota, R., & Whinston, A. B. (1997). *Electronic commerce: a manager's guide*. Addison-Wesley Professional.
- Kotler, P., Armstrong, G., Saunders, J., Wong, V., Miquel, S., Bigné, E., & Cámara, D. (2000). *Introduction to marketing*. Pearson Prentice Hall.
- Similarweb (2022). similarweb.com. Accessed on Mar 30, 2022
- UNCTAD NEWS (2020,October 08) COVID-19 has changed online shopping forever,survey shows <https://unctad.org/news/covid-19-has-changed-online-shopping-forever-survey-shows>. Accessed on Nov 21,2021.
- Victoria,Fryer.(2021) Understanding COVID-19's Impact on Ecommerce and Online ShoppingBehavior <https://www.bigcommerce.com/blog/covid-19-ecommerce/#covid-ecommerce-trends>. Accessed on Nov 21, 2021.

- Zwass, V. (1996) "Electronic commerce: structures and issues". *International Journal of Electronic Commerce*. 1(1), Fall, 3-23.

### 作者簡介



林裕森

現任國立高雄師範大學人力與知識管理研究所教授兼所長。

成功大學學士、美國匹茲堡大學工程博士、企管碩士。



施文妮

現職於金豐企業管理顧問有限公司 業務助理。

文藻外語大學國際企業管理學士、國立高雄師範大學人力與知識管理研究所碩士。

## INSTRUCTIONS TO AUTHORS

### *Submission of papers*

The International Journal of Systematic Innovation is a refereed journal publishing original papers four times a year in all areas of SI. Papers for publication should be submitted online to the IJoSI website (<http://www.ijosi.org>) In order to preserve the anonymity of authorship, authors shall prepare two files (in MS Word format or PDF) for each submission. The first file is the electronic copy of the paper without author's (authors') name(s) and affiliation(s). The second file contains the author's (authors') name(s), affiliation(s), and email address(es) on a single page. Since the Journal is blind refereed, authors should not include any reference to themselves, their affiliations or their sponsorships in the body of the paper or on figures and computer outputs. Credits and acknowledgement can be given in the final accepted version of the paper.

### *Editorial policy*

Submission of a paper implies that it has neither been published previously nor submitted for publication elsewhere. After the paper has been accepted, the corresponding author will be responsible for page formatting, page proof and signing off for printing on behalf of other co-authors. The corresponding author will receive one hardcopy issue in which the paper is published free of charge.

### *Manuscript preparation*

The following points should be observed when preparing a manuscript besides being consistent in style, spelling, and the use of abbreviations. Authors are encouraged to download manuscript template from the IJoSI website, <http://www.ijosi.org>.

1. *Language.* Paper should be written in English except in some special issues where Chinese maybe acceptable. Each paper should contain an abstract not exceeding 200 words. In addition, three to five keywords should be provided.
2. *Manuscripts.* Paper should be typed, single-column, double-spaced, on standard white paper margins: top = 25mm, bottom = 30mm, side = 20mm. (The format of the final paper prints will have the similar format except that double-column and single space will be used.)
3. *Title and Author.* The title should be concise, informative, and it should appear on top of the first page of the paper in capital letters. Author information should not appear on the title page; it should be provided on a separate information sheet that contains the title, the author's (authors') name(s), affiliation(s), e-mail address(es).
4. *Headings.* Section headings as well as headings for subsections should start front the left-hand margin.
5. *Mathematical Expressions.* All mathematical expressions should be typed using Equation Editor of MS Word. Numbers in parenthesis shall be provided for equations or other mathematical expressions that are referred to in the paper and be aligned to the right margin of the page.
6. *Tables and Figures.* Once a paper is accepted, the corresponding author should promptly supply original copies of all drawings and/or tables. They must be clear for printing. All should come with proper numbering, titles, and descriptive captions. Figure (or table) numbering and its subsequent caption must be below the figure (or table) itself and as typed as the text.
7. *References.* Display only those references cited in the text. References should be listed and sequenced alphabetically by the surname of the first author at the end of the paper. For example:


Altshuller, G. (1998). *40 Principles: TRIZ Keys to Technical Innovation*, Technical Innovation Center.  
Sheu, D. & Lee, H. (2011). A Proposed Process for Systematic Innovation, International Journal of Production Research, Vol. 49, No. 3, 2011, 847-868.

**The International Journal of Systematic Innovation  
Journal Order Form**

<b>Organization Or Individual Name</b>	
<b>Postal address for delivery</b>	
<b>Person to contact</b>	Name: _____ e-mail: _____ Position: _____ School/Company: _____
<b>Order Information</b>	<b>I would like to order ___ copy(ies) of the <i>International Journal of Systematic Innovation</i>:</b> <b>Period Start: 1<sup>st</sup>/ 2<sup>nd</sup> half ____, Year: ____ (Starting 2010)</b> <b>Period End : 1<sup>st</sup>/ 2<sup>nd</sup> half ____, Year: ____</b> <b>Price:</b> <b>Institutions: US \$150 (yearly) / NT 4,500 (In Taiwan only)</b> <b>Individuals: US \$50 (yearly) / NT 1500 (In Taiwan only)</b> (Local postage included. International postage extra) <b>E-mail to: <a href="mailto:IJoSI@systematic-innovation.org">IJoSI@systematic-innovation.org</a> or fax: +886-3-572-3210</b> Air mail desired <input type="checkbox"/> (If checked, we will quote the additional cost for your consent)
<b>Total amount due</b>	<b>US\$</b>
<b>Payment Methods:</b>	
<ol style="list-style-type: none"> <li><b>Credit Card (Fill up the following information and e-mail/ facsimile this form to The Journal office indicated below)</b></li> <li><b>Bank transfer</b></li> <li><b>Account:</b> The Society of Systematic Innovation</li> <li><b>Bank Name:</b> Mega International Commercial BANK</li> <li><b>Account No:</b> 020-53-144-930</li> <li><b>SWIFT Code:</b> ICBCTWTP020</li> <li><b>Bank code :</b> 017-0206</li> <li><b>Bank Address:</b> No. 1, Xin'an Rd., East Dist., Hsinchu City 300, Taiwan (R.O.C.)</li> </ol>	

**VISA / Master/ JCB/ AMERICAN Cardholder Authorization for Journal Order**

**Card Holder Information**

Card Holder Name	(as it appears on card)		
Full Name (Last, First Middle)			
Expiration Date	/ (month / year)	Card Type	<input type="checkbox"/> VISA <input type="checkbox"/> MASTER <input type="checkbox"/> JCB
Card Number	□□□□-□□□□-□□□□-□□□□	Security Code	□□□ 
Amount Authorized		Special Messages	
Full Address (Incl. Street, City, State, Country and Postal code)			

Please Sign your name here \_\_\_\_\_ (same as the signature on your card)

**The Society of Systematic Innovation**  
6 F, #352, Sec. 2, Guanfu Rd,  
Hsinchu, Taiwan, 30071, R.O.C.