Quarterly Journal of
Optimization In Soft Computing

Vol. 2, Issue. 2, Summer 2024

Journal of Optimization of Soft Computing (JOSC)

Vol. 2, Issue 2, pp: (1-7), Summer-2024

Journal homepage: http://josc.rimag.ir/en

Research paper

Investigating the Impact of Distributed Generation (DG) in Radial Distribution Networks and Optimizing Protective Devices Using the PSO Optimization Algorithm

Narjes Mohsenifar [1*] and Najmeh Mohsenifar [2]

1. Department of Electrical Engineering, Shahrekord Branch, Islamic Azad University, Shahrekord, Iran.
2. Chaharmahal and Bakhtiari Provincial Electricity Distribution Company, Shahrekord.

Article Info

Article History:
Received: 2024/04/28
Revised: 2024/09/08
Accepted: 2024/09/09

DOI:

Keywords:
Distributed energy resources, radial distribution network, coordination of protective systems, DigSilent software, PSO optimization algorithm, Matlab software

**Corresponding Author's Email Address: mohsenifar86@gmail.com*

Abstract

Integrating distributed energy resources into existing power networks can lead to significant challenges. These challenges often arise from the coordination of protective devices within the network. In this paper, we investigate the impact of connecting Distributed Generation (DG) to radial distribution networks and propose a novel approach for optimizing protective device settings to enhance their coordination. By analyzing different relay settings in the DigSilent software under various fault conditions, we determine the optimal configurations for the system in the presence of DG. The Particle Swarm Optimization (PSO) algorithm is employed using MATLAB. The proposed approach is applied to a radial distribution network, and the results are evaluated under different fault scenarios. In each scenario, a short circuit happens and relays operate in order; the operation intervals for relays are 0.02s, approximately.

Introduction

The demand for electrical energy is increasing day by day. To meet this demand, new technologies called Distributed Generators (DGs) are being introduced into the system. These include renewable energy sources such as photovoltaic (PV), wind, fuel cells, biomass, etc. Integrating DGs into existing networks enhances network robustness, reduces losses, and operational costs at peak times, while improving voltage profiles and load factors [1]. In this situation, the reliability of electricity supply to consumers is enhanced.

However, placing new energy sources close to consumers introduces various operational challenges in the system. Implementing Distributed Generation (DG) in the system can create significant issues in distribution networks because existing distribution systems are designed for radial flow. The presence of DG may impact or violate existing planning and operational procedures [2]. Integrating distributed generation and other storage devices into radial networks alters the traditional unidirectional power flow, significantly affecting the coordination of protective devices used in the system.

By adding DG to existing feeders, nominal current or fault current is redistributed through the feeder. This can significantly impact load current and fault current seen by protective devices. In such scenarios, variations in fault current levels occur. Increased short-circuit capacity and changes in fault current direction in the distribution network affect the coordination of protective relays installed in the distribution system, thereby compromising the performance, selection, and reliability of protection schemes [3].

M. H. Bresten and et al, investigate the aspects of grid-connected converters and their inherent influence on the power grid [4] and in 2023, laboratory tests of distribution feeder protection response was done with inverter-based resources [5].

Optimal settings for protective coordination devices in DG-connected systems are essential (Abbay, 2007). To achieve this, relay operation

times have been optimized using various optimization methods. Common trial-and-error approaches are also used to determine parameter settings for different operational conditions. Artificial intelligence methods have proven highly effective for optimizing coordination parameters in DG-connected networks [7].

In this article, the main objective is to utilize an artificial intelligence-based algorithm for optimizing the coordination parameters of protection systems in distribution networks connected to DG. Protection plays a crucial role in these networks and is essential in the power industry.

Genetic algorithms and Particle Swarm Optimization (PSO) are computational intelligence techniques proposed for solving optimization problems. The genetic algorithm mimics evolutionary biology to find approximate optimal solutions [8]. While it can quickly find good solutions, it has drawbacks, including convergence toward local optima instead of global solutions and difficulties in execution due to dynamic data sets.

In specific optimization problems with computational time constraints, simpler optimization methods may yield better results than this algorithm. As reported in studies by Hasan et al, (2005) [9] and Arya et al, (2010) [10], the PSO algorithm consistently achieves superior results compared to other optimization methods, especially genetic algorithms.

The advantages of the PSO method over other approaches, particularly genetic algorithms, are as follows:

Executing PSO with fewer parameters for tuning is easier.

1. Executing PSO with fewer parameters for tuning is easier.
2. The memory capability of PSO is more effective than the genetic algorithm because each particle is able to memorize its previous best position and local location in the best possible way.
3. PSO is more efficient for maintaining diversity within a population. This is because PSO utilizes the most successful information to move toward better solutions, similar to the social behavior of a community. In contrast, genetics disregards inferior solutions and only retains good ones.

This article focuses on the impact of DG on the coordination of protection systems in radial distribution networks. An optimal protection scheme has been developed to determine the best settings for protective devices. A 4-bus system is studied, simulated using DigSilent software, and optimized using the PSO algorithm in Matlab. The structure of the article is as follows: In the second section, various protection issues in radial distribution networks with distributed generation (DG) are discussed. The third section introduces the PSO optimization algorithm and explains how it relates to the two software programs. The fourth section describes a 4-bus system case study and presents the simulation results. Finally, the article concludes with suggestions for future research in this field.

Protective Issues in the Presence of DG

For reliable operation of DG connected to the system, proper coordination with the network is essential. DG is connected to the network through the Point of Common Coupling (PCC). This PCC must be adequately protected to prevent damage to DG equipment and auxiliary tools. Therefore, appropriate coordination of protective devices is necessary. Coordination of protective equipment in simple radial networks is straightforward because the network has a single source [2], resulting in unidirectional fault current flow. The entry of DG into the distribution network results in bidirectional current flow and complexity in network topology. Protective device settings and features must be adjusted according to the new topology.

DG can be utilized at different locations in the distribution network:

1) End of the line with fault creation in the direction of the main source and distributed generation.

2) Between the end of the line and the main network.

3) At the end of the line and the occurrence of an error in a branch other than the main source and DG.

Therefore, to analyze the effects of DG, a three-bus general network is examined under three different scenarios.

First System Configuration

The first state of the 3-bus test system is shown in Figure (1). In this system, an error has occurred at bus 2.

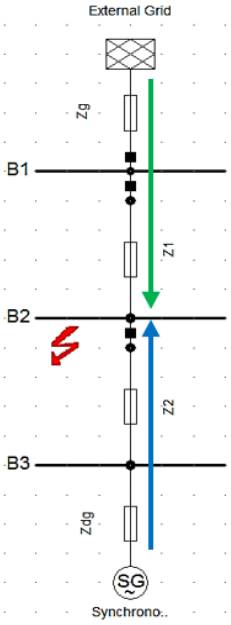


Figure 1: First case study of a 3-bus system.

Considering the circuit in Figure 1, the three-phase short-circuit current is the sum of the short-circuit currents from the network and the DG at the fault point. These are represented by the green and blue directional lines, respectively. The short-circuit currents from the network (I_g) and DG (I_{dg}) can be expressed as follows:

$$I_g = \frac{U_g}{Z_g + Z_1} \quad (1)$$

$$I_{dg} = \frac{U_{dg}}{Z_{dg} + Z_2} \quad (2)$$

In the formulas (1) and (2), U_g and U_{dg} , respectively represent the external network voltage and the DG-generated voltage. Z_g, Z_{dg} , Z_1 and Z_2 , respectively denote the internal impedance of the network, the internal impedance of DG, and the impedances of the first and second lines.
The value of the internal impedance of distributed generation can be calculated as follows:

$$Z_{dg} = X''_{dg}\frac{S_{sys}}{S_{dg}} \quad (3)$$

X''_{dg}, Underpass reactance of DG. S_{sys} denotes the system base power in MVA. S_{dg} is DG capacity (MVA).

Here, the fault current exists in both directions relative to the fault point. Therefore, bidirectional relays are necessary to detect reverse current.

second System Configuration:
The second case study involves a three-bus experimental system, as shown in Figure 2. In this scenario, a fault is created at bus 3.

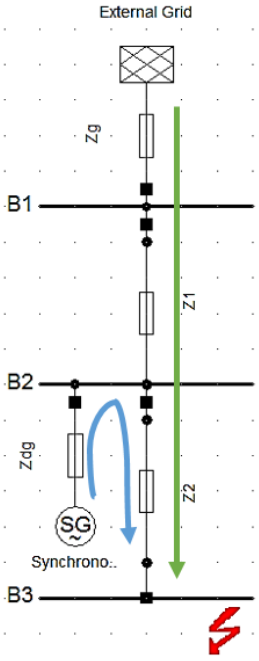


Figure 2: Second case study of a 3-bus system.

In the circuit shown in Figure 2, the three-phase short-circuit current at bus 3 is equal to the sum of the short-circuit currents from the network and DG. The short-circuit currents from the network (I_g) and DG (I_{dg}) can be expressed as follows:

$$I_g = \frac{Z_{dg}}{Z_{dg}(Z_g + Z_1 + Z_2) + Z_2(Z_g + Z_1)}U_g \quad (4)$$

$$I_{dg} = \frac{1}{Z_{dg} + Z_2}\left[U_{dg} - \frac{Z_2 Z_{dg}}{Z_{dg}(Z_g + Z_1 + Z_2) + Z_2(Z_g + Z_1)}U_g\right] \quad (5)$$

As observed, the main network current at the fault point depends on the presence of DG and its location. This may impact the sensitivity and selection of the relay connected to the system. The lower value of the main network current at the fault location affects the performance of the feeder end relay. The relevant relay does not turn off when it sees the fault or is disconnected after a long delay.

Third System Configuration
Figure 3 illustrates the third case study of a three-bus experimental system. In this scenario, a fault

occurs at bus 3 in a branch separate from the network path to DG.

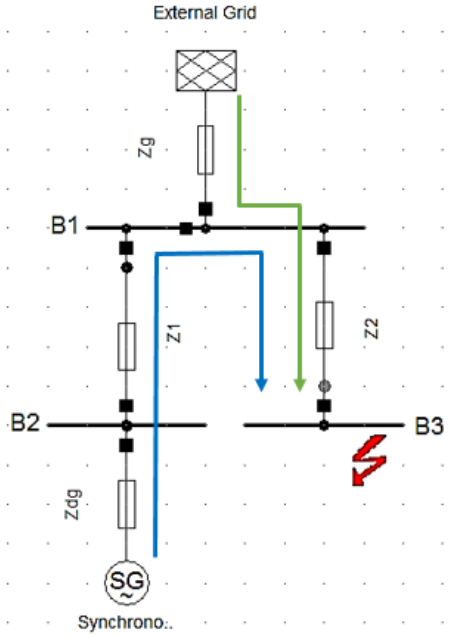


Figure 3: Third case study of a 3-bus system.

As observed in the circuit shown in Figure 3, the three-phase short-circuit current at bus 3 is equal to the sum of the short-circuit currents from the network (green) and DG (blue). The short-circuit currents from the network (I_g) and DG (I_{dg}) can be expressed as follows:

$$I_g = \frac{(Z_g + Z_1 + Z_2)U_g - Z_2 U_{dg}}{(Z_1 + Z_{dg})(Z_2 + Z_g) + Z_g} \quad (6)$$

$$I_{dg} = \left[\frac{\left(Z_g - Z_2(Z_g + Z_2)\right)U_g + Z_2(Z_g + Z_2)U_{dg}}{(Z_1 + Z_{dg})(Z_g + Z_2) + Z_g}\right] \quad (7)$$

The impact of this system can be considered as an example of false tripping. For a significant share of DG current at the fault point, Feeder 2 trips instead of Feeder 3, which is unacceptable.

DG Impact on Distribution Networks

Considering the three systems examined in the previous subsections, the effects of Distributed Generation (DG) in the distribution network can be summarized as follows. Considering the presence of Distributed Generation (DG) in the network, the fault current direction changes, and this topology causes bidirectional fault current to vary. The fault current level increases or decreases, which leads to a loss of sensitivity and selectivity in the protective device. With an increase in the nominal current, the protective device issues incorrect trip commands. Conversely, when the network current decreases, blind spots are created in protection. The fault current level that is less than the instantaneous pickup value affects the trip time. These issues [9] directly or indirectly related to each other. Changes in the short-circuit current level are the main cause of protection coordination mismatch.

Study Methodology

In this section, the PSO method in Matlab software and the connection between Matlab and DigSilent are presented.

PSO Optimization Method

Mathematically, the search process can be expressed using simple equations involving position vector $X_i = [x_{i1}.x_{i2}.\ldots.x_{id}]$ and velocity vector $V_i = [v_{i1}.v_{i2}.\ldots.v_{id}]$ in a d-dimensional search space. The optimality of the solution in the PSO algorithm depends on the update of particle positions and velocities, which are calculated using equations (8) and (9) (Del Valle et al., 2008).

$$V_i^{k+1} = wV_i^k + c_1 r_1\left[X_{pbest}^k - X_i^k\right] + c_2 r_2\left[X_{gbest}^k - X_i^k\right] \quad (8)$$

$$X_i^{k+1} = X_i^k + V_i^{k+1} \quad (9)$$

In the above equations, i represents the particle index. V_i^k and X_i^k are, respectively, the velocity and position of the i-th particle in the k-th iteration. w is the inertia constant, typically within the range [0, 1]. $c_1$and c_2 are convergence coefficients within the interval [0, 2]. r_1 and r_2 are random values generated for updating the velocity. X_{gbest} and X_{pbest} represents the global best position across all iterations, and represents the local best position in the current iteration.

The connection between Matlab and DigSilent software:

In this article, an algorithm for optimizing the existing relay settings in medium-voltage networks has been developed using the PSO method within the Matlab software. The calculations related to short-circuit analysis, fault current magnitude, and total relay operating time are performed using the DigSilent software. Figure 4 illustrates the input and output connections between the DigSilent and Matlab software platforms.

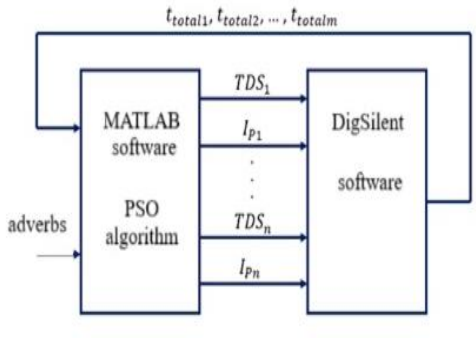


Figure 4: Illustration of the communication between DigSilent and Matlab software platforms.

In Figure 4, m, n and t_{total}, represent the number of buses (or the number of short-circuit calculations created in each bus), the number of relays, and the total relay operating time, respectively. These values are calculated using Equation (10).

$$t_{totalh} = t_{1h} + t_{2h} + \cdots + t_{nh} \quad . \quad 1 < h < m \qquad (10)$$

In the given equation, t_1 to t_n, represent the operating time of the first to the n-th relay present in the network. In the designed PSO algorithm, there exists a constraint and an objective. The defined constraint pertains to the first relays that must operate after a fault occurs. The objective is to minimize the total relay operating times for fault occurrences in each bus.

Simulation Conducted

This section comprises two subsections, which include the studied system and simulation results.

Studied System

The studied system is depicted in Figure 5.

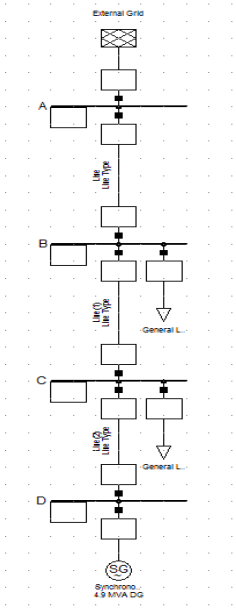


Figure 5: The studied system.

This distribution system consists of 4 buses, a main network, two loads, and one DG unit. Additionally, there is a relay at the beginning and end of each line. The relays present in the system are depicted in Figure 6.

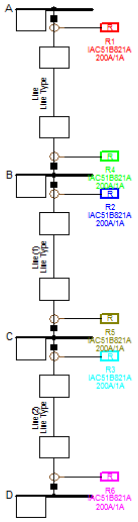


Figure 6: Relays present in the studied system.

Simulation Results

The applied constraint in the PSO algorithm is listed in Table 1.

By running 1178 iterations in the PSO algorithm, the results obtained for the optimal values of TDS and I_p for each relay are presented in Table (2).

In figures (7) to (10), the results obtained in the Digsilent software are visible.

Table (1): Constraints Applied in the PSO Algorithm

The sequence of operation of the bus low-side relay.	The sequence of operation of the bus high-side relay	The bus where the error occurred
1-4-2-5-3-6	-	A
2-5-3-6	4-1	B
3-6	5-2-4-1	C
-	6-3-5-2-4-1	D

Table (2): Results obtained from the designed PSO algorithm

I_p	TDS	Relay
0.39	0.051	R_1
0.5	0.05	R_2
0.55	0.051	R_3
1.2	0.05	R_4
0.9	0.05	R_5
0.7	0.052	R_6

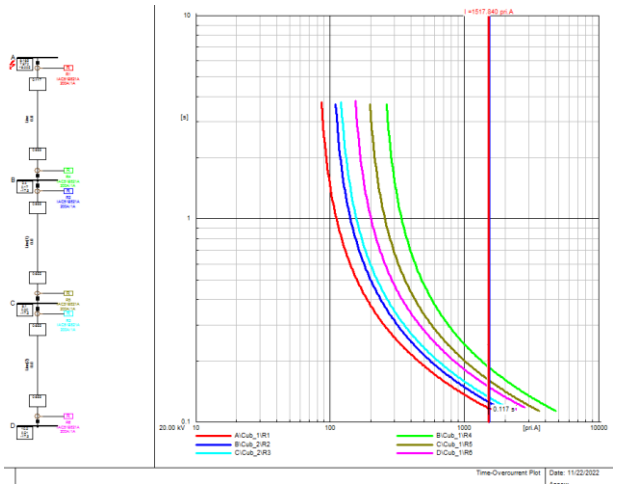


Figure (7): The results of relay performance with an error in the first bus.

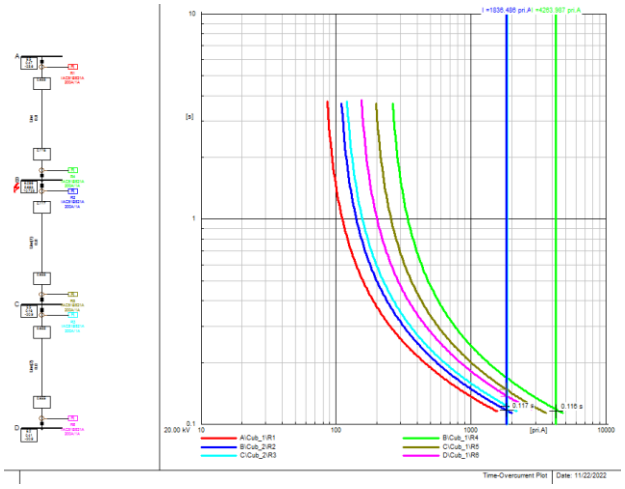


Figure (8): The results of relay performance with an error in the second bus

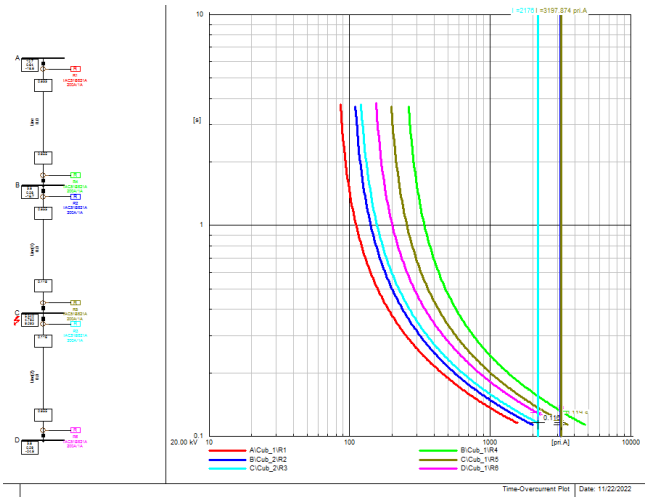


Figure (9): The results of relay performance with an error in the third bus

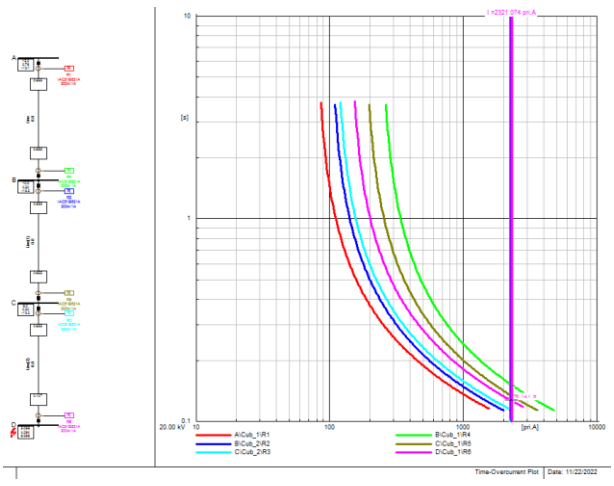


Figure (10): The results of relay performance with an error in the fourth bus

As observed in figures (7) to (10), the relay performance sequence has been correctly adhered to for each fault.

Conclusion and Recommendations:

The objective of this article is to determine the optimal settings for relays in the network when DG is introduced. As seen in section 4, with the integration of DG into the network, the entire relay settings require review. The PSO algorithm was employed to identify the best settings. By introducing faults in each bus, the affected bus is isolated by the existing relays, while the rest of the network remains energized by the main network and DG.

In future work, additional constraints, such as relay operation time intervals, could be incorporated into the PSO algorithm, or an alternative generalized optimization method could replace the algorithm used in this study.

References

[1] H. Zhan *et al.*, "Relay Protection Coordination Integrated Optimal Placement and Sizing of Distributed Generation Sources in Distribution Networks," IEEE Transactions on Smart Grid, vol. 7, pp. 1-1, 04/27 2015, doi: 10.1109/TSG.2015.2420667.

[2] E. Coster, J. M. A. Myrzik, W. L. Kling, and D. N. Gaonkar, "Effect of dg on distribution grid protection distributed generation," *InTech,* pp. 93-118, 01/01 2010.

[3] L. A. Kojovic and J. F. Witte, "Improved relay coordination and relay response time by integrating the relay functions," *2000 Power Engineering Society Summer Meeting (Cat. No.00CH37134),* vol. 2, pp. 1202-1207 vol. 2, 2000.

[4] M. H. Brestan and R. Schuerhuber, "Aspects of grid-connected converters and their inherent influence on the power grid," in *2023 23rd International Scientific Conference on Electric Power Engineering (EPE)*, 24-26 May 2023 2023, pp. 1-5, doi: 10.1109/EPE58302.2023.10149242.

[5] N. Gruman and P. Moses, Laboratory Tests of Distribution Feeder Protection Response with Inverter-Based Resources. 2023, pp. 1-5.

[6] C. Abbay, Protection coordination planning with distributed generation (Abbay). Canada, 2007.

[7] S. Katyara, L. Staszewski, and Z. Leonowicz, "Protection Coordination of Properly Sized and Placed Distributed Generations–Methods, Applications and Future Scope," Energies, vol. 11, no. 10, p. 2672, 2018. [Online].

[8] J. Nocedal and S. Wright, Numerical Optimization. Springer New York, 2006.

[9] R. Hassan, B. Cohanim, O. d. Weck, and G. Venter, "A Comparison of Particle Swarm Optimization and the Genetic Algorithm," in 46th AIAA/ASME/ASCE/AHS/ASC Structures, Structural Dynamics and Materials Conference, 2005.

[10] L. D. Arya, L. S. Titare, and D. P. Kothari, "Improved particle swarm optimization applied to reactive power reserve maximization," International Journal of Electrical Power & Energy Systems, vol. 32, no. 5, pp. 368-374, 2010/06/01/ 2010.

Journal of Optimization of Soft Computing (JOSC)

Vol. 2, Issue 2, pp: (8-17), Summer-2024

Journal homepage: http://josc.rimag.ir/en

Research paper

Using Soft Computing and Chaos Theory in investigating the Deformed Stadium

Rashid Riahi

Energy and Environment Research Center, Shahrekord Branch, Islamic Azad University, Shahrekord, Iran.

Article Info

Article History:
Received: 2024/06/23
Revised: 2024/08/14
Accepted: 2024/09/08

DOI:

Keywords:
Chaos, Billiard, Poincaré map, Cross section, Birkhoff mapping

**Corresponding Author's Email Address: r.riahi@iaushk.ac.ir*

Abstract

This paper analyzed the dynamic system of billiards from a classic perspective. For this purpose, mapping and cross-section methods were first employed to study the behavior of this system and the results indicated that it was a chaotic one. Then a deformed stadium was introduced and its long-term behavior was analyzed. Considering changes in the behavior of this system following the slightest deformation at the boundaries, Poincaré map was used to demonstrate the occurrence of regular and irregular motions, indicating the completely chaotic behavior of the system. The shape of the cross-section of the regular motion shows that the points of contact with the boundary are located on a line in the phase space. On the other hand, the cross-sectional surface of a chaotic motion, the surface is covered with collision points and the empty spaces are surrounded by invariant curves. These spaces are also filled in case of $n \to \infty$ and they eventually disappear and the surface is covered with collision points, completely. This behavior is characteristic of chaotic systems

1. Introduction

"Chaos" has a Greek origin, denoting a gaping void or a chasm that existed before all things. Romans applied the word to the rough, shapeless mass from which architects of the world create order and harmony. In modern language, chaos is used to imply disorder and lawlessness. Upon the introduction of Newton's laws in 1687, scholars used them to solve copious problems. Due to the large variety of these problems, they believed that the subsequent states of the system could be reached at any other given point in time if the initial conditions existed.

By the late 19th century, Poincaré showed that the temporal evolution of some systems created by the Hamiltonian equations could have chaotic motions[1]. In 1963, Lorenz demonstrated that a simple set of three linear first-order differential equations could produce completely chaotic trajectories. He found one of the first examples of algebraic chaos in dissipative systems. Chaos can be detected in a non-linear system where dynamic rules uniquely determine its temporal evolution through initial conditions[2]. In recent years, new theoretical findings along with high-speed computers and experimental results have helped us realize that nature prevents these phenomena. Furthermore, non-linearity is a necessary but not sufficient condition for chaos to occur. Chaotic motions are not observed due to external noise sources, infinite degrees of freedom of the system, or uncertainty in quantum mechanics. However, the main source of irregularity is a property of

nonlinear systems that exponentially separates initial paths that are very close to each other in the phase space[3]. Therefore, it is impossible to predict the behavior of such systems for long periods, as errors grow exponentially with the limited accuracy of the initial conditions. Lorenz called this sensitivity to initial conditions the butterfly effect because the results of equations can change by flaps of a butterfly's wings[4]. Billiards is a dynamic system studied in classic and quantum mechanics[5]. This paper aimed to analyze the dynamic system of billiards from a classic perspective. To this end, first dynamic systems were introduced and examined their properties. By analyzing the trajectories of these systems in the phase space, the classic properties of chaotic systems were introduced. Next, building on our knowledge of classic chaotic motions, the chaotic system of billiards was introduced, proposing a method for studying various motions of this system. Then different types of motion in the stadium billiards were examined. Finally, a specific billiard system was introduced and investigated its various observed motions.

2. Dynamic Systems

Newton's laws are employed to analyze dynamic systems, being the base from which describing equations of these systems are derived. However, the number of dynamic systems that can be fully analyzed to obtain explicit solutions is very limited. Most dynamic systems are non-integrable, and their behaviors must be studied through numerical methods.

Dynamic systems are characterized by two specific features: 1) the states of the system at each moment are determined by the values of N variables $x_1, x_2, ..., x_N$; 2) the evolution of the system is determined by N differential equations. In other words:

$$\frac{dx_i}{dt} = f_i(x_1, ..., x_N) \quad i = 1, ..., N \tag{1}$$

N is the order of the dynamic system, and N variables x_i, represent physical quantities such as position and velocity. If $\vec{X}$ is defined with $x_1, x_2, ..., x_N$ components and $\vec{F}$ with $f_1, f_2, ..., f_N$ components, the differential equations are written more simply as[6-9]:

$$d\vec{X}/dt = \vec{F}(\vec{X}) \tag{2}$$

This equation is completed with the following initial conditions:

$$\vec{X}(t = 0) = \vec{X}_0 \tag{3}$$

Its product would be an integral curve passing through $\vec{X}_0$. There are two states for the product of a dynamic system. First state: the overall product is explicitly written. In this case, the product is written as $X(a_1, a_2, ..., a_N, t)$, where *a*'s are integration constants. Second state: the overall product is not explicitly known, in which case, the product can be divided into two categories: First: The product is valid only within a limited time interval, such as calculating the positions of planets in the next few years, where direct numerical integration leads to the desired product. Second: The product is acceptable for a relatively long time, such as the long-term stability of the solar system. In such problems, the asymptotic behavior of the product in $t \to \infty$ is examined. The paths resulting from such problems are divided into two categories: 1) Nonreversible paths that never return to their initial position; and 2) Reversible paths that return to their initial position after a limited period[6].

2.1. Hamiltonian Systems

Hamiltonian systems are a special case of dynamic systems. The first feature of these systems is their even dimensionality, $N = 2n$ [6,7,10]. n is the number of degrees of freedom of the system and N dimensions of the phase space. The 2n variables that make up the phase space are $q_1, ..., q_n, p_1, ..., p_n$. The system is described by a 2n-dimensional function (instead of N functions in the general state), called the Hamiltonian $H(p_1, ..., q_n)$, and major differential equations for the variables are:

$$\frac{dq_i}{dt} = \frac{\partial H}{\partial p_i}, \quad \frac{dp_i}{dt} = -\frac{\partial H}{\partial q_i} \quad i = 1, ..., n \tag{4}$$

q_i and p_i are called conjugate variables. Hamiltonian is an accessible integral and can be demonstrated H fixes on a path, using equations (2). Therefore, the order of the system is reduced to 2n-1[6,7]. By introducing a cross-section for the system, the problem is reduced to studying the system in a 2n-2-dimensional space. An appropriate method for reducing the dimension of the problem under study is to first eliminate a variable like p_i using the integral H, and then define the cross-section using the equation $q_i = 0$. In this method, one pair of variables is eliminated, and n-1 pairs of conjugate variables determine the cross-section[6-13].

The motion of Hamiltonian systems can be divided into two categories: 1) Regular motion: These motions can be described using Newton's equations, such as the motion of a simple harmonic oscillator in one dimension and the motion of planets if disturbances from other planets are overlooked. In regular motion systems, paths with close initial conditions linearly diverge. 2) Irregular motion: Motions such as the motion of gas molecules when the molecules are confined to a plane and all molecules except one are fixed. In this case, a completely unpredictable motion with two degrees of freedom exists. In irregular motion systems, paths with roughly similar boundary conditions exponentially diverge and are highly sensitive to initial conditions. The difference between regular and irregular motions becomes apparent in the geometry of paths in phase space in the long-term[7,11].

2.2. Integrable Systems

Canonical transformations can help simplify the Hamiltonian concept. Using these transformations, variables $p_1,...,p_n$ and $q_1,...,q_n$ are transformed into new variables $P_1,...,P_n$ and $Q_1,...,Q_n$, and motion equations are derived from the new Hamiltonian, $H(p_1,...,q_n) \to H(P_1,...Q_n)$ If the canonical transformation is such that one of the variables does not appear in H, then the Hamiltonian will be simpler. If H does not depend on variable Q_i, then:

$$\frac{dp_i}{dt} = -\frac{\partial H}{\partial Q_i} = 0 \qquad (5)$$

As a result $P_i(t) = P_i(0) = \text{const}$. This constant value is a parameter that, if known, the Hamiltonian will depend on 2n-2 variables; meaning that there are n-1 conjugate variables and the degrees of freedom of the system decrease by two. Whenever there exists a canonical transformation under which the Hamiltonian does not depend on any Q_i s, that is: $H(P_1,...,P_n)$ then, P_i s are considered actions, Q_i s. angles,

$$P_i(t) = C_i \qquad i = 1,...,n \qquad (6)$$

$$\frac{dQ_i}{dt} = \frac{\partial H}{\partial P_i} = \omega(C_1,...,C_n) \qquad (7)$$

$$Q_i(t) = \omega_i t + D_i \qquad (8)$$

where C_i s and D_i s are 2n constants of integration. If the above conditions hold for the Hamiltonian of a system, the system is considered to be in a normal state. If a system's Hamiltonian can be brought into the normal state, the system is integrable. $P_1,...,P_n$ actions are integrals of the system and have constant values along any path. Conversely, if there are n specified integrals in a Hamiltonian system, there exists a canonical transformation whose resulting P_i s will be the integrals of the system[6-12]. For a system with a time-independent Hamiltonian, the Hamiltonian itself is an integral of the system. All systems with n=1 and a time-independent Hamiltonian are integrable[14].

3.2. Phase Space

In solving Hamiltonian equations for q and p as a function of time, given the initial conditions q_1 and p_1 at time t_1, the trajectory of the motion for any time t_2 can be determined. This p-q space is called the phase space of the system. A good way to present a dynamic system is by using phase space. Each state of the system at each point of time is presented by a point in the phase space. This point evolves concerning time and its velocity is $\vec{F}$,

whose components are determined by equation (1). The geometric location of the points corresponding to the transformation of a system forms a curve in the phase space, where the velocity vector is at a tangent at every point in time. Therefore, by drawing the velocity vector in the phase space without integration, the trajectory can be determined because the equation (2) is independent of time and the first order[6-10]. The integral curves of equation (2) create a flux in the phase space, only one of the curves being a solution to condition (3). The created flux in the phase space has the following properties: 1) The time evolution of each path is uniquely determined as a function of the initial conditions; 2) The equation (2) is also integrable in time reversal; that is, two different paths never collide; and 3) Paths limited to a boundary in a region of the phase space remain limited to the boundary over time[14].

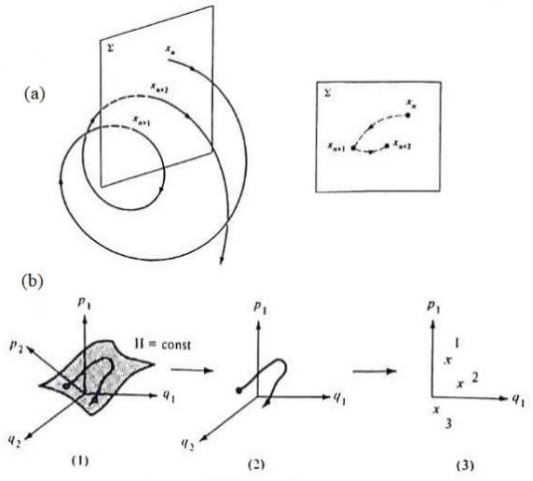


Figure 1Motion in the phase space and definition of the Poincaré cross-section: a) Points of collision of the path with the cross section; b) Motion with two degrees of freedom. 1) Four-dimensional phase space, 2) Image of the path in the volume (q_1, q_2, q_3), 3) Points of successive collisions of the path with the cross section $q_2 = \text{const}$.

2.4. Cross section and Mapping

Since the aim of this study is analyzing the long-term and asymptotic behavior of the path of the dynamic system, following the path continuously is not needed; instead, the path discretely point by point is traversed in time. This idea is based on the cross section method. If N=4 is considered as figure. 1, then a four-dimensional phase space was existed. In this space, the cross section surface is a two-dimensional plane $\sum$. The consecutive intersection points of the path with this plane are denoted as $x_1, x_2, \ldots$. Since the path is reversible, there are points $x_0, x_{-1}, \ldots$ in the return path. Using this property, if a point x_i is known, the point x_{i+1} can be determined. By following the path from the point x_i by integrating the differential equations until it collides with the plane $\sum$ again, a new point x_{i+1} is obtained. This gives us a G mapping of the plane $\sum$ to itself, called the Poincaré map. In general state:

$$G : \sum \to \sum \quad \text{and} \quad x_{i+1} = G(x_i) \tag{9}$$

in general

$$x_{i+j} = G^j(x_i) \tag{10}$$

Since the path can be followed in both time directions, the inverse mapping is defined as follows:

$$x_{i-1} = G^{-1}(x_i) \tag{11}$$

Overall, the equation (10) can be considered for positive and negative j. For any N, in an N-dimensional phase space, a N-1 dimensional subspace is considered and the collision points with $\ldots, X_{-2}, X_{-1}, X_0, X_1, X_2, \ldots$ are denoted The M=N-1 subspace should be called cross space, but for similarity with the case of N=4, called the cross section as well. In this space, the $x_1, x_2, \ldots, x_M$ coordinate system is introduced and the coordinates of each point X_i with $x_{i1}, x_{i2}, \ldots, x_{iM}$ is denoted The G mapping of this space is as follows:

$$x_{i+1,1} = g_1(x_{i,1}, \ldots, x_{i,M}), \ldots, x_{i+1,M} = g_M(x_{i,1}, \ldots, x_{i,M}) \tag{12}$$

In this method, consecutive points of X_i and overlook other details of the path are considered. These consecutive points are specified using G mapping and not using differential equations; thus, these equations are left out. Cross section and mapping methods are preferred for the following reasons: 1) the inherent properties of the dynamic system can be seen in the mapping and cross

section equations. For example, a simple periodic path that returns to the initial point after one round corresponds to a fixed point in the G mapping; that is, the periodic path is stable if and only if the fixed point is constant; $X_i = G^j(X_i)$ 2) the new problem is much simpler because instead of differential equations, mapping equations are examined. Moreover, in an N-dimensional case, the investigated space has N-1 dimensions, making theoretical and numerical studies easier; 3) the inherent properties of the system are clearly shown in the long-term behavior, but the details of the short-term evolution are omitted. Therefore, the cross section should not be used to study the system in short periods; 4) Graphical display of results is much easier. For example, N=4 state is a two-dimensional cross section and its representation is much simpler than four-dimensional space.[6,7]

2.5. Ergodic Systems

In a Hamiltonian system, the path in the phase space is confined to a fixed subspace, $H = \text{const}$. This subspace is called the energy level. In an ergodic system, each path fills its energy level and the collision points cover all surfaces in the cross sectional space[6,8,10].

2.6. Chaotic Systems

Integrability is an exceptional property for Hamiltonian systems with over two degrees of freedom. Integrable systems are so rare that it is impossible to approximate a non-integrable system with a series of integrable ones[15]. Therefore, in most problems, numerical methods are used to obtain the solution of Hamiltonian equations, where small changes in initial conditions lead to significant changes in the obtained solutions. There is a class of dynamic systems where the particle passes through every point in the phase space. These systems are called chaotic systems. In a chaotic system, small changes in initial conditions cause paths to exponentially diverge, whereas paths diverge linearly in integrable systems. In the cross section of chaotic systems, chaotic regions can be seen, which are separated from each other by invariant and regular curves. However, the presence of these empty areas in the cross section does not contradict the ergodicity of the system, as these areas disappear with the mapping for $N \to \infty$, and the entire cross section surface is filled[6].

2.7. Liapunov Exponent

In a chaotic motion, the mapping points $x_{n+1} = G(x_n)$ diverge exponentially. The Liapunov exponent defines this divergence. As shown in the Figure 2:

$$\varepsilon \exp\left(N\lambda(x_0)\right) = \left|G^N(x_0+\varepsilon) - G^N(x_0)\right| \qquad (13)$$

$$\lambda(x_0) = \lim_{N\to\infty}\lim_{\varepsilon\to\infty}\frac{1}{N}\log\left|\frac{G^N(x_0+\varepsilon) - G^N(x_0)}{\varepsilon}\right| \qquad (14)$$

$$\lambda(x_0) = \lim_{N\to\infty}\frac{1}{N}\log\left|\frac{dG^N(x_0)}{dx_0}\right| \qquad (15)$$

That is, $\exp(\lambda(x_0))$ is the average amount by which the distance between two neighboring points changes after one iteration[16]. Depending on the value of $\lambda(x_0)$, there are three different situations: 1) If it is positive, the two paths diverge exponentially; 2) If it is negative, the two paths converge; and 3) If it is zero, the distance between the two paths remains constant[17].

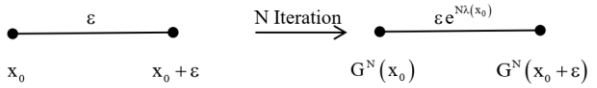


Figure 2 Liapunov Exponent

3. Billiards

Billiards is an important class of dynamic systems. By definition, billiards is a dynamic system with a closed environment (usually two-dimensional) that includes a free and dimensionless particle [6,11,15,16,18]. In billiards, the particle undergoes elastic collision with the boundary, therefore, energy conservation results velocity conservation. The particle behaves like a geometric ray with uniform angle of incidence and reflection at each collision point. A small group of billiards has interesting features. In this group, all velocity components are reversed at the reflection point. This type of reflection is Andreev reflection and billiards with this reflection is called Andreev billiards. Andreev billiards are of interest in

condensed matter physics, especially in superconductivity [19]. To examine billiards, they are classified into two categories: 1) Integrable billiards: systems with n motion constants like rectangular and circular billiards; 2) Non-integrable billiards: systems where only energy conservation holds. These types of billiards are ergodic and chaotic systems [10]. Sinai and Bunimovich billiards are examples of these billiards. By defining billiards in a 2D state, Sinai proved that billiards is an ergodic and chaotic system and the moving particles of billiards behave chaotically[18]. Bunimovich also defined billiards with two degrees of freedom called the stadium billiards and proved that the stadium is an ergodic and chaotic system [20,21].

3.1. Birkhoff Mapping

In billiards, the movement of the particle starts from a point on the boundary, and the path is extended according to the differential equations describing the path until the particle hits the boundary at another point and reflects. This point is the end of the previous path and the beginning of the new one. Therefore, with the coordinates of this point, the path of the motion can be determined and it seems that the billiard wall is a suitable cross section to describe the motion of the particle. To investigate the behavior of the Hamiltonian billiard system, the state of the system with the coordinates of the reflection points on the boundary and in the direction of the motion is described. Thus, the phase space is determined as a mapping between consecutive and discrete collisions. The location of the reflection point on the billiard boundary is determined by the length of the arc S along the boundary and the direction of the motion after reflection by the angle Ψ, the angle of the velocity vector, $\vec{V}$, and the tangent vector on the boundary. If the total length of the boundary is L, the length of the arc is normalized to $s=\frac{S}{L}$. The conjugate coordinate s, is the tangential component of momentum, $P=\cos\Psi$. If the magnitude of the velocity is considered 1 in the selection of units, $0\le\psi\le180^{\circ}$ then the cross-sectional area is limited to a rectangle with the following dimensions:

$$-1\le p\le 1 \quad , \quad 0\le s\le 1 \tag{16}$$

s and p are defined coordinates of Birkhoff and the defined mapping, called the Birkhoff mapping, which is a class of Poincaré mappings[22,23,24]. Paths obtained are divided into three categories: 1) A finite set of N points, $(s_0,p_0),(s_1,p_1),...,(s_{N-1},p_{N-1})$, is obtained, which correspond to a closed path, and since for each closed path, there is:

$$(s_{n+N},p_{n+N})=M^N(s_n,p_n)=(s_n,p_n) \tag{17}$$

each of these N points is a fixed point of the mapping; 2) Repeating (s_0,p_0) forms a smooth and well-behaved curve in the phase space. Thus, the curve is called an invariable curve because, under the M mapping, each point on the curve returns to a point on the curve. This behavior can be seen in integrable systems, where there is a motion constant as a function $F(s,p)$ that: $F(s_0,p_0)=F(s_1,p_1)$ and each invariant curve is a contour of $F(s,p)$; 3) The repetition of (s,p) fills a certain level in the phase space, and in this case, the path is not limited by any constant quantity and is very sensitive to initial conditions (s_0,p_0). All three types of paths are observed in the study of a billiard dynamic system[19,23]. By comparing the solution method of differential equations and numerical integration and the mapping method, it becomes apparent that using the mapping method, the volume of computations and errors decreases significantly. Therefore, studying billiards using mapping is a suitable method for investigating dynamic systems.

3.2. Stadium

Stadium is an example of non-integrable and chaotic billiards (Bunimovich billiards). The stadium consists of two semicircles with the same radius, which are separated from each other by two parallel line segments of length a. The parameter $\eta=\frac{a}{R}$ is called the characteristic parameter of the stadium. Due to the above change, a stadium is

perturbed by circular billiards, which turns into a stadium for non-zero values of η [16,20,23]. In the stadium, in addition to observing chaotic motions, in some of the initial conditions, completely regular and predictable motions are observed, which may disappear and be replaced by chaotic motions upon a small change in the initial conditions.

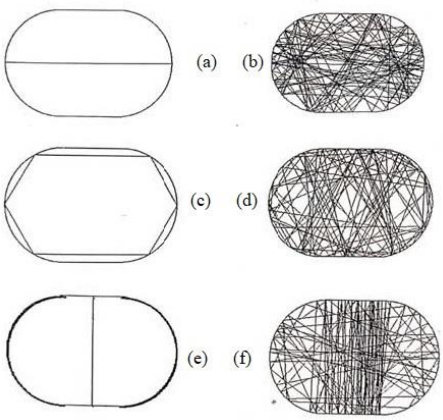


Figure 3 Regular and irregular motion in stadium. (a) Singular regular motion $\psi_0 = 90^\circ$,(b) chaotic motion $\psi_0 = 89^\circ$, (c) Singular regular motion $\psi_0 = 30^\circ$,(d) chaotic motion $\psi_0 = 29^\circ$, (e) Non-singular regular motion $\psi_0 = 90^\circ$,(f) chaotic motion $\psi_0 = 89^\circ$,

Figure 3 shows three examples of regular motion states of a particle in the stadium. With a small change in the initial conditions, the trajectory undergoes a major change and covers the entire surface of the stadium. Regular motions are divided into two categories: 1) Singular regular motion: there is a regular path corresponding to certain initial conditions, figure 3(a) and (c); 2) Non-singular regular motion: there is a set of regular trajectories corresponding to certain initial conditions, figure 3(e). Singular and non-singular paths are both unstable, and a small change in the location or momentum can cause the path to exponentially diverge from the regular state and cover the entire energy surface[10]. In examining the cross section area of the motion of the particle in the stadium using Birkhoff mapping, the obtained cross-sectional area can be seen in two forms: 1) The collision points lie on a one-dimensional curve shown in figure 4(a), which corresponds to the regular motion. 2) The collision points cover a two-dimensional surface that corresponds to the irregular motion of the particle in the stadium, shown in figure 4(b).

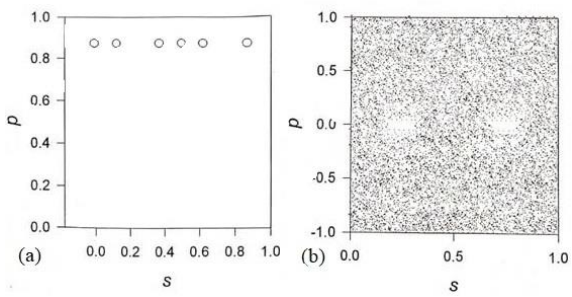


Figure 4 Collision cross section (a) regular, (b) irregular motion

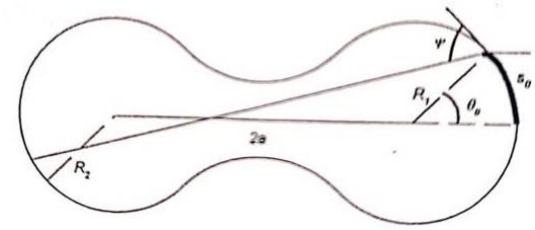


Figure 5 deformed stadium

3.3. Deformed Stadium

To better investigate the chaotic behavior of billiards, the effect of changing various parameters and changing the shape of the billiard boundary, a deformed stadium that specifies two arcs of a circle representing the boundary instead of two line segments (dumbbell shape) is introduced. To investigate this problem from a classic point of view, the free motion of a particle in billiards is considered. Using the Birkhoff map, points of collision of the particle with the boundary are determined and the cross section is plotted. The path corresponding to the cross section at the billiard level can also be determined. For this purpose, collision points are obtained using repeated calculations of obtained mapping equations. This eliminates the need for solving differential equations, integration, or dealing with errors resulting from these stages. To calculate the collision points of the particle with the billiard boundary, the relationship between p_n and p_{n+1} or (x_n, y_n) and (x_{n+1}, y_{n+1}) can be determined. Assuming the starting point of the movement is

point (x_0, y_0) at Cartesian coordinates and (s_0, p_0) in the mapping coordinates, where $p_0 = \cos\psi_0$. Given the coordinates of (x_0, y_0) θ_0 can be defined as the polar angle relative to the positive x-axis, shown in figure 5. The motion equations for the particle are as follows,

$$x = x_0 + v_{0x}t \quad , \quad y = y_0 + v_{0y}t \tag{18}$$

where the velocity $v = \left(v_{0x}^2 + v_{0y}^2\right)^{1/2}$ is a constant value. By removing the time parameter from the equations (18), the path equation is obtained as follows:

$$y = y_0 + (x - x_0)\frac{v_{0y}}{v_{0x}} \tag{19}$$

quantity v_{0y}/v_{0x} represents the slope of the particle's path. Therefore, for two consecutive points, there are:

$$y = y_0 + (x - x_0)\tan\beta \quad , \quad \beta = \frac{\pi}{2} + \theta_0 + \psi_0$$

$$x = x_0 - (y - y_0)\tan(\theta_0 + \psi_0) \tag{20}$$

for two points in a row

$$x_{n+1} = x_n - (y_{n+1} - y_n)\tan(\theta_n + \psi_n) \tag{21}$$

for the collision point on one of the two arcs, the y_{n+1}-coordinate boundary is determined by calculating the collision point of the path with the equation of boundary, and for the collision point on the two ends:

$$y_{n+1} = R_1 \sin\theta_{n+1} \quad , \quad y_{n+1} = R_2 \sin\theta_{n+1} \tag{22}$$

using the above equations, the mapping equations are obtained and the coordinates and collision points are determined by repetitive computer calculations. The particle's trajectory and the collision cross-section are thus determined without solving differential equations describing the path.

4. Results

Figure 6(a) depicts the trajectory of a regular motion, whereas figure 6(b) displays the trajectory of a chaotic motion resulting from changes in the initial conditions. The cross-sectional surfaces of these two motions are also shown in figure 7. The shape of the cross-section of the regular motion shows that the four points of contact with the boundary are located on a line in the phase space. Figure 7(a) presents the cross-sectional surface of a chaotic motion, where the surface is covered with collision points and the empty spaces are surrounded by invariant curves. These spaces are also filled in case of $n \to \infty$ and they eventually disappear. In Figure 8, 9, 10 and 11, the cross-section of the motion are determined with different initial conditions, from which the following results can be obtained: 1) As shown in

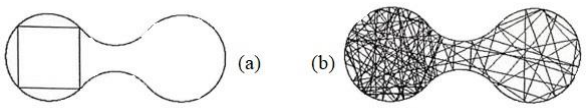


Figure 6 (a) regular motion $\psi = 45^\circ, n = 100$ (b) chaotic motion $\psi = 46^0, n = 100$

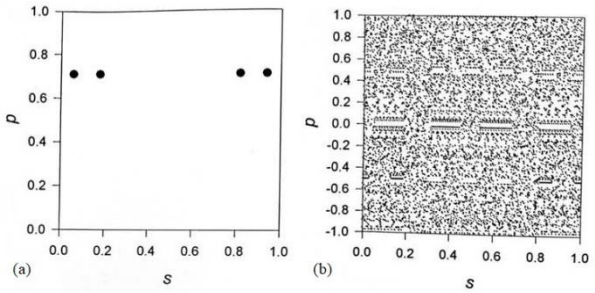


Figure 7 (a) cross section of 6(a), (b) cross section of 6(b).

figures 6 and 7, like the stadium, for some initial conditions, regular paths are obtained, and with a slight change in the initial conditions, these paths disappear, and irregular paths fill the entire billiard surface, covering the entire cross-sectional surface in phase space; 2) In contrast to the stadium case, where there are only two empty areas around two fixed points,

$$\left(s = \frac{3}{4}, p = 0\right) , \left(s = \frac{1}{2}, p = 0\right)$$

which correspond to the non-singular motion perpendicular to two line segments, in the deformed stadium, the number and size of the empty areas in the cross-section depend on the parameters defining the shape of the boundary. For

example, with the increase of the angle α, the angle of the perpendicular line to the point of contact of two parts of the boundary with the positive direction of the axis y, the number of empty areas increases and they are obtained around points,

$$\left(\frac{1}{2}+\frac{\pi R_2}{2L},0\right),\left(\frac{\pi R_1}{2L},0\right),\left(1-\frac{\pi R_1}{2L},0\right),...$$

which are surrounded by invariant curves.

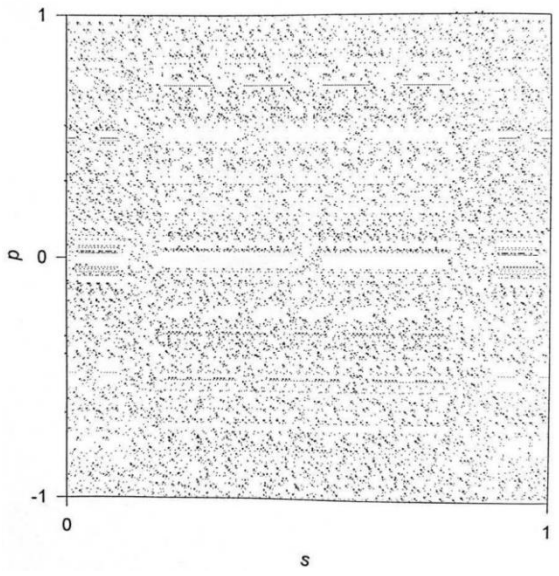


Figure 8 chaotic motion in the deformed stadium for initial conditions $s=0.1,\Psi=30^{\circ},\alpha=55^{\circ}$.

Conclusions

A deformed stadium that specifies two arcs of a circle representing the boundary instead of two line segments (dumbbell shape) is introduced to investigate the chaotic behavior of billiards. The results show:

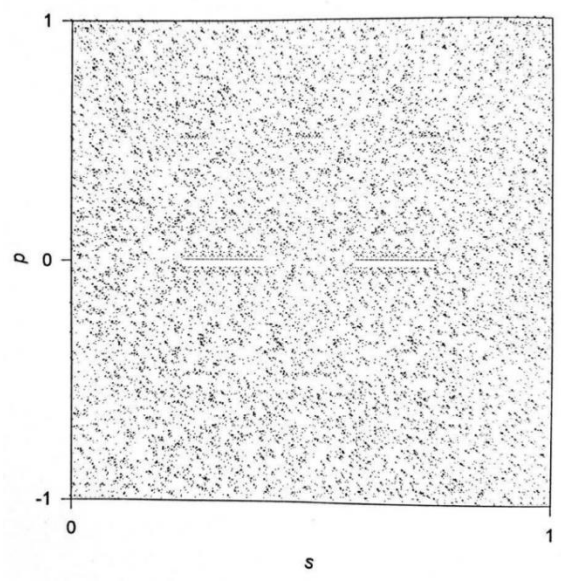


Figure 9 chaotic motion in the deformed stadium for initial conditions $s=0.1,\Psi=30^{\circ},\alpha=10^{\circ}$.

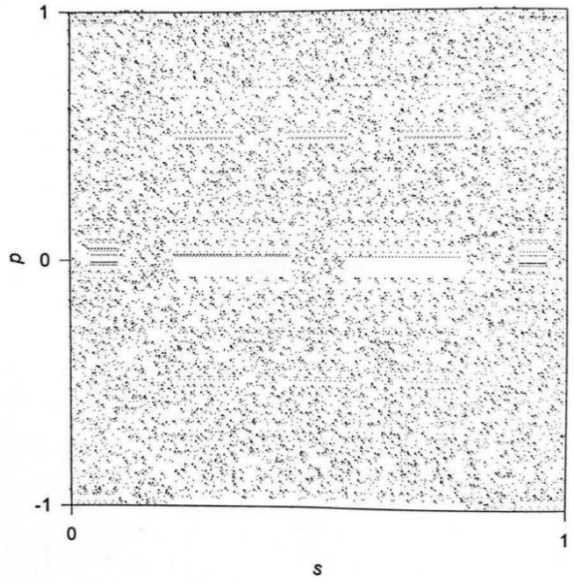


Figure 10 chaotic motion in the deformed stadium for initial conditions $s=0,\Psi=45^{\circ},\alpha=35^{\circ}$.

. The shape of the cross-section of the regular motion is located on a line in the phase space.

1) The cross-sectional surface of a chaotic motion is covered with collision points and the empty spaces are surrounded by invariant curves.
2) The empty spaces are also filled in case of $n\to\infty$ and they eventually disappear.
3) Like the stadium, for some initial conditions, regular paths are obtained, and with a slight change in the initial

conditions, these paths disappear, and irregular paths fill the entire billiard surface, covering the entire cross-sectional surface in phase space

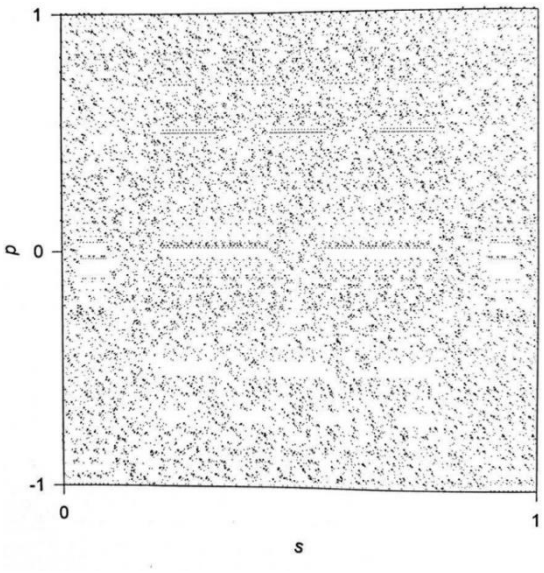


Figure 11 chaotic motion in the deformed stadium for initial conditions $s = 0.5, \Psi = 40^{\circ}, \alpha = 35^{\circ}$

4) In contrast to the stadium case, where there are only two empty areas around which correspond to the non-singular motion perpendicular to two line segments, in the deformed stadium, the number and size of the empty areas in the cross-section depend on the parameters defining the shape of the boundary.

The observation of chaotic motions in classic mechanics prompts the question of how this randomness manifests in quantum mechanics. To find the answer to this question, the wave equation for these systems shall be studied in the following.

References

[1] H. Poincaré, "New methods of celestial mechanics" *National Aeronautics and Space Administration*, 1967.
[2] J. Gleick, "Chaos: Making a new science" *Penguin* 2008.
[3] S. H. Strogatz, "Nonlinear dynamics and chaos: with applications to physics, biology, chemistry, and engineering" *CRC press* 2018.
[4] E. N. Lorenz, "Deterministic nonperiodic flow" *Journal of atmospheric sciences*, Vol *20*(2), pp.130 1963.
[5] E. Ott, "Chaos in dynamical systems" *Cambridge university press* 2002.
[6] R. C. Hilborn, "Chaos and Nonlinear Dynamics: An Introduction for Scientists and Engineers" *Oxford University Press*, 2000.
[7] J. R. C. Piqueira, C. E. N. Mazzilli, C. P. Pesce, and G. R. Franzini, "Lectures on Nonlinear Dynamics" *Springer International Publishing AG*, 2024.
[8] T. Huang, L. Dai, and H. Zhang, "An approach combining periodicity ratio and secondary Poincaré map for characteristics diagnosis of nonlinear oscillatory systems" *Nonlinear Dynamics*, vol 84, pp. 959, 2016.
[9] A. Shahhosseini, M. H. Tien, and K. D'Souza "Poincare maps: a modern systematic approach toward obtaining effective sections" *Nonlinear Dynamics*, vol 111(1), pp. 529 2023.
[10] A. Lichtenberg, A. M. Liberman, "Regular and stochastic Motion" *Springer-Verlag*, New York, 1983.
[11] M. Berry, "Semiclassical Mechanics of Regular and Irregular Motion" *North-Holland* 1983.
[12] H. Goldstein, C. Poole and J. Safko, "Classical Mechanics" *Addison-Wesley*, 2001.
[13] M, C. Gutzwiller, "Chaos in Classical and Quantum Mechanices" *Springer-Verlag*, 1990.
[14] B. L. Hao, "Directions in Chaos" *World Scientific*, Vol. 1, 1987.
[15] B. L. Hao, "Chaos II" *World Scientific*, 1990.
[16] H. G. Schuster and W. Just, "Deterministic Chaos: An Introduction 4th Revised and Enlarged Edition" *Wiley-VCH*, 2005.
[17] H. Haken, "At least one Lyapunov exponent vanishes if the trajectory of an attractor does not contain a fixed point." *Phys. Lett. A.* vol. 94(2), pp. 71-2, Feb 1983.
[18] Y. G. Sinai, "Dynamical systems with elastic reflections." *Russ. Math. Surv.* Vol. 25(2), pp. 137 Apr 1970.
[19] I. Kosztin, D. L. Maslov and P. M. Goldbart "Chaos in Andreev billiards." *Phys. Rev. Let.* vol. 75(9), pp. 1735 Aug 1995.
[20] L. A. Bunimovich, "On ergodic properties of certain billiards." *Functional Analysis and Its Applications.* Vol. 8(3), pp. 254-5, Jul 1974.
[21] L.A. Bunimovich, "On the ergodic properties of nowhere dispersing billiards" *Commun. Math. Phys.* vol. 65, pp. 295-312, Oct 1979.
[22] Z. Shahriari, S. D. Algar, D. M. Walker, and M. Small "Ordinal Poincaré sections: Reconstructing the first return map from an ordinal segmentation of time series" *Chaos: An Interdisciplinary Journal of Nonlinear Science*, vol 33(5) 2023.
[23] M. V. Berry, "Regularity and chaos in classical mechanics, illustrated by three deformations of a circular billiard" *Eur. J. Phys.* Vol. 2(2), pp. 91 Apr 1981.
[24]W. Schlei, K. C. Howell, X. Tricoche, and C. Garth,. "Enhanced visualization and autonomous extraction of poincaré map topology". *The Journal of the Astronautical Sciences*, Vol 61, pp.170, 2014.

Journal of Optimization of Soft Computing (JOSC)

Vol. 2, Issue 2, pp: (18-22), Summer-2024

Journal homepage: http://josc.rimag.ir/en

Research Paper

Simulation of crack influence on the free vibration of a rectangular plate using the finite element method (FEM)

Ahmad Haghani[1,2*] and Soleyman Esmaeil zadeh[3]

1. Department of Mechanics, Faculty of Engineering, Shahrekord Branch, Islamic Azad University, Shahrekord, Iran.
2. Energy and Environment Research Center, Shahrekord Branch, Islamic Azad University, Shahrekord, Iran.
3. M.Sc. Student in Mechanical Engineering, Faculty of Engineering, Shahrekord Branch, Islamic Azad University, Shahrekord, Iran.

Article Info

Article History:
Received: 2024/07/24
Revised: 2024/08/31
Accepted: 2024/09/13

DOI:

Keywords:
Free Vibration, Crack, Natural frequency.

**Corresponding Author's Email Address:a.haghani@iaushk.ac.ir*

Abstract

Today, sheet metals are extensively utilized in various industries as one of the most crucial components. The presence of a crack in a structural element reduces local stiffness and consequently weakens the structure's resistance. Any change in local stiffness affects modal factors including mode shapes, natural frequencies, and structural damping. A major challenge in structural health monitoring is identifying the severity and location of potential cracks. Continuous evaluation is essential to ensure the proper functioning of many structures. This study presents an engineering perspective on the influence of cracks on vibration frequencies considering crack dimensions and locations. Finite element simulations, a widely accepted computational tool, were employed for this investigation. After verifying the convergence of the solution method, the simulation results were compared with those found in other sources, showing good agreement. Finally, the impact of crack orientation and position on the natural frequencies of the system was analyzed.

1. Introduction

The occurrence of sudden failures due to crack growth in structures has always been a challenging and investigated topic. When a part of a structure is damaged by a crack, the stiffness in that area decreases, consequently increasing the natural period of the structure and reducing its natural vibration frequency. These damages can also lead to changes in mass distribution and structural damping characteristics. Such defects predominantly affect regions near cracks under severe stress concentration factors, diminishing gradually as they move away from the crack. Many researchers have proposed methods for determining the location and characteristics of cracks, as well as understanding how these damages propagate.

The dynamic behavior of cracked structures has been extensively studied using various mathematical, numerical, and experimental methods. Much of this research has focused on modeling cracks in plates under different boundary conditions. Some notable studies investigating the properties and effects of cracks on the mechanical characteristics of plates include:

Xiong et al. [1] analyzed the path of actual crack propagation and changes in resonance frequencies under intensified conditions for a plate. They initially proposed a simulation analysis method for crack propagation and validated their proposed method through crack propagation experiments. Finally, they studied the relationship between crack propagation length and resonance frequencies.

Wang et al. [2] developed a nonlinear dynamic model for thin cylindrical shells prone to crack under long-term loading and external impact, using partial Fourier transformation and residual theorem

to examine nonlinear forced vibrations in a cracked cylindrical shell.

Wu et al. [3] proposed a new model for a breathing crack with axial bending (ABCBCM) for rotating blades. They derived the governing equations based on Timoshenko beam theory and Castigliano's principle, solved them using the proposed model, and then validated the results with FEM and experimental tests. The findings indicated that the axial vibration reaction of the blade is more sensitive to the nonlinearity caused by the breathing crack compared to the bending response.

Tho et al. [4] applied the third-order shear deformation theory to simulate the free vibration behavior and static bending of multilayer composite plates containing fractures in the core layer. They showed that with changes in crack dimension, the natural frequency and the highest displacement of the plate do not change significantly.

Hu et al. [5] investigated and provided new analytical solutions for the vibration behavior of robust rectangular plates in free conditions with edge cracks. Finally, they presented the natural frequency results for different vibration modes of thick plates with edge cracks and examined the high accuracy and fast convergence of the solutions.

Khoram-Nejad et al. [6] investigated and analyzed the free vibration of a cracked FGM plate under uniaxial compressive load. They obtained the nonlinear differential equations of motion using the Mindlin plate theory for an imperfect primary plate and solved them using the differential quadrature method. The results were in strong agreement with those obtained from the FEM analysis.

Taima et al. [7] examined the lateral vibration of cracked thick isotropic beams using Timoshenko beam theory and the third-order shear deformation theory. The results indicate that the discrepancy between the analytical and experimental findings is minimal, which confirms the validity of the solution.

Wu et al. [8] evaluated the simulation of crack growth in curved steel tensile specimens using cohesive zone modeling.

Citarella and Giannella [9] examined advanced numerical approaches for crack growth simulation. Additionally, Alshoaibi [10] analyzed fatigue crack spread under uniform amplitude loading by the FEM.

Singh et al. [11] investigated the simulation of crack growth in an FGM plate by extended FEM.

In this article, the effect of crack position and size on the natural frequency of a simply supported plate is analyzed through simulation. In this simulation, a single element is used to investigate the stress and its concentration at the crack tip. Then, the convergence and independence of the solution from the mesh are examined, and then validate the results with other references. Finally, the results of the simulation are presented.

2. Geometry, Boundary Conditions, and Mechanical Properties

Figure 1 shows the geometry of the cracked plate under consideration. As shown in this figure, the crack is at the edge of the plate, and its position is specified by three parameters: a, c, and θ.

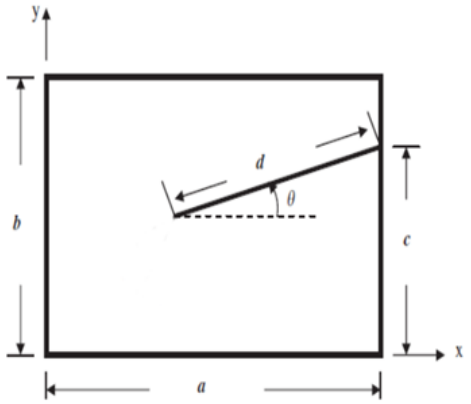


Figure 2. Geometry of the cracked plate.

In this study, the plate boundary conditions are assumed to be simply supported, as shown in Figure 2. It is also assumed that the plate is square with a side length of 0.1 meters and a thickness of 1 millimeter.

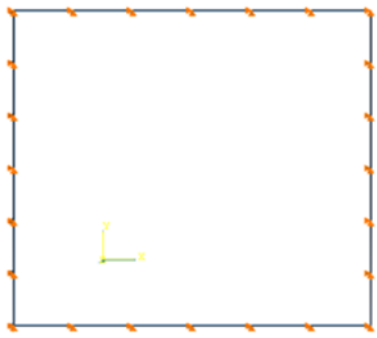


Figure 2. Boundary conditions of the plate.

The elastic properties of the steel utilized in the present work are shown in Table 1.

Table 1. Steel Elastic Properties

Elastic Properties	value
E (GPa)	204
ν	0.3
ρ (kg/m^3)	7860

3. Stress Intensity Factor (SIF)

A significant crucial factor that should typically be considered in crack analysis is the SIF. In fact, the SIF represents the crack propagation resistance of the material. Fracture has three modes: opening, sliding, and tearing. For example, the SIF in the opening mode is calculated using the following equation [12].

$$K_I = \sigma\sqrt{\pi a} \quad (1)$$

In this equation, K_I is the SIF in the first mode of fracture, σ is the stress, and a is the length of crack.

4. Convergence and Validation of the Solution

To ensure that the problem is not sensitive to the number of elements, the SIF in the first mode is calculated for different numbers of elements. To non-dimensionalize the SIF in the first mode, it is sufficient to divide this factor, calculated by the finite element software, based on the right-side expression in Eq. (1). Table 2 shows the changes of the non-dimensional SIF in the first mode with respect to the number of elements. As indicated in the table, the dimensionless SIF in the first mode experiences negligible change with increasing the number of elements to 1.6 million elements, and therefore, the problem is not dependent on the number of elements and is convergent.

Table 2. Sensitivity of the Problem to the Number of Elements

non-dimensional SIF	Number of Elements (Millions of Elements)
0.094	0.8
0.185	1
0.186	1.3
0.188	1.6
0.188	2

To validate the solution, the results obtained from the simulation for the elastic properties stated in Table 1 and the geometry shown in The results in Figure 3 are contrasted with those from Ref. [13]. The findings of this comparison are presented in Table 3. The natural frequency is non-dimensionalized by dividing the frequency of the cracked plate by that of the uncracked plate. As shown in the table, the findings are in close agreement with those of Ref. [13].

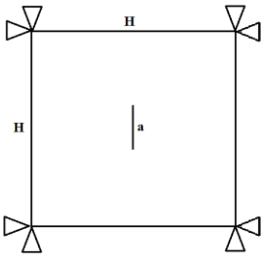


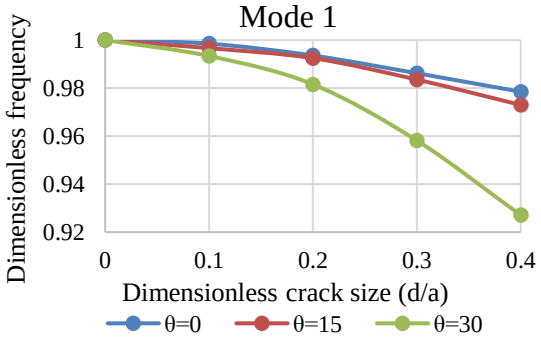
Figure 3. Plate with Simply Supported Boundary Condition and Vertical Central Crack.

Table 3. Validation of solution

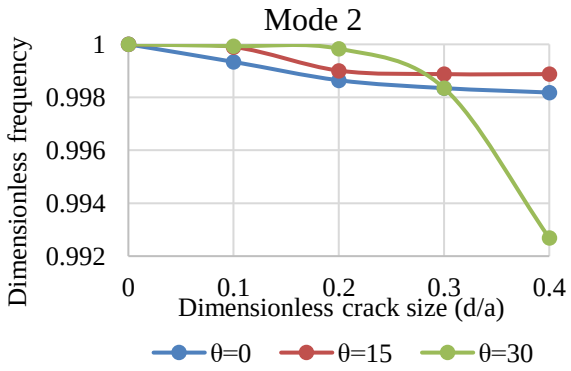
Nondimensionalized First Mode Natural Frequency	2a/H	
	0.1	0.2
Ref. [13]	0.9942	0.9806
Present work	0.9982	0.995
% Difference	0.4	1.45

5. Results and Discussion

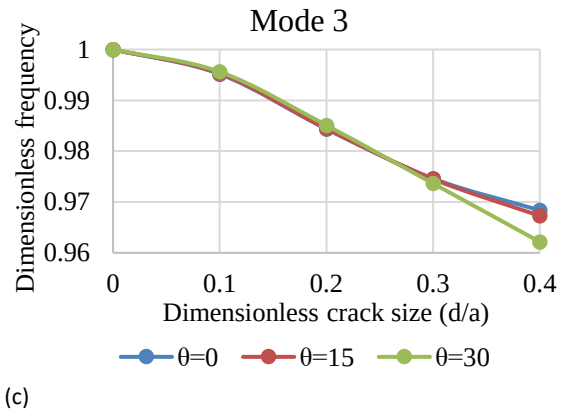
The natural frequencies results obtained from the FEM for the first five vibration modes, for different d/a ratios at crack angles of 0, 15, and 30 degrees, are presented. These results were calculated for c/a=0.5, and Fig. 4 shows a graph of these results. As indicated by the figure, in the first, second, and third modes, the frequency reduction is more pronounced at a 30-degree angle as the crack length increases. In contrast, in the fourth and fifth modes, the frequency reduction is more noticeable at a 0-degree angle with increasing crack length.



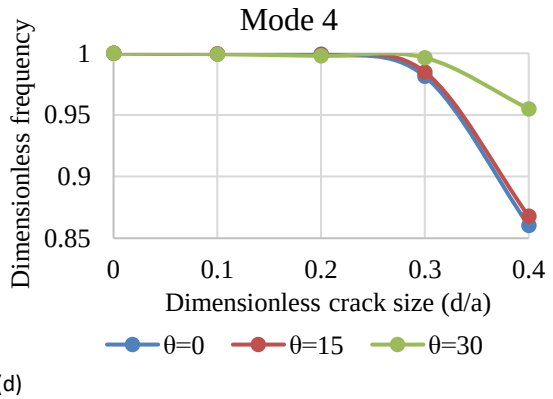
(a)



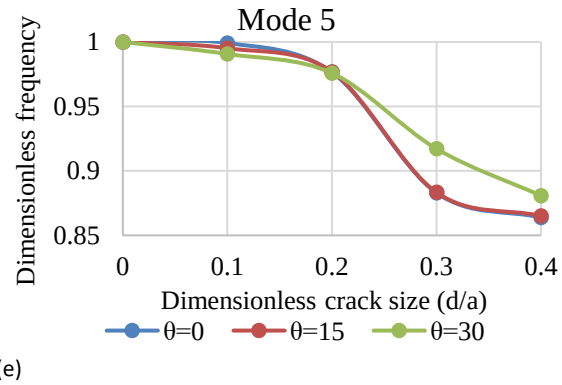
(b)



(c)



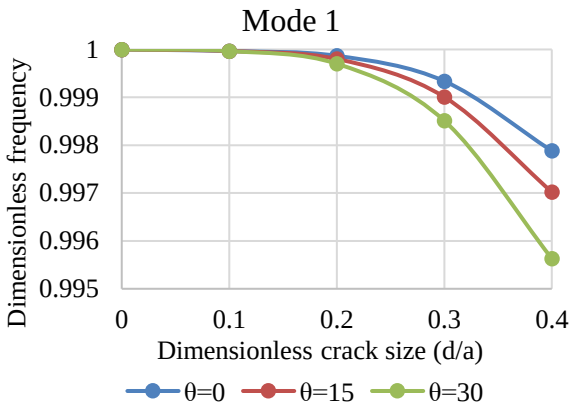
(d)



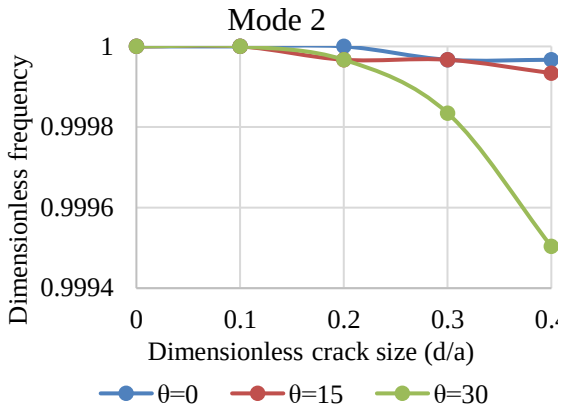
(e)

Figure 4. Effect of Changing the *d*/*a* Ratio on Natural Frequency: a) First, b) Second, c) Third, d) Fourth, and e) Fifth Modes at Angles of 0, 15, and 30 Degrees for an Edge Crack with Angle ϑ.

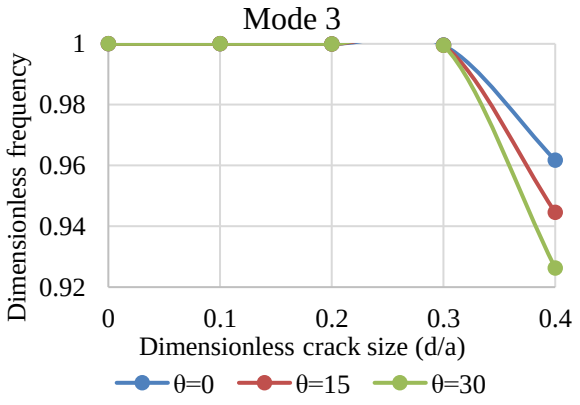
Figure 5 shows the impact of a central crack on the first five vibration frequencies of the plate. These results are calculated for a *c*/*a*=0.5 ratio and different crack lengths at angles of 0, 15, and 30 degrees. As observed in the figure, for the first through fourth modes, the frequency reduction is more pronounced at a 30-degree angle as the length of the crack grows, whereas in the fifth mode, the frequency reduction is more noticeable at a 0-degree angle with increasing crack length.



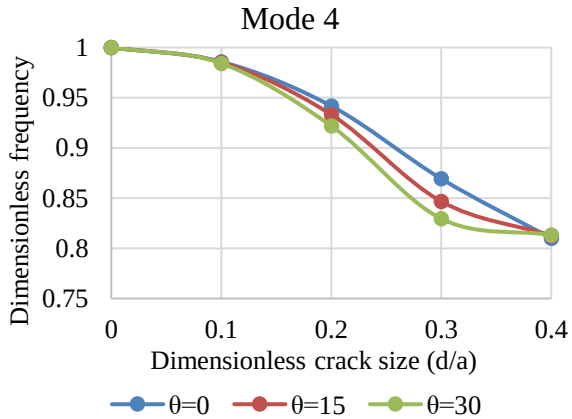
(a)



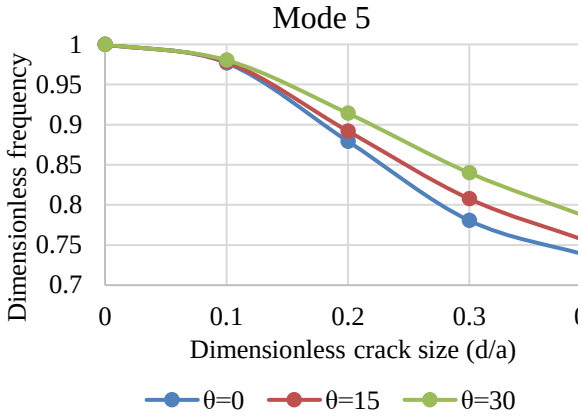
(b)



(c)



(d)



(e)

Figure 4. Effect of Changing the d/a Ratio on Natural Frequency: a) First, b) Second, c) Third, d) Fourth, and e) Fifth Modes at Angles of 0, 15, and 30 Degrees for an Edge Crack with Angle ϑ.

From observing Figures (4) and (5), the data suggest that increasing crack length leads to a reduction in natural frequency. This happens due to the presence of a crack reduces the stiffness of the plate, leading to a decrease in the natural frequency.

6. Conclusion

In this study, the free vibrations of a square plate with edge and central cracks were investigated. The simulation was performed using the Abaqus FEM software, and after examining mesh independence, the solution was validated, showing strong agreement with the findings from other references. The findings from this study are as follows:

- Based on the results, among the crack parameters, the crack dimension has the most important impact on frequency reduction. Additionally, the crack length has the most substantial impact on the first and fifth frequencies.
- As the crack angle decreases, the frequency also decreases. The reduction in frequency is more noticeable in the first mode compared to other frequencies.
- The closer the crack is to the center of the plate, the lower the natural frequency. Therefore, when the crack is at the center and in the longitudinal direction of the plate, the natural frequency is more significantly reduced.
- The natural frequency changes depending on the crack location in specific modes. These modes can be considered as patterns that are continuously monitored to detect cracks in the plate.

References

[1] Q. Xiong, H. Guan, H. Ma, Z. Wu, J. Zeng, W. Wang and H. Wang, "Crack propagation and induced vibration characteristics of cracked cantilever plates under resonance state: Experiment and simulation," *Mechanical Systems and Signal Processing*, vol. 201, p.110674, 2023.

[2] T. Wang, C. Wang, Y. Yin, Y. Zhang, L. Li and D. Tan, "Analytical approach for nonlinear vibration response of the thin cylindrical shell with a straight crack," *Nonlinear Dynamics*, vol. 111(12), pp.10957-10980, 2023.

[3] Z.Y. Wu, H. Yan, L.C. Zhao, G. Yan, Z.B. Yang, H.F. Hu and W.M. Zhang, "Axial-bending coupling vibration characteristics of a rotating blade with breathing crack," *Mechanical Systems and Signal Processing*, vol. 182, p.109547, 2023.

[4] N.C. Tho, P.H. Cong, A.M. Zenkour, D.H. Doan, and P. Van Minh, "Finite element modeling of the bending and vibration behavior of three-layer composite plates with a crack in the core layer," *Composite Structures*, vol. 305, p.116529, 2023.

[5] Z. Hu, Z. Ni, D. An, Y. Chen and R. Li, "Hamiltonian system-based analytical solutions for the free vibration of edge-cracked thick rectangular plates," *Applied Mathematical Modelling*, vol. 117, pp.451-478, 2023.

[6] E.S. Khoram-Nejad, S. Moradi and M. Shishehsaz, "Effect of crack characteristics on the vibration behavior of post-buckled functionally graded plates," *Structures*, vol. 50, pp. 181-199, 2023.

[7] M.S. Taima, T.A. El-Sayed, M.B. Shehab, S.H. Farghaly and R.J. Hand, "Vibration analysis of cracked beam based on Reddy beam theory by finite element method," *Journal of Vibration and Control*, vol.29(19-20), pp.4589-4606, 2023.

[8] S. Wu, J. Pan, P.S. Korinko and M.J. Morgan, "Simulations of Crack Extensions in Arc-Shaped Tension Specimens of Uncharged and Tritium-Charged-and-Decayed Austenitic Stainless Steels Using Cohesive Zone Modeling," In *Pressure Vessels and Piping Conference*, USA, 2020.

[9] R. Citarella and V. Giannella, "Advanced Numerical Approaches for Crack Growth Simulation," *Applied Sciences*, vol. 13(4), p.2112, 2023.

[10] A.M. Alshoaibi, and A.H. Bashiri, "Adaptive finite element modeling of linear elastic fatigue crack growth," *Materials*, vol. 15(21), p.7632, 2022.

[11] A.P. Singh, A. Tailor, C.S. Tumrate and D. Mishra, "Crack growth simulation in a functionally graded material plate with uniformly distributed pores using extended finite element method," *Materials Today: Proceedings*, vol. 60, pp.602-607, 2022.

[12] E.E. Gdoutos, "Fracture mechanics: an introduction," In *Springer*, 3rd ed., Switzerland AG 2020, ch. 2, p. 15.

[13] M. Krawczuk, A. Żak, and W. Ostachowicz, "Finite element model of plate with elasto-plastic through crack," *Computers & Structures*, vol. 79(5), pp.519-532, 2001.

Journal of Optimization of Soft Computing (JOSC)

Vol. **2**, Issue **2**, pp: (23-35), **Summer**-2024

Journal homepage: http://josc.rimag.ir/en

Research paper

Key Pre-distribution Based on Block Complementation Design in Internet of Things Security

Vahid Chegeni[1*], Hamid Haj Seyyed javadi[2], Mohammad Reza Moazami Goudarzi[3]

1. Department of Computer Engineering, Khorramabad Branch, Islamic Azad University, Khorramabad, Iran
2. Department of Mathematics and Computer Science, Shahed University, Tehran, Iran
3. Department of Mathematics, Borujerd Branch, Islamic Azad University, Borujerd, Iran

Article Info

Article History:
Received: 2024/08/12
Revised: 2024/09/18
Accepted: 2024/09/22

DOI:

Keywords:
Internet of Things (IoT), Combinatorial Designs, Residual Design, Block Complementation, key management, cryptography.

**Corresponding Author's Email Address: vahid.chegeni@iau.ac.ir*

Abstract

The Internet of Things is a network of smart devices that can connect and exchange data with other things. Due to the heterogeneous nature of IoT devices and constrained resources, creating a secure connection between IoT devices is very important. The use of previous algorithms for encryption, such as RSA and AES, involves complex and heavy computation and is unsuitable. Therefore, lightweight encryption methods are required. This paper presents a new and essential pre-distribution scheme proposed to attain high security. This scheme is based on a design derived from combinatorial algebra, namely the residual design. According to this scheme, each device in IoT will have a set of keys called the key ring from a key pool assigned to it. It should be noted that the residual design that is built from block complementation is being used in the IoT for the first time. A basic mapping from residual design to key pre-distribution is illustrated. Another advantage of this approach is improving the IoT resilience while maintaining high scalability. The evaluations performed indicate that our approach leads to an improvement in secure connectivity and an increase in IoT scalability with high resilience.

1. Introduction

The Internet of Things (IoT) is defined as a network of smart devices that share information through interacting with one another. "Things" refers to any physical object with a device with a unique IP address. This device can connect to a network to send and receive data. The IoT is now used to define several things, such as the convergence of multiple technologies, real-time analytics, machine learning, object sensors, and implant systems. The IoT can be considered a framework for the Smart City and Smart Energy Management Systems that are widely used today. Each IoT has various devices, such as sensors, actuators, RFID tags, and smartphones or backend servers, which vary in size, capability, and functionality. The 6LoWPAN has recently been used to help even the smallest devices connect to the Internet. The idea behind 6LoWPAN is that everything is expected to support the TCP/IP protocol stack and join the IoT. Making the IPv6/RPL connected 6LoWPANs secure is challenging since the devices are connected to the untrusted Internet. Also, the resources used are constrained, and the communication links are lossy. A second challenge is related to the limitations and constraints of IoT devices in terms of memory and processing power. These

limitations, in turn, indicate a restriction in the size of keys, IDs, and Rings [1].

The sensors on the IoT, in addition to the senses, can process and store sensed events. They can even intelligently recognize if a sensed event is a repeating one. IoT combines multiple technologies such as RFID, wireless sensor networks (WSN), NFC, etc. WSN is a subset of IoT. IoT is responsible for data processing, manipulation, and decision-making. In IoT, the data is sent to the Internet in only one hop. First, in IoT, routing is not implemented. Sensors send their data directly to the Internet because they have an Internet connection. In IoT, each device is identifiable with a unique ID: its IP address [2].

IoT is gradually becoming a significant part of different aspects of our lives. It is used in smart homes, wearable devices, healthcare, etc. Its wide range of applications yields common data, such as the enormous value of user's private information. Hence, the security of this information is very consequential. Several factors, such as data confidentiality, data integrity, authentication, access control, and privacy, are required to provide security for the IoT. It should also be noted that the authentication of IoT devices is of particular significance [3].

We need to use proper cryptographic methods to achieve high security in IoT. Cryptographic methods are divided into two categories, including symmetric key and asymmetric key cryptography. Asymmetric key cryptography, such as RSA and ECC (Elliptic Curve Cryptography), requires high computational cost, more processing time, and larger key sizes. Thus, for these reasons, the use of asymmetric key cryptography in IoT is limited, and as a result, symmetric-key cryptography is recommended [4].

Since IoT security depends on the method used to distribute keys between IoT devices, an effective key establishment method is required to distribute the cryptographic keys between the IoT devices. As already mentioned, public-key cryptography requires high computational costs. So, key pre-distribution is a solution to the key establishment problem in IoT, where each device is pre-loaded with a finite set of keys before deployment. The key pre-distribution scheme (KPS) determines which nodes store which keys [5]. Traditional cryptographic methods cannot provide authentication in their present form for the expected 50 billion devices. Using related technologies such as DES, 3DES, and AES to encrypt resource-constrained IoT devices requires too much energy [6] [7].

An IoT has various criteria that analyze key distribution solutions, such as memory overhead, connectivity, scalability, resilience, and communication overhead. Memory overhead is the required memory to store keys in every device. Connectivity refers to the probability of a shared key between two nodes. Resilience is the persistence of the IoT against node capture. Scalability is the maximum network size supported by a KPS, and communication overhead is the number of messages sent between nodes [8].

A KPS uses three methods: random, deterministic, and hybrid. The first schemes require that the keys be selected randomly from a key pool and stored in each object. This method will not guarantee the direct communication of each two nodes. Lack of direct communication creates a path between the two nodes, reducing communication speed. Also, a deterministic method should be used to design a key pool and key rings to achieve better key connectivity. A combination of both the deterministic and random approaches creates a hybrid method that can be used to improve scalability and resiliency [9].

The key pre-distribution scheme is a good solution for IoT security and is used in most research studies. Each KPS has three phases:

1) Pre-distribution
2) Shared-key discovery
3) Path-key establishment

A key pool is produced during the first phase. Subsequently, a subset of the key pool, namely the *key ring,* is assigned to each sensor node. The second phase is carried out after the deployment of sensor nodes. Each pair of nodes must communicate with each other to find at least one shared key between them. In the final phase, the two nodes without a shared key that want to communicate with each other may create a secure path using one or more intermediate nodes in which each pair of nodes shares a standard key [10].

Applying combinatorial designs in KPS with proper parameters can cause a decrease in the length of the key path and increase the maximum connectivity. The solid mathematical structure of combinatorial designs results in the communicational algorithms that can be reduced to $O(1)$ in the path-key establishment and shared-key discovery phases.

The IoT wants to convert traditional devices into connected devices by using interchanging data and communications to monitor and control the devices. To get the required security on the IoT, we must consider the following challenges [11] [12]:

- **Resource constraints:** IoT devices frequently operate on channels with low-bandwidth communication. Therefore, it is impossible to execute directly standard conventional security protocols of the Internet in the context of IoT.
- **Resilience to attacks:** IoT devices are typically small and inexpensive, with low physical protection. For example, a mobile device can be stolen, or fixed devices can be moved.
- **Scalability:** The IoT is universally composed of a large number of devices. The proposed security approach must be able to scale all those included.

This paper focuses on combinatorial constructions for key pre-distribution schemes in IoT. To improve IoT resilience while maintaining high scalability and secure

connectivity, we illustrated a novel basic mapping from residual design based on block complementation to key pre-distribution. The residual design is constructed from block complementation and is used on the IoT for the first time. The new approach has been analyzed and compared analytically and experimentally with other state-of-the-art KPSs, examining various evaluation criteria. It was indicated that the proposed scheme amended network scalability and decreased memory overhead compared to other works [13].

The rest of the paper is organized as follows. In section 2, related work is summarized. In section 3, we provide a brief overview of key pre-distribution and combinatorial design theory. In sections 4 and 5, we introduce and analyze the proposed scheme and present how to map the residual design to key distribution. Section 6 presents the implementation and simulation results. Finally, Section 7 concludes the paper.

2. RELATED WORK

A q-composite random key pre-distribution scheme has been proposed by Chan et al. [14]. This scheme can enhance the security of communication between the two nodes. Based on this scheme, every two nodes may create a secure link on condition that they have at least *q* shared keys. Qian proposed A key pre-distribution scheme [15]. This scheme contained a hash function to improve the resilience against node capture attacks. Recently, binational design in key pre-distribution has been proposed as a solution. In this study, a -PBIBD combinatorial design is introduced and constructed, and the mapping of such design as a key pre-distribution scheme in the resource-constrained IoT network is explained. Using such a pre-distribution scheme, more keys are obtained for communication between two devices in the IoT network [16] [17].

A new key pre-distribution scheme named POK (adaPtive and rObust Key pre-distribution) is presented in [18]. POK improves the way keys are generated and pre-loaded in the sensor nodes. The main idea of the POK is that newly added sensor nodes will be pre-loaded with pairwise keys computed by using a hash function and having knowledge of the number of future post-deployments. A comparison study with related works concludes that POK offers less communication overhead and doesn't require time synchronization, leading to an energy-efficient scheme.

Different encryption and hash algorithms were proposed by Vinayaga et al. [19] to enhance the security of smart home systems. Their algorithms were designed to secure any communication between the devices within an IoT System. Thus, a hash algorithm was created based on RC4, and its efficiency was measured against the existing hash algorithms.

In [20], a key pre-distribution scheme has been proposed based on the combinatorial design for IoT. This scheme has increased the scalability of the network. For the proposed scheme, a kind of mapping from the unital design to the key establishment has been proposed, which yields a network with high scalability. The results indicate that the proposed scheme increases network scalability considerably with high resilience.

To combine security in an IoT-based Smart Home System, Santoso et al. [21] proposed a method to maintain user comfort. Their paper explained how to implement a WiFi-based IoT Smart Home system, including IoT devices such as sensors, actuators, and equipment. These devices were connected to the Home Gateway over the Home network. They designed a user device to control and monitor the system. This device was connected to a Home gateway over the Internet. The home gateway made it possible for IoT devices to communicate securely. Also, it allowed users to access, configure, and control the system via the user interface. It is an open-source IoT framework containing various libraries from cryptography (ECC, AES, etc).

In [22], the current cryptographic methods, such as the Advanced Encryption Standard (AES) and the Elliptic Curve Cryptography (ECC), are expounded, and their functionality, together with their advantages and disadvantages, are discussed. Also, this paper highlights the need for more flexible cryptographic suites.

The term security is a vital issue in any sensor network. In these networks, key management is considered the main security service. Due to the limitations on sensor nodes, traditional key management techniques do not fit with sensor networks. A new key pre-distribution scheme was proposed in [23] using multivariate polynomials to establish the pairwise key in sensor networks. Based on this approach, the combinatorial design theory must be applied in the multivariate key pre-distribution scheme. In this scheme, the common multivariate polynomials can be stored in sensor nodes before deploying the network. This idea is done using the identifier of sensors and the combinatorial design. Also, compared to previous schemes, the proposed approach receives better security in terms of resilience against node capture with the exclusion of additional communication overhead.

In [24], an advanced key administration framework for remote sensor networks is proposed, consolidating fuzzy logic and AES encryption to improve the performance of the WSNs. The proposed framework uses fuzzy logic for cluster formation and head rotation and utilizes the AES algorithm to encode the information. It falls in the classification of techniques that depend on hierarchical structures, in which the sensor nodes use pre-distribution and post-deployment mechanisms to distribute keys. The proposed key management uses fuzzy logic, which enhances security and energy efficiency. Thus, the energy utilized by the network is reduced, and the network's lifetime is improved.

A key pre-distribution scheme was presented in [25] for a clustered heterogeneous WSN using transversal designs. In this novel scheme, key rings are assigned to sensor nodes before the network is deployed, and a key pool of each cluster is separated by adding a pseudo-random-generated number after the network is deployed. The efficiency evaluation and security analysis results suggested that the proposed scheme, compared to other key management schemes, can provide better security and considerably reduce communication overhead and memory space without losing connectivity.

The residual design, a novel combinatorial approach, was proposed in [26] for key establishment. This approach requires that WSNs have a highly scalable key management scheme. The scheme is intended to provide highly secure connectivity. This scheme implies that the residual design undergoes a basic key pre-distribution mapping with high network scalability. It should be noted that this mapping lacks high resilience. Accordingly, a new approach should be designed for key pre-distribution based on the residual design to improve network resilience, maintain connectivity, and high scalability. Results suggest that the use of this approach leads to a reduction in computational cost and memory overhead. Although this approach provides the same connectivity based on the first scheme, the analysis and numerical results suggest that the optimized approach yields better network resilience. At the same time, it leads to lower network scalability against the residual design key pre-distribution scheme at an equal key-ring size [27].

In [28], the authors propose a new key pre-distribution scheme for wireless sensor networks based on combinatorial design. The proposed scheme divides the WSN into cells of the same size, where the sensor nodes are distributed evenly. Each cell has two types of sensor nodes, including the cluster head and the sensor node. The communication within the cell is direct; the communication between the nodes of the different cells is done through the cluster head. This scheme would reduce the key storage overhead and increase overall network resistance.

With a symmetric key, shared key allocation methods could be accomplished in cryptography before or after the network deployment. The one that occurs before the deployment is called the key pre-distribution. The Key Pre-distribution Schemes (KPSs) are the most desirable choices due to their limited computational costs and constrained energy and communication capacities of end devices. Therefore, keys are assigned to the end device's memory before their distribution in the network. According to these schemes, every pair of nodes can usually communicate securely because of the shared common credential(s) [29].

Cryptography schemes such as asymmetric or public keys normally facilitate secure communication between objects. Of course, it is not advisable to use these schemes for the sake of deployment on low-power battery operating devices. This is because they are required to compute costly cryptographic operation(s). However, the approaches proposed in [30] to reduce the number of exchanged messages are intended to design asymmetric key schemes for environments with resource constraints, such as the IoT. Other researchers [31] revealed that asymmetric solutions should be used for resource-constrained devices. The reason is that they have acceptable flexibility and scalability regarding shared key management.

Camtepe and Yener have proposed combinatorial designs for key pre-distribution in WSNn [32]. Their paper presented a new deterministic KPS based on the Symmetric Balanced Incomplete Block Design (SBIBD). The SBIBD is mapped onto the key pre-distribution to create $m^2 + m + 1$ key-rings from a key pool S of $m^2 + m + 1$ keys. There are $k = m + 1$ keys in each key-ring. Also, precisely one common key is shared by every two key rings. The main advantage of the Camtepe scheme is that every two nodes share exactly one common key. However, SBIBD schemes do not match extensive networks. To construct roughly $m^2 + m + 1$ key-rings, key rings of $m+1$ keys should be used. In the article [33], the SBIBD-based key pre-distribution was used to guarantee intra-region secure communications in grid group WSNs.

In [14], a perfect network resilience was proposed by Chan et al. aimed at obtaining network scalability of $O(k)$ where k is the key-ring size. The SBIBD [32] could also obtain network scalability of $O(k^2)$. For this reason, the unital design theory was used to pre-distribute keys. Their paper proposed mapping from units to key pre-distribution to achieve a good trade-off between scalability and connectivity. Hence, the method proposed in their paper was designed to improve network resilience against node capture attacks.

Contrary to wireless sensor network security, security in the IoT involves end-to-end communications. The IoT devices deny the possibility of defining static client and server roles. The devices in IoT act alternatively as a client and a server. Every IoT device has four criteria: the number of exchanged messages, the required bandwidth, the complexity of computations, and the possibility of pre-computations. These criteria are important in the cryptographic protocol. They only matter when they have to be implemented by highly resource-constrained devices. A good metric for these nodes is the overall energy consumption induced by both computations and message exchanges. Fig. 1 shows some applications of IoT devices. As can be seen from Fig. 1, secure communication is vital in every IoT device.

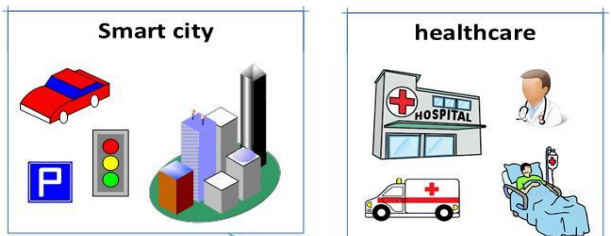


Fig. 1. Secure communication is essential in the IoT

3. Key Pre-Distribution and Combinatorial Design

3.1. Key Pre-distribution

Key management is the techniques and procedures for establishing secure communications between authorized parties. It is vital for a secure connection in IoT. Creating secret keys between sensor nodes is exceptionally challenging due to resource constraints (energy, CPU, and memory) on the nodes.

Key management includes four essential functions: analysis, assignment, generation, and distribution of network keys, such as Fig 2. A central server is responsible for storing and distributing the key pool. In a symmetric key algorithm, the keys must be chosen carefully, distributed, and stored securely.

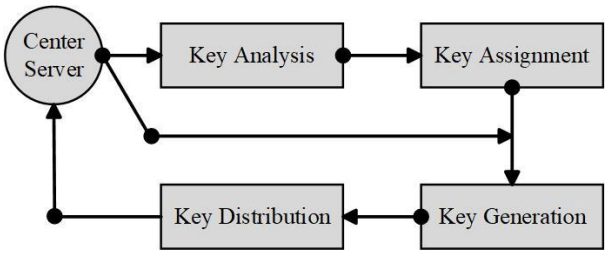


Fig. 2. Key Management process.

1. **Key analysis:** First, the number of keys required for the network, as well as the number of keys needed for each node, are analyzed.
2. **Key assignment:** This step refers to mapping keys to different parts. In this case, a key assignment manager assigns the key to the parts that want to create a secure communication channel. In this case, a key manager determines how many keys are assigned to each node to create a secure communication channel.
3. **Key generation:** This step may occur once or several times over the network's life. In the static key distribution scheme, the keys are generated by a central server and loaded in nodes before network development.
4. **Key distribution:** This step involves delivering the generated keys to predefined nodes.
   When an attack occurs, the above steps are repeated to ensure the network's security.

Key pre-distribution is the most effective technique to establish secure communication between nodes. Based on the key pre-distribution scheme, we should assign determined keys to each sensor before deployment. Before deployment, each sensor node should be pre-loaded with a set of keys from the large pool. Based on key pre-distribution, every two nodes with at least one shared key can create a communication path with another node. Given the features of the IoT, the use of key pre-distribution yields better results. A KPS includes three phases: key pre-distribution, shared key discovery, and path-key establishment. Security keys should be created and allocated to nodes during these three phases. Two nodes should detect one or more shared keys to make a secure connection. These keys are diverse in each KPS, and then communication is done between nodes using these shared keys.

The IoT includes devices with constrained resources that suffer from low memory capacity. Nonetheless, in most methods, the size of the key rings is related to the network size. Most of the existing techniques suffer from low scalability and memory overload. This problem led us to use a combinatorial design, especially residual design theory. To further expand, we will start with the definition of *block complementation* and the features of *residual design* theory. Afterward, we will propose the basic mapping from residual design to key pre-distribution and evaluate its performance metrics [34].

3.2. Combinatorial Design

Combinatorial design theory deals with arranging elements into subsets satisfying some generalized concepts of balance and symmetry. We focus primarily on the definition and properties of a particular kind of design, Balanced Incomplete Block Designs (BIBD) and symmetric BIBD. This paper defines a projective plane and blocks its complementation. Then, we build a residual design from block complementation.

Symmetric BIBD

A BIBD is a design (X, A) with positive integer parameters v, k, and λ such that $v > k$. Therefore, a $(v, k, \lambda) - BIBD$ is a design that $|X| = v$, and each block includes exactly k elements (points), and every both distinct points is included in precisely λ block. In definition, X is a set of points $X = \{x_1, x_2, \dots, x_v\}$, and $A = \{A_1, A_2, \dots, A_b\}$ is a collection of non-empty subsets of X called blocks. Generally, a BIBD contains v distinct objects into b blocks with size k, so each object includes exactly r various blocks, and every different

point occurs together in exactly λ blocks. Then, the design is explained as (v, k, λ), or equivalently (v, b, r, k, λ), where [35] [36]:

$$\lambda(v-1) = r(k-1) \qquad bk = vr \tag{1}$$

A *Symmetric BIBD* or *Symmetric Design* is a *BIBD* with $b = v$ and therefore $k = r$. In *Symmetric Design*, every block includes $k = r$ points, every object is contained in $r = k$ blocks, and every pair of objects is included in λ blocks, and finally, every pair of blocks intersects in λ objects. This paper uses a subset of Symmetric Designs called a Projective Plane.

block complementation

In this study, we use a *projective plane* with parameters $(q^2 + q + 1, q + 1, 1)$ where $q \geq 2$ and q is a prime number. Here, we state one method of constructing new BIBDs from old BIBD that is called *block complementation*. Suppose (X, A) is a $(v, b, r, k, \lambda) - BIBD$, where $k \leq v$-2. Then *block complementation* is done by replacing every block $A_i \epsilon A$ by $X \backslash A_i$ for $1 \leq i \leq v$. This created design is a $BIBD$ with parameters $(v, b, b - r, v - k, b - 2r + \lambda)$ [37].

Example 1: consider a *projective plane* with order $q = 2$, $(7,3,1) - BIBD$; then we construct *block complementation* with parameters $(7,4,2) - BIBD$. The element set and blocks of both designs are as follows:

Projective plane $(7,7,3,3,1) - BIBD$:

$$X = \{1,2,3,4,5,6,7\}$$
$$A_1 = \{1,2,4\}, A_2 = \{2,3,5\}, A_3 = \{3,4,6\},$$
$$A_4 = \{4,5,7\}, A_5 = \{1,5,6\}, A_6 = \{2,6,7\},$$
$$A_7 = \{1,3,7\} \tag{2}$$

Block complementation $(7,7,4,4,2) - BIBD$:

$$X = \{1,2,3,4,5,6,7\}$$
$$A'_1 = \{3,5,6,7\}, A'_2 = \{1,4,6,7\}, A'_3 = \{1,2,5,7\},$$
$$A'_4 = \{1,2,3,6\}, A'_5 = \{2,3,4,7\}, A'_6 = \{1,3,4,5\},$$
$$A'_7 = \{2,4,5,6\} \tag{3}$$

Then, using block complementation, we build residual design sets, as described in the following section.

Residual Design

Given a symmetric $(v, k, \lambda) - BIBD$ with elements $X = \{x_1, x_2, \dots, x_v\}$ and blocks $A = \{A_1, A_2, \dots, A_v\}$, then for every $1 \leq i \leq v$, fixing a block A_i Deleting this block and its elements from all other blocks of SBIBID constructs a new BIBD called *Residual Design*. That is, for any i, $\{A_1 \backslash A_i, A_2 \backslash A_i, \dots, A_v \backslash A_i\}$ are the blocks of a $(v - k, v - 1, k, k - \lambda, \lambda) - BIBD$ of the element set $X \backslash A_i$ Provided that $\lambda \neq k - 1$ [38] [39].

Based on the projective plane with parameters $(q^2 + q + 1, q + 1, 1)$ and the definition of the block complementation, we create a $BIBD$ with parameters $(q^2 + q + 1,\ q^2, q^2 - q)$. Then, we create a residual design based on the new $BIBD$.

Example 2: Consider $(7,4,2) - SBIBD$ in example 1. Then, we can create seven classes of residual sets for any A_i, where each building is a $(3, 6, 4, 2, 2) - BIBD$ over the element set $X \backslash A_i$. Therefore, by considering A_i as a fixed block in each class C_i, we have seven classes of residual sets, including the following blocks:

$C_1 = X \backslash A_1 = \{1,2,4\}, A_2 \backslash A_1 = \{1,4\}, A_3 \backslash A_1 = \{1,2\},$
$A_4 \backslash A_1 = \{1,2\}, A_5 \backslash A_1 = \{2,4\}, A_6 \backslash A_1 = \{1,4\}, A_7 \backslash$
$A_1 = \{2,4\}.$
$C_2 = X \backslash A_2 = \{2,3,5\}, \quad A_1 \backslash A_2 = \{3,5\}, \quad A_3 \backslash A_2 =$
$\{2,5\}, \quad A_4 \backslash A_2 = \{2,3\}, \quad A_5 \backslash A_2 = \{2,3\}, \quad A_6 \backslash A_2 =$
$\{3,5\}, A_7 \backslash A_2 = \{2,5\}.$
$C_3 = X \backslash A_3 = \{3,4,6\}, \quad A_1 \backslash A_3 = \{3,6\}, \quad A_2 \backslash A_3 =$
$\{4,6\}, \quad A_4 \backslash A_3 = \{3,6\}, \quad A_5 \backslash A_3 = \{3,4\}, \quad A_6 \backslash A_3 =$
$\{3,4\}, A_7 \backslash A_3 = \{4,6\}.$
$C_4 = X \backslash A_4 = \{4,5,7\}, \quad A_1 \backslash A_4 = \{5,7\}, \quad A_2 \backslash A_4 =$
$\{4,7\}, \quad A_3 \backslash A_4 = \{5,7\}, \quad A_5 \backslash A_4 = \{4,7\}, \quad A_6 \backslash A_4 =$
$\{4,5\}, A_7 \backslash A_4 = \{4,5\}.$
$C_5 = X \backslash A_5 = \{1,5,6\}, \quad A_1 \backslash A_5 = \{5,6\}, \quad A_2 \backslash A_5 =$
$\{1,6\}, \quad A_3 \backslash A_5 = \{1,5\}, \quad A_4 \backslash A_5 = \{1,6\}, \quad A_6 \backslash A_5 =$
$\{1,5\}, A_7 \backslash A_5 = \{5,6\}.$
$C_6 = X \backslash A_6 = \{2,6,7\}, \quad A_1 \backslash A_6 = \{6,7\}, \quad A_2 \backslash A_6 =$
$\{6,7\}, \quad A_3 \backslash A_6 = \{2,7\}, \quad A_4 \backslash A_6 = \{2,6\}, \quad A_5 \backslash A_6 =$
$\{2,7\}, A_7 \backslash A_6 = \{2,6\}.$
$C_7 = X \backslash A_7 = \{1,3,7\}, \quad A_1 \backslash A_7 = \{3,7\}, \quad A_2 \backslash A_7 =$
$\{1,7\}, \quad A_3 \backslash A_7 = \{1,7\}, \quad A_4 \backslash A_7 = \{1,3\}, \quad A_5 \backslash A_7 =$
$\{3,7\}, A_6 \backslash A_7 = \{1,3\}.$ (4)

In this study, we build the residual design by symmetric BIBD with parameters $(q^2 + q + 1,\ q^2, q^2 - q)$. Consider the *i*th class of the residual design that is created by the select block A_i as a fixed block, therefore, the element set of each class builds a BIBD with parameters $(v, b, r, k, \lambda) = (q + 1, q^2 + q, q^2, q, q^2 - q)$. In this paper, the focus is on a residual design that runs for q as a prime power. The $v \times b$ incidence matrix, named *M*, may define a residual. In this matrix, rows represent the x_i points and columns represent the A_j blocks. Subsequently, matrix *M* can be defined as:

$$M = \delta_{ij} = \begin{cases} 1 & if\ x_i \in A_j \\ 0 & otherwise \end{cases} \tag{5}$$

properties

- **Maximum network size that is supported in residual design is $N = (q^2 + q + 1)(q^2 + q)$:** Since the number of classes in residual design is exactly $q^2 + q + 1$, and each class forms $(q + 1, q^2 + q, q^2, q, q^2 - q) - BIBD$, therefore, we can have in total $(q^2 + q + 1)(q^2 + q)$ blocks for support nodes.
- **Any two classes have only one common element:** whereas the element set for each class C_i is the same corresponding with the block A_i in the projective plane. Therefore, as defined for the projective plan, every two blocks have exactly one common element.
- **Each element in the residual design is included in**

exactly $q^2(q+1)$ block: Due to the point set each class in residual design with parameters $(q+1, q^2+q, q^2, q, q^2-q)$, each element is included in $q+1$ classes, and each element is repeated in q^2 block in each class; therefore, each element is included in $q^2(q+1)$ blocks.

Block designs are precisely relevant to key pre-distribution schemes. Key rings are assigned to devices in a KPS system proposed for an IoT. We suppose a key matches with a point, and a key ring matches with a block. For example, a residual design based on KPS creates $(q^2+q+1)(q^2+q)$ key rings from a key pool with q^2+q+1 keys. In IoT, if two key rings have at least one shared key, the corresponding two devices can be directly and securely connected since they have at least one common key.

4. The Proposed Approach

4.1. mapping from residual design to key pre-distribution in IoT

In our scheme, we consider an IoT of *N* nodes (devices), where each node is assigned a key ring from a key pool. As already mentioned, we build a residual design using a form of symmetric BIBD, including parameters $(q^2+q+1, q+1, 1)$ where q is a prime number.

We proposed a basic mapping in which a distinct key matches a residual point. the key ring also corresponds to each block, and the key pool matches the global set of points. Then, we can create $N=(q^2+q+1)(q^2+q)$ key rings from a key pool with $|X|=q^2+q+1$ keys. The size of each key ring is $k=q$ keys. We select q as a prime number in such a way that $(q^2+q+1)(q^2+q)\geq N$.

This residual design contained q^2+q+1 classes where each class has q^2 elements. In total, it creates $(q^2+q+1)(q^2+q)$ blocks of size q. Then constructed blocks as key-rings are assigned to N devices.

We have indicated basic mapping from the residual design to key pre-distribution in Table 1. First, we create the residual blocks according to key-rings. Then, we allocate a distinct key ring to each node in which each key ring has a key identifier. After the assignment, every two adjacent devices exchange their key identifiers to determine a shared key. In our approach, every two devices share at most one common key. According to residual features, every two points are included together in exactly one block, which results in the fact that two blocks cannot have more than one common point. Therefore, if two adjacent devices have one common key, the key is selected as a pairwise key that is further used to create secure communication. Otherwise, devices are required to determine secure paths, including some secure communications.

Table 1 Mapping from residual design to key pre-distribution

Residual Design	Key pre-distribution
Point Set (S)	Key-Pool (*P*)
Blocks	Key-Rings
Object Set Size ($\|X\|=v=q^2+q+1$)	Key-Pool Size $\|P\|$
Number of Blocks $b=(q^2+q+1)(q^2+q)$	Number of Key-Rings (*N*)
Size of a Block $k=q$	Size of a Key-Ring (*K*)
Number of Blocks that an Object is in $r=q^2$	Number of Key-Rings that a Key is in
Two Blocks share $\lambda=q^2-q$ Objects	Two Key-Rings share λ Keys

To create an IoT with N devices, we require N key-rings; therefore, a residual design with $b=N$ blocks and set X with $|X|=v$ points need to be constructed. Hence, with prime number q, we have $v=q^2+q+1$ and $b=(q^2+q+1)(q^2+q)$. Each point in X can be related to a distinct random key, and each block can be associated with a key ring. Residual design guarantees that every two blocks have λ points in common; each key-rings (or device) has λ common keys. Table 2 indicates notations that are used in the remainder of the paper. The key pre-distribution approach proposed in this paper for an IoT of size *N* can be explained briefly in Algorithm 1.

Table 2 List of used notations.

Notation	Definition
N	Total number of nodes in the IoT
N_{CRD}	Number of supported nodes in CRD
l	The key size
k	Key-ring size & Block size of a given design
q	The design order (a prime number)
C_i	*i-th class of residual design*
B_{ij}	*j-th block in class i*
P_{CRD}	The probability that two nodes can establish a secure link
$P(L\|C_x)$	The network resiliency when *x* nodes are captured

The most important advantage of our approach is improving the probability of a shared key. As explained in the next section, our approach allows us to obtain highly secure connectivity coverage and network scalability since a block with $k=q$ disjoint keys is assigned to each device. Also, this solution provides good network resiliency due to the pairwise secret keys, which augment secure communications. Moreover, this approach illustrates that our solution can achieve higher network scalability than the existing solutions.

5. Analyses of the proposed scheme

5.1. Theoretical analysis

This section analyzes the proposed scheme, considering four important metrics: network connectivity, memory overhead, network scalability, and *resilience* against node capture attacks.

1) ***Connectivity*:** Connectivity is the probability of every two nodes sharing at least one common key. We assume that B_{ij} (block *j* in class *i*) and $B_{i'j'}$ they are two blocks of the residual design. These two

blocks are either in the same class or in different classes. We continue to consider the probability of a shared key in both cases as follows:

i. **Same class:** any two blocks are included in the same class ($i = i'$, e.g. C_i). In this case, the probability that every two blocks from the same class have at least a common key is 1.

Proposition 1. The probability Q_{SC} that each pair of blocks is in the same class is calculated as follows:

$$Q_{SC} = \frac{\binom{q^2+q}{2}}{\binom{(q^2+q)(q^2+q+1)}{2}} \quad (6)$$

Proof. We have $q^2 + q + 1$ classes and each class has $q^2 + q$ blocks. □

ii. **Different Classes:** Each of the two blocks is included in different classes ($i \neq i'$, e.g. C_i, $C_{i'}$). In this case, the probability of a shared key of block B_{ij} is checked with blocks of two categories of classes:

1. Classes whose point set includes exactly one point of the block B_{ij}.
2. Classes that do not contain any points in the block B_{ij}.

In case 1: Each block in class C_i has exactly one common point with point set q^2 other classes. Therefore, the probability of selecting one of these classes is $\frac{q^2}{q^2+q}$.

Given the definition of the residual design, in each of these classes, there are q^2 blocks that contain one common key with block B_{ij} from the class C_i. That is, the probability of choosing a block with a shared key in each of these classes is $\frac{q^2}{q^2+q}$.

Proposition 2: in case of 1**,** let's assume that B_{ij} is a block of class C_i, and block $B_{i'j'}$ is of another class, then we can calculate the probability of B_{ij} and $B_{i'j'}$ having at least one common key using the following relation:

$$P_{dcA} = \frac{q^2}{q^2+q} \times \frac{q^2}{q^2+q} \quad (7)$$

Algorithm 1. Mapping from Residual Design based on Block Complementation to key pre-distribution in IoT

Require: N {Total number of devices}

1. Find the minimum prime number q such that $(q^2 + q + 1)(q^2 + q) \geq N$.
2. Generate the projective plane of order q (Symmetric Design) with parameters $(q^2 + q + 1, q + 1, 1)$.
   - $v = q^2 + q + 1$ element $X = \{x_1, x_2, \dots, x_v\}$.
   - $b = v$ blocks $A = \{A_1, A_2, \dots, A_v\}$ of size $q + 1$.
3. Build block complementation, based on the projective plane, with parameters $(q^2 + q + 1,\ q^2, q^2 - q)$.
   - $v = q^2 + q + 1$ element $X = \{x_1, x_2, \dots, x_v\}$.
   - $b = v$ blocks $A' = \{A'_1, A'_2, \dots, A'_v\}$ of size q^2.
4. Create a residual design based on block complementation containing $q^2 + q + 1$ class, each class with parameters $(q + 1, q^2 + q, q^2, q, q^2 - q)$.
   - Blocks $A_{ij} = A'_i \backslash A'_j$: j-th block in class i.
5. Delete repeated blocks in all classes.
6. Assign blocks to specified devices.

Proof. The probability of selecting one class in case 1 is $\frac{q^2}{q^2+q}$. Also, by definition of the residual design, the probability of choosing a block with a common key in case 1 is $\frac{q^2}{q^2+q}$. □

In case 2: in this case, the probability that two blocks in different classes have at least one common key is zero.

Proposition 3: the probability Q_{DC} that different classes own every two blocks can be calculated as follows:

$$Q_{DC} = \frac{\binom{q^2+q}{1}\binom{q^2+q}{1}}{\binom{(q^2+q)(q^2+q+1)}{2}} \quad (8)$$

Proof. We have $q^2 + q + 1$ classes and each class has $q^2 + q$ blocks. Therefore, we select two blocks in two different classes. □

Proposition 4 The probability that each pair of blocks has one or more common keys in residual design is calculated as follows:

$$P_{CRD} = 1 \times Q_{SC} + \left(\frac{q^2}{q^2+q} \times \frac{q^2}{q^2+q}\right) \times Q_{DC} \quad (9)$$

Proof it follows from Proposition 1 to 3. □

2) ***Memory overhead*:** In the key pre-distribution scheme, each node needs the memory to store keys. When using the proposed residual design of order q, one key ring was assigned to each node. Therefore, there were q disjoint keys in each node. Also, to store keys in each node, the memory required was $l \times q$, where the key size was denoted by l.

3) ***Network scalability*:** network scalability means the maximum number of nodes a network can support. This number is similar to the key rings that design can

support. Hence, based on our approach, we can use the relation $N = (q^2 + q + 1)(q^2 + q)$ calculate the total number of possible key rings when the residual design is used. Since each block of residual design is repeated q times in each class, and disjoint blocks from N possible blocks of residual design are assigned to each node; the maximum number of nodes that we can support is equal to:

$$N_{CRD} = \frac{(q^2 + q + 1)(q^2 + q)}{q} = (q^2 + q + 1)(q + 1) \quad (10)$$

This section will calculate the maximum network size based on the compromised nodes to keep the network secure.

4) **Resilience:** Network resiliency is defined as the fractions of secure external links that are uncompromised when x sensor nodes are captured. In terms of resilience, we are interested in solving the probability $P(L|C_x)$ that is calculated as:

$$P(L|C_x) = \sum_{\forall j} P(l_j|l)P(D_j|C_x) \quad (11)$$

Equation 11 shows the probability of a link L being compromised when x randomly selected nodes and their related key-rings are captured by an adversary. In Equation 11, C_x is the number of times that x nodes are captured, l_j is the number of times that a given link is secured with key j, and finally, D_j is the event that a key ring, including key j is compromised.

First, we assume two nodes, v, and u are not captured. If x node is attacked and decrypted, the probability of an attacker's decrypting the communication between v and u can be calculated using Equation 11. The probability that communication is secured with key j can be calculated as:

$$P(l_j|l) = \frac{\binom{q^2(q+1)}{2}}{\binom{(q^2+q+1)(q+1)}{2}} \quad (12)$$

Also, the probability that D_j key-ring includes key j is compromised, with x captured nodes computed as:

$$P(D_j|C_x) = 1 - \frac{\binom{(q^2+q+1)(q+1)-q^2(q+1)}{x}}{\binom{(q^2+q+1)(q+1)}{x}} \quad (13)$$

Finally, the probability that a link is compromised when x nodes are captured by an adversary can be computed using the following relation:

$$P(L|C_x) = \sum_{j-1}^{q^2+q+1} \frac{\binom{q^2(q+1)}{2}}{\binom{(q^2+q+1)(q+1)}{2}} \left(1 - \frac{\binom{(q^2+q+1)(q+1)-q^2(q+1)}{x}}{\binom{(q^2+q+1)(q+1)}{x}}\right) \quad (14)$$

In our approach, the resilience against node capture is a significant parameter. An attacker may attack our proposed idea in two ways. First, the attacker agrees with the link key between nodes without capturing them. Second, sensor nodes may be captured by the attacker to prevent creating pairwise keys. Therefore, our main metrics of interest include the fraction of compromised secure links between pairs of uncompromised nodes and the fraction of compromised keys.

6. Implementation

The Contiki Operating System was used to develop the simulation. Contiki is created to run on hardware devices that are severely limited in terms of memory, power, processing power, and communication bandwidth. There is a network simulator called Cooja in every Contiki system, which simulates network nodes. Contiki is designed for small-scale systems. It has merely a few *kilobytes* of memory available. The recently standardized IETF protocols for low-power IPv6 networking, including the 6LoWPAN adaptation layer, the RPL IPv6 multi-hop routing protocol, and the CoAP RESTful application-layer protocol, are supported by this system.

Eschenauer and Gligor proposed A key pre-distribution algorithm [40]. In the context of RPL, the performance was examined using a simulation experiment. The experiment explicitly explores the percentage of leaves sharing a key in the RPL routing table.

We compared essential metrics of the proposed scheme (CRD) against other methods such as SBIBD, Combinatorial Trade, and Residual Design (RD) with a similar approach (combinatorial design) in a key establishment.

6.1. Performance Comparison

In this section, the CRD approach proposed is compared with the existing schemes in terms of different criteria. The parameters of different existing schemes, such as symmetric BIBD, combinatorial trade and residual design (RD) are summarized in table 3.

Table 3 parameters of SBIBD, Trade, RD, CRD

Design	v	b	r	k	λ
SBIBD	q^2+q+1	q^2+q+1	$q+1$	$q+1$	1
Trade	q^2+q+1	$2(q^2+q+1)$	$2(q+1)$	$q+1$	1
RD	q^2+q+1	$(q^2+q+1)(q+1)$	$q(q+1)$	q	1
CRD	$q+1$	$(q^2+q+1)(q+1)$	q^2	q	q^2-q

Combinatorial trade: In each $t-(v,k)$ trade (also known as the combinatorial trade), there are $T = \{T_1, T_2\}$ collections where T_i $(i = 1,2)$ is a collection of m blocks with the size of k (*k-subsets*) that are selected from X that the T_1 blocks are different from the T_2 $(T_1 \cap T_2 = \emptyset)$ blocks. Also, each *t-set* that is selected from X happens in the same number of blocks of T_1 similar to those of T_2. Therefore, upon noticing a $t-(v,k)$ Steiner trade of volume m, all the k-subsets of $T_1 \cup T_2$ as blocks of the design can be considered. It should be noted that any t-subset of elements happens in either 2 or no blocks. As soon as it is mapped onto the key pre-distribution, v is the size of the key pool, and T_1 and T_2 are the sensors holding k keys [41].

The key pre-distribution scheme will have a key pool size of q^2+q+1, provided that q is a prime power, the maximum number of nodes in the network is $2(q^2+q+1)$, and the key-ring size is $q+1$.

1) *Scalability*

We compared the scalability of our proposed scheme with three existing methods, namely SBIBD [32], Trade [41] and RD [26] in Fig 3. As can be seen from the figure, considering a similar key-ring size, the scheme proposed here leads to a significant increase in scalability compared to the two methods, namely SBIBD and Trade. However, its scalability is the same as the RD method. Therefore, simulation results suggest that considering similar network sizes, using RD and CRD schemes reduces the key-ring size compared to the other schemes.

In our proposed scheme, an equal number of key rings and devices that can be supported by design was used. For instance, in our scheme and the RD scheme, in case a network requires 2500 devices, the smallest prime number that satisfies this requirement is $q = 13$, which results in 2562 nodes. However, in SBIBD and Trade, to support this network with 2500 devices, the smallest prime number must be $q = 53$ and $q = 37$, respectively.

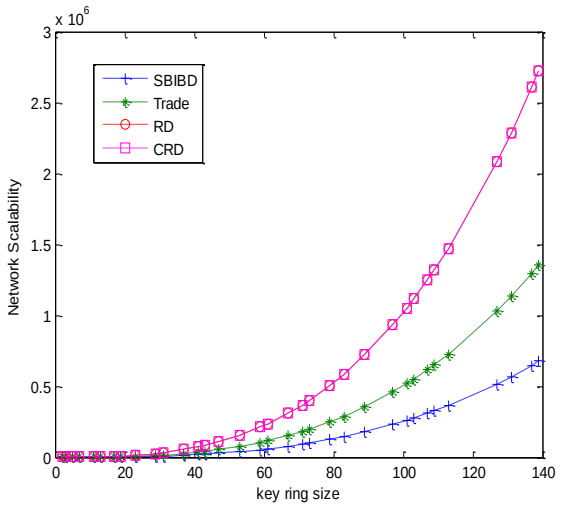


Fig. 3. A comparison of various schemes regarding Scalability

2) *Connectivity*

Two neighboring nodes can have at least one shared key to communicate with each other directly on an IoT device. Fig. 4 compares the probability of a shared key between nodes for four specified methods. It can be seen from the figure that the SBIBD method has a perfect probability of key sharing. The reason is that every two nodes in this method have at least one shared key. Results suggest that our proposed scheme improves connectivity compared to the RD and the Trade methods.

3) *Resilience*

With a similar approach, the resilience of our proposed scheme was compared against node capture attacks using the SBIBD, Trade and RD methods.

In Fig 5, all four methods are compared at an equal number of compromised nodes for a key-ring size of $k = 24$ and $k = 42$. To calculate the resilience of three methods SBIBD, Trade, and RD, we used [32], [41] and [26], respectively. It was found that our scheme, compared to the three other methods for a compromised node number (CNN) larger than 25, provides good resilience considering the same key-ring size $k = 24$. Also, the CRD resilience is compared with that of other schemes considering the same key-ring size $k = 42$. For compromised nodes numbers bigger than 47, CRD has an advantage in terms of resilience.

6.2. Discussion

In Tables 4 and 5, we illustrated the numerical results comparing scalability, connectivity, and resilience of the four schemes, namely SBIBD, Trade, RD, and CRD, considering similar key-ring sizes. The proposed scheme provides the maximum number of supported nodes for network scalability. For example, if the key-ring size were equal to $k = 90$, the CRD method would generate nodes more than 90 times the SBIBD and more than 45 times the Trade. Also, numerical results show that the scheme proposed in this paper is better than the other three schemes regarding network resilience. For example, considering key-ring size $k = 42$ and CNN=85, the CRD resilience=0.773, SBIBD=0.835, Trade=0.872, and RD=0.783. Also, our scheme increases the probability of shared key compared to the two methods, namely Trade and RD.

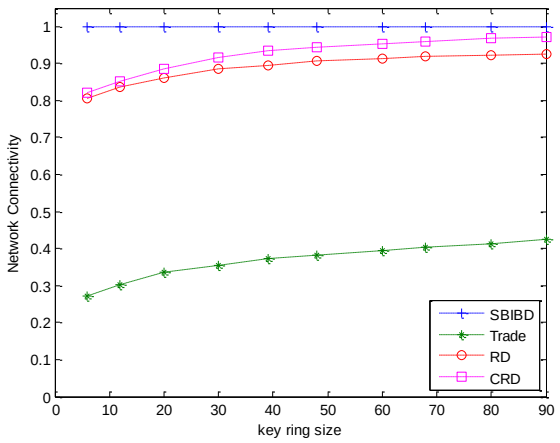


Fig. 4. Connectivity comparison our scheme with three existing methods. SBIBD has perfect connectivity and our scheme is better than Trade and RD.

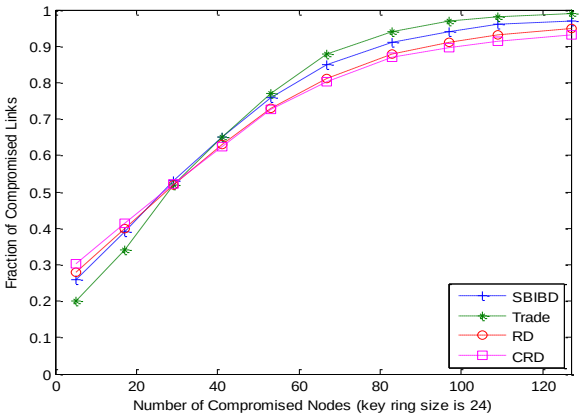
Table 4 Simulation results of different schemes in terms of connectivity and scalability

Method	key-ring size=20	key-ring size=30	key-ring size=48	key-ring size=68	key-ring size=90

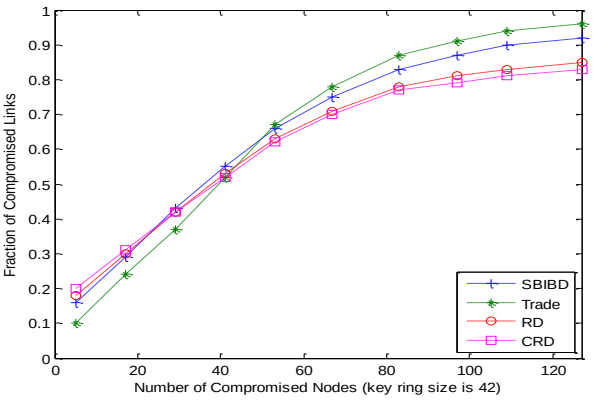
	Number of nodes	P_C	Number of nodes	P_C	Number of nodes	P_C	Number of nodes	P_C	Number of nodes	P_C
SBIBD	381	1	871	1	2257	1	4557	1	8011	1
Trade	762	0.335	1742	0.356	4514	0.381	9114	0.403	16022	0.425
RD	7620	0.860	26130	0.886	108336	0.908	309876	0.920	720990	0.927
CRD	7620	0.885	26130	0.916	108336	0.945	309876	0.960	720990	0.971

Table 5 Simulation results of different schemes in terms of resilience

KRS-CNN	10	25	40	55	70	85	100
SBIBD							
24	0.331	0.532	0.658	0.764	0.856	0.913	0.945
42	0.254	0.432	0.557	0.660	0.758	0.835	0.877
Trade							
24	0.275	0.523	0.657	0.772	0.884	0.947	0.975
42	0.176	0.374	0.520	0.671	0.783	0.872	0.913
RD							
24	0.346	0.521	0.637	0.734	0.812	0.885	0.910
42	0.241	0.426	0.533	0.630	0.715	0.783	0.817
CRD							
24	0.353	0.523	0.625	0.728	0.802	0.870	0.896
42	0.263	0.423	0.527	0.625	0.704	0.773	0.794



(a)



(b)

Fig. 5. Resilience Comparison. (a) Resilience of our proposed is compared with SBIBD, Trade and RD with same key-ring size k=24. (b) Resilience of our proposed is compared with other existing methods with same key-ring size k=42.

7. Conclusion

This paper proposed a developing and highly scalable key pre-distribution scheme for an IoT device. The block complementation theory was used to build a residual design for the first time. We showed that a mapping from residual design to key pre-distribution is needed to achieve profoundly high network scalability while at the same time degrading the key sharing probability. In Figs 3 to 5, we demonstrated comparison results of scalability, connectivity, and resilience of four methods (SBIBD, Trade, RD, and CRD) considering similar key-ring sizes. The maximum network scalability was obtained using the RD and CRD methods. Also, the proposed scheme is better than the other three schemes regarding network resilience. Our proposed scheme increases the probability of network connectivity more than the two methods, Trade and RD, but its connectivity is less than that of the SBIBD method.

8. References

1.

[1] T. Gomes, F. Salgado, S. Pinto, J. Cabral and A. Tavares, "A 6LoWPAN Accelerator for Internet of Things Endpoint Devices," *IEEE Internet of Things Journal*, vol. 5, no. 1, pp. 371-377, 2018, **DOI:** 10.1109/JIOT.2017.2785659.

[2] M. Husamuddin and M. Qayyum, "Internet of Things :A Study on Security and Privacy Threats," in *2nd International Conference on Anti-Cyber Crimes (ICACC)*, Abha, 2017, **DOI:** 10.1109/Anti-Cybercrime.2017.7905270.

[3] M. Saadeh, Z. Sleit, M. Qatawneh and W. Almobaideen, "Authentication Techniques for the

Internet of Things: A Survey," in *Cybersecurity and Cyberforensics Conference*, Amman, Jordan, 2016, **DOI:** 10.1109/CCC.2016.22.

[4] M. Malik, M. Dutta and J. Granjal, "A Survey of Key Bootstrapping Protocols Based on Public Key Cryptography in the Internet of Things," *IEEE Access,* vol. 7, pp. 27443 - 27464, 2019, **DOI:** 10.1109/ACCESS.2019.2900957.

[5] M. Anzani, H. Haj Seyyed Javadi and V. Modiri, "Key-management scheme for wireless sensor networks based on merging blocks of symmetric design," *Wireless Networks,* vol. 24, no. 8, p. 2867–2879, 2017, DOI:10.1007/s11276-017-1509-y.

[6] M. Tajeri, H. H. S. Javadi, M. Bayat and M. E. Shiri, "Pre-Distribution Encryption Key Scheme for Communicating between IoT Device Layer and Fog Layer," *Cybernetics and Systems,* pp. 1-25, 2022, DOI:10.1080/01969722.2022.2145665.

[7] T. Dargahi, H. H. S. Javadi and M. Hosseinzade, "Application-specific hybrid symmetric design of key pre-distribution for wireless sensor networks," *Security and Communication Networks,* vol. 8, no. 8, pp. 1561-1574, May 2015, DOI:10.1002/sec.1104.

[8] A. Pattanayak and B. Majhi, "Key pre-distribution schemes in distributed wireless sensor network using combinatorial designs revistied," 2009.

[9] S. Akhbarifar, H. Haj Seyyed Javadi, A. M. Rahmani and M. Hosseinzadeh, "Hybrid Key Pre-distribution Scheme Based on Symmetric Design," *Iranian Journal of Science and Technology, Transactions A: Science,* vol. 28, no. 39, p. 1–8, 2019, DOI:10.1007/s40995-019-00703-7.

[10] H. Haj Seyyed Javadi and M. Anzani, "Hybrid Key Pre-distribution Scheme for Wireless Sensor Network Based on Combinatorial Design," *Journal of Advances in Computer Engineering and Technology,* vol. 1, no. 3, pp. 33-38, 2015, DOI:10.13140/RG.2.2.13619.63520.

[11] S. H. Erfani, H. Haj Seyyed Javadi and A. M. Rahmani, "Analysis of Key Management Schemes in Dynamic Wireless Sensor Networks," *Advances in Computer Science: an International Journal,* vol. 4, no. 1, pp. 117-121, 2015.

[12] S. H. Erfani, H. Haj Seyyed Javadi and A. M. Rahmani, "A dynamic key management scheme for dynamic wireless sensor networks," *Security and Communication Networks,* vol. 8, no. 6, pp. 1040-1049, 2014, DOI:10.1002/sec.1058

[13] D. Chen, G. Chang, D. Sun, J. Jia and X. Wang, "Lightweight key management scheme to enhance the security of internet of things," *International Journal of Wireless and Mobile Computing,* vol. 5, no. 2, pp. 191-198, 2012, DOI:10.1504/IJWMC.2012.046773.

[14] H. Chan, A. Perrig and D. Song, "Random Key pre-distribution Schemes for Sensore Networks," in *Symposium on Security and Privacy*, Berkeley, CA, USA, 2003, **DOI:** 10.1109/SECPRI.2003.1199337.

[15] S. Qian, "A Novel Key Pre-distribution for Wireless Sensor Networks," in *International Conference on Solid State Devices and Materials Science*, 2012, DOI: 10.1016/j.phpro.2012.03.368.

[16] N. Solari Esfehani and H. Haj Seyyed Javadi, "A survey of key pre-distribution schemes based on combinatorial designs for resource-constrained devices in the IoT network," *Wireless Networks,* vol. 27, no. 4, pp. 3025-3052, 2021, DOI: 10.1007/s11276-021-02629-8.

[17] A. Morshed Aski and H. Haj Seyyed Javadi, "A novel key pre-distribution scheme based on -PBIBD combinatorial design in the resource-constrained IoT network," *arXiv preprint arXiv:2102.07137,* 2021, DOI:10.48550/arXiv.2102.07137.

[18] M.-L. Messai, "A Self-Healing Pairwise Key Pre-Distribution Scheme in IoT-based WSNs," in *International Wireless Communications and Mobile Computing (IWCMC)*, Marrakesh, Morocco, 2023, **DOI:** 10.1109/IWCMC58020.2023.10183198.

[19] B. Vinayaga Sundaram, M. Ramnath, M. Prasanth and J. Varsha Sundaram, "Encryption and Hash based Security in Internet of Things," in *3rd International Conference on Signal Processing, Communication and Networking (ICSCN)*, Chennai, 2015, **DOI:** 10.1109/ICSCN.2015.7219926.

[20] V. Chegeni, H. Haj Seyyed Javadi, M. R. Moazami Goudarzi and A. Rezakhani, "A scalable key pre-distribution scheme based on the unital design for the internet of things security," *IETE Journal of Research,* pp. 1-12, 2021, DOI: 10.1080/03772063.2021.1933626.

[21] F. K. Santoso and N. C. H. Vun, "Securing IoT for smart home system," in *International Symposium on Consumer Electronics (ISCE)*, Madrid, 2015, **DOI:** 10.1109/ISCE.2015.7177843.

[22] N. Sklavos and I. D. Zaharakis, "Cryptography and Security in Internet of Things (IoTs) : Models, Schemes and Implementations," in *8th IFIP International Conference on New Technologies, Mobility and Security (NTMS)*, Larnaca, 2016, **DOI:** 10.1109/NTMS.2016.7792443.

[23] M. Anzani, H. Haj Seyyed Javadi and A. Moeni, "A deterministic Key Predistribution Method for Wireless Sensor Networks Based on Hypercube Multivariate Scheme," *Iranian Journal of Science and Technology, Transactions A: Science,* vol. 42, no. 2, p. 777–786, June 2018, DOI:10.1007/s40995-016-0054-3.

[24] Y. Kim, E. Lim and T. Kwon, "On the Impact of Deployment Errors in Location-Based Key Pre-distribution Protocols for Wireless Sensor Networks," *IEEE Access,* vol. 12, pp. 35765 - 35778, 2024, **DOI:** 10.1109/ACCESS.2024.3372653.

[25] M. Javanbakht, H. Erfani, H. Haj Seyyed Javadi and P. Daneshjoo, "Key Pre-distribution Scheme for Clustered Hierarchical Wireless Sensor Networks based on Combinatorial Designs," *Security and Communication Networks,* vol. 7, no. 11, pp. 2003-2014, November 2014, DOI: 10.1002/sec.914.

[26] V. Modiri, H. Haj Seyyed Javadi and M. Anzani, "A Novel Scalable Key Pre-distribution Scheme for Wireless Sensor Networks Based on Residual Design," *Wireless Personal Communications,* vol. 96, no. 2, p. 2821–2841, September 2017, DOI:10.1007/s11277-017-4326-9.

[27] V. Modiri, H. Haj Seyyed Javadi and M. Anzani, "Using Residual Design for Key Management in

Hierarchical Wireless Sensor Networks," *Journal of Information Systems and Telecommunication (JIST),* vol. 8, no. 1, pp. 53-61, 2020, https://civilica.com/doc/1352392/.

[28] A. Kumar, N. Bansal and A. R. Pais, "New key pre-distribution scheme based on combinatorial design for wireless sensor networks," *IET Communications,* vol. 13, no. 7, p. 892 – 897, 2019, DOI:10.1049/iet-com.2018.5258.

[29] C. Y. Chen and H. C. Chao, "A survey of key distribution in wireless sensor networks," *Security and Communication Networks,* vol. 7, pp. 2495-2508, 2014, DOI: 10.1002/sec.354.

[30] G. Gaubatz, J.-P. Kaps, E. Ozturk and B. Sunar, "State of the art in ultra-low power public key cryptography for wireless sensor networks," in *Third IEEE International Conference on Pervasive Computing and Communications Workshops*, Kauai Island, 2005, **DOI:** 10.1109/PERCOMW.2005.76.

[31] S. Raza, H. Shafagh, K. Hewage, R. Hummen and T. Voigt, "Lithe: Lightweight Secure CoAP for the Internet of Things," *IEEE Sensors Journal,* vol. 13, no. 10, pp. 3711-3720, 2013, **DOI:** 10.1109/JSEN.2013.2277656.

[32] S. A. Camtepe and B. Yener, "Combinatorial Design of Key Distribution Mechanisms for Wireless Sensor Networks," *IEEE/ACM Transactions on Networking,* vol. 15, no. 2, pp. 346 - 358, April 2007, **DOI:** 10.1109/TNET.2007.892879.

[33] S. Ruj and B. Roy, "Key pre-distribution using combinatorial designs for grid-group deployment scheme in wireless sensor networks," *ACM Trans. Sensor Netw,* vol. 6, no. 4, pp. 1-4, Jan 2010, DOI:10.1145/1653760.1653764.

[34] A. Morshed Aski, H. Haj Seyyed Javadi and G. H. Shirdel, "A Full Connectable and High Scalable Key Pre-distribution Scheme Based on Combinatorial Designs for Resource-Constrained Devices in IoT Network," *Wireless Personal Communications,* vol. 114, no. 3, 2020, DOI: 10.1007/s11277-020-07466-0.

[35] A. Dey, Theory of block designs, J. Wiley, 1986.

[36] C. J. Colbourn and J. H. Dinitz, Handbook of combinatorial designs, CRC press, 2010.

[37] D. R. Stinson, Combinatorial Designs: Constructions and Analysis, New York: Springer, 2004, DOI: 10.1145/1466390.1466393.

[38] H. R. Sharifi, H. Haj Seyyed Javadi, A. Moeini and M. Hosseinzadeh, "Residual design of sink localization algorithms for wireless sensor networks," *Journal of High Speed Networks,* vol. 25, no. 1, pp. 87-99, 2019, DOI:10.3233/JHS-190605.

[39] P. Nikkhah Bahrami, H. Haj Seyyed Javadi, T. Dargahi, A. Dehghantanha and K. K. Raymond , "A Hierarchical Key Pre-Distribution Scheme for Fog Networks," *Concurrency and Computation: Practice and Experience,* 2018, DOI: 10.1002/cpe.4776.

[40] L. Eschenauer and V. D. Gligor, "A key-management scheme for distributed sensor networks," in *Proceedings of the 9th ACM conference on Computer and communications security*, Washington, DC, USA, 2002, DOI:10.1145/586110.586117.

[41] S. Ruj, A. Nayak and I. Stojmenovic, "Pairwise and Triple Key Distribution in Wireless Sensor Networks with Applications," *IEEE Transactions on Computers,* vol. 62, no. 11, pp. 2224 - 2237, 2013, **DOI:** 10.1109/TC.2012.138.

Journal of Optimization of Soft Computing (JOSC)

Vol. 2, Issue 2, pp: (36-43), Summer-2024

Journal homepage: http://josc.rimag.ir/en

Review paper

The relationship between important achievements in soft computing technology with the transportation and logistics industry in Iran: A review

Ali Shahabi[1, *]

1. Department of Industrial Engineering, Central Tehran Branch, Islamic Azad University, Tehran.

Article Info

Article History:
Received: 2024/08/24
Revised: 2024/09/25
Accepted: 2024/09/27

DOI:

Keywords:
Soft Computing; Transportation; Logistics; Digital Transformation; Communications.

**Corresponding Author's Email Address: eng_shahabi@yahoo.com*

Abstract

The transportation and logistics industry in Iran plays a pivotal role in the nation's economic development, yet it faces significant challenges, including inefficiencies and high operational costs. This review explores the critical relationship between advancements in soft computing technologies-such as fuzzy logic, neural networks, and genetic algorithms and their transformative impact on this sector. The necessity of this study arises from the urgent need for innovative solutions to enhance decision-making processes, optimize resource allocation, and improve overall service delivery within the Iranian transportation and logistics framework. This review synthesizes existing literature and case studies to highlight the novelty of integrating soft computing techniques in addressing complex logistical challenges. By employing a systematic methodology that includes qualitative analysis of recent advancements and their practical applications, this study identifies key trends and successful implementations within the industry. The findings reveal that soft computing technologies significantly enhance predictive modeling, route optimization, and demand forecasting, leading to improved efficiency and reduced costs. This research underscores the importance of embracing these technologies to foster sustainable growth in Iran's transportation and logistics sector, ultimately contributing to the broader economic landscape.

1. Introduction

The relationship between important achievements in soft computing technology and the transportation and logistics industry in Iran has become a crucial area of study, reflecting a significant evolution in operational efficiency and decision-making capabilities within this sector. Soft computing encompasses methodologies such as fuzzy logic, neural networks, and genetic algorithms, which have proven essential in optimizing logistics processes and enhancing intelligent industrial control systems. This integration has catalyzed substantial advancements in the transportation industry, facilitating digital transformation and improving service delivery across various logistical frameworks [1-3].

Historically, Iran's computing landscape has undergone profound changes over the last few decades, influenced by global technological trends and a growing emphasis on digitalization. The emergence of soft computing techniques has addressed challenges like information uncertainty and complex problem-solving in logistics operations, ultimately reshaping how transportation companies manage supply chains and respond to market demands [4, 5]. As these technologies gain traction, they have enabled improved data utilization, leading to better resource allocation and reduced operational costs [6].

Despite the benefits, the adoption of soft computing technologies is not without challenges. Issues such as data quality, interoperability among systems, and high initial implementation costs present significant hurdles that industry stakeholders must navigate [7, 8]. Moreover, research gaps persist regarding the operational-level integration of these technologies, which limits their full potential in optimizing logistics performance. As the industry moves forward, addressing these challenges will be pivotal in harnessing the capabilities of soft computing to foster a more efficient and sustainable transportation ecosystem in Iran [9, 10].

The ongoing digital transformation within the transportation and logistics industry heralds a promising future, as advancements in soft computing are expected to enhance operational practices and contribute to sustainable development. With an increasing focus on intelligent transportation systems and data analytics, the potential for innovation and growth within this sector is vast, positioning Iran as a key player in the evolution of logistics technology on a global scale [11-13].

Soft computing techniques have garnered significant attention in transportation and logistics due to their ability to address complex optimization problems and uncertainty [14, 15]. These methods, including fuzzy logic, neural networks, and evolutionary computing, offer flexibility and dynamism in supply chain management and tackle challenges in planning and decision-making [15, 16]. Applications of soft computing in maritime logistics, supply chain management, and industrial systems have shown promising results in improving efficiency and effectiveness [14, 17]. These techniques have been employed to enhance customer service levels, reduce operational costs, and maintain profit margins [17]. While soft computing has been widely applied in various aspects of logistics and transportation, some areas, such as customer relationship management and reverse logistics, remain underexplored, presenting opportunities for future research [17].

The present work provides an overview of the significant achievements in soft computing technologies and the necessity of utilizing them in the transportation and logistics industry in Iran. Conducting this research, being the first of its kind, will provide insights for decision-makers in the fields of transportation, logistics, and communications in Iran. The study includes a review of concepts, an examination of key soft computing methods used in the transportation and logistics industry, case studies conducted, the impact of soft computing technology on performance and cost improvement, a discussion of the challenges and limitations ahead, as well as an analysis and prediction of future trends.

2. Historical Context

The evolution of soft computing technologies has significantly impacted the transportation and logistics industry in Iran. Over the past several decades, key historical events and advancements in computing have shaped the industry's landscape.

2.1. Development of Computing in Iran

Iran's journey into modern computing began with its ancient empires, notably the Achaemenid Empire, which is recognized as one of the earliest civilizations to employ basic forms of calculation and record-keeping [1]. This historical foundation laid the groundwork for more complex computing systems that would emerge later.

2.2. Key Historical Events

The last 60 years have seen transformative developments in the logistics and supply chain sectors, influenced by major technological advancements. A pivotal study by Herold et al. (2021) highlights specific field-configuring events that facilitated the emergence and adoption of digitalization in these industries [2]. These events not only advanced computing methodologies but also provided insights into the evolution of a digitalization logic that has redefined operational frameworks.

2.3. Soft Computing Techniques

Soft computing techniques, including fuzzy logic, neural networks, and genetic algorithms, have emerged as essential tools for enhancing intelligent industrial control systems within the transportation and logistics sectors [3, 4]. These methodologies address the challenges faced in modern industrial environments, such as information acquisition and system optimization, thereby playing a crucial role in the evolution of the logistics landscape in Iran.

2.4. Current Implications

The integration of soft computing solutions into supply chain management has provenbeneficial in optimizing decision-making processes and improving efficiency [5]. As the logistics industry continues to evolve, the historical context of computing in Iranreveals a legacy of innovation that underpins current advancements in soft computing technology, ultimately contributing to the industry's growth and development.

3. Key Soft Computing Technologies

Soft computing encompasses a variety of methodologies designed to handle imprecise information and emulate complex human reasoning processes. In the context of the transportation and logistics industry in Iran, several key soft computing technologies have been pivotal in enhancing operational efficiency and decision-making capabilities.

3.1. Fuzzy Logic

Fuzzy logic is particularly useful in scenarios where information is uncertain or imprecise. It allows for the modeling of complex systems by incorporating linguistic variables and rules. Applications in the industry include fuzzy PI control for weigh belt feeders, which optimize material handling processes and improve accuracy in logistics operations [3]. The capability of fuzzy logic to deal with ambiguity makes it an ideal candidate for real-time decision-making in transportation systems [4].

3.2. Genetic Algorithms

Genetic algorithms (GAs) are optimization techniques inspired by natural selection that are particularly effective for solving complex problems with numerous variables. These algorithms have been applied to optimize logistics operations, including route planning and inventory management, significantly reducing costs and improving efficiency [3]. The integration of GAs with fuzzy systems has also shown promising results in enhancing decision-support mechanisms within logistics frameworks [4].

To provide a simple example, a genetic algorithm is used to optimally solve the following transportation problem in Figure 1. If there is a truck with a capacity of 50 cubic meters, which packages of various sizes can be placed in it to optimize the usable space?

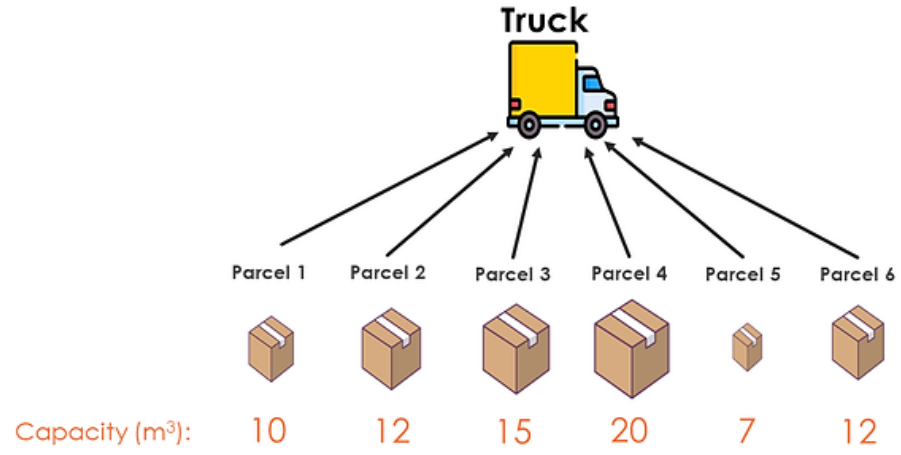


Figure 1. Example of Genetic Algorithm applied to transportation problem

3.3. Neural Computing

Neural computing employs artificial neural networks (ANNs) to learn from data and make predictions or decisions based on patterns. Recent advancements in neural computing have led to improved adaptive control systems in various applications, such as automated fault diagnosis and control in mechanical systems [3]. The adaptability of these systems enhances their utility in dynamic environments typical of transportation and logistics operations in Iran [4].

3.4. Hybrid Approaches

The combination of fuzzy logic, neural networks, and genetic algorithms forms hybrid approaches that leverage the strengths of each technique. For instance, fuzzy neural networks have been utilized to develop robust adaptive control systems for robot manipulators, showcasing their effectiveness in real-world applications such as automated warehousing and transportation systems [3]. This hybridization allows for the development of intelligent systems capable of tackling the multifaceted challenges faced in the transportation and logistics industry.

4. Case Studies

4.1. Digital Transformation in Railways

A significant trend in the railway sector is the integration of digitalization and intelligent technologies, which is pivotal for modernizing operations. This includes advancements in intelligent construction, equipment, and operational aspects. These initiatives are not merely about upgrading existing systems but are crucial for enhancing the safety, efficiency, and

overall competitiveness of the railway networks globally [6].

4.2. Challenges and Responses

In recent years, railway systems have faced multiple challenges such as the need for modernization in engineering projects, effective management of intelligent digital railway equipment, and addressing the increasing demand for both passenger and freight transport. To tackle these issues, the shift toward intelligent digital transformation has become essential. This transformation not only aims to elevate the travel experience for passengers but also catalyzes collaborative development within the railway sector on a global scale [6, 7].

4.3. Practical Applications

One illustrative case is the use of a survey conducted in Singapore, which utilized a Fleet Management System platform to monitor the activity of heavy goods vehicles. The study indicated challenges such as unverified stop records and an imbalance in activity reporting. To address this, a predictive model was developed, employing gradient-boosting techniques to classify activity types based on collected data. This integration of technology significantly improved the quality and efficiency of data collection and insights into vehicle movement patterns [7].

4.4. Future Predictions

The future of railway digitization suggests a continued emphasis on intelligent technologies that can enhance various facets of operations, including passenger services and freight logistics. Industry stakeholders are focusing on harnessing these advancements to not only improve service quality but also to facilitate more sustainable and efficient railway systems globally [6].

By leveraging such digital transformations, the railway sector is expected to overcome existing challenges while simultaneously setting a foundation for future innovations.

5. Impact on Efficiency and Cost-Effectiveness

The integration of soft computing technologies in the transportation and logistics industry in Iran has significantly enhanced operational efficiency and cost-effectiveness. By leveraging advancements in digital connectivity and the Internet of Things (IoT), companies have improved data-driven decision-making processes, which are crucial for optimizing production rates according to fluctuating customer demand. This shift in operational strategy allows for real-time adjustments, ensuring that production aligns closely with market needs and reducing inventory holding costs [8, 9].

5.1. Enhanced Data Utilization

The transition towards data-driven planning is pivotal in the optimization of logistics operations. Traditional supply chains often suffer from inefficiencies due to linear information flow, leading to the "bullwhip" effect, where minor fluctuations in demand can create substantial inventory variances further upstream [10]. However, with the application of soft computing methods, stakeholders can access real-time data, facilitating quicker responses to changes and minimizing disruptions across the supply chain. This results in improved service levels and better resource allocation, ultimately driving down operational costs [9-11].

5.2. Cost Considerations

Despite the advantages, the adoption of IoT and other advanced technologies presents challenges, particularly concerning high upfront costs. Manufacturing firms in Iran, similar to their global counterparts, face significant investments for implementing these solutions, which include expenses for systems integration, workforce training, and security measures. Nevertheless, the long-term benefits, such as improved efficiency and reduced operational costs, often outweigh these initial hurdles [8]. For instance, advancements in Radio Frequency Identification tracking have led to greater accuracy in inventory management, thus diminishing instances of stock-outs and excess inventory, which can be financially burdensome [9].

5.3. Strategic Decision-Making

As companies adopt these technologies, they also enhance their strategic decision-making capabilities. The improved visibility across various processes allows businesses to make informed decisions about resource allocation and operational adjustments, fostering a more agile and responsive supply chain environment [10]. By utilizing optimization algorithms that can process vast amounts of data, transportation and logistics firms are better equipped to handle dynamic market conditions, ultimately leading to more cost-effective operations [11].

The strategic decision-making process is shown in the Figure 2.

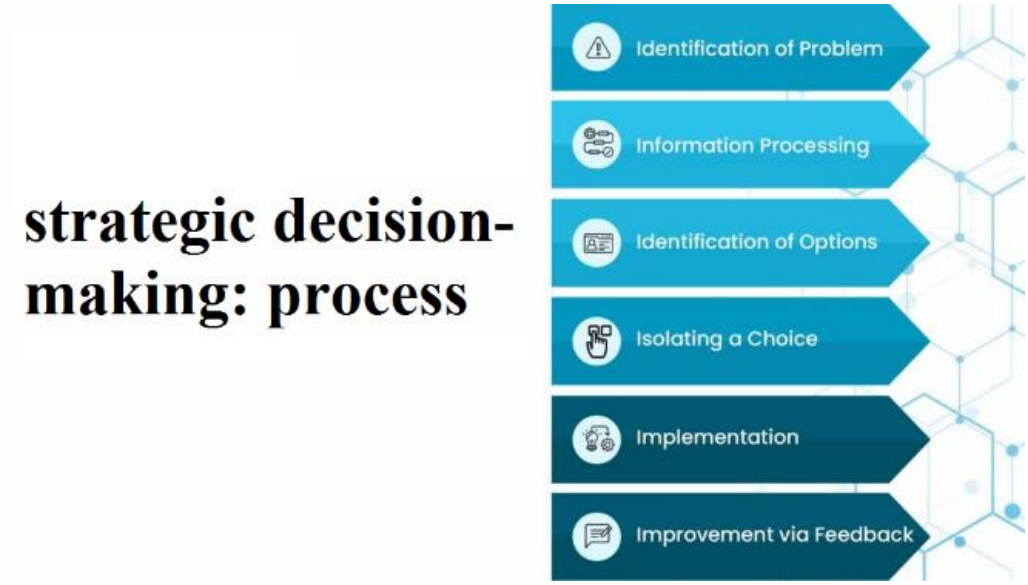


Figure 2. The strategic decision-making process

6. Challenges and Limitations

The integration of Intelligent Transportation Systems (ITS) in the transportation and logistics industry has introduced several significant challenges and limitations that impact the effectiveness of soft computing technologies in this sector. While ITS data has improved modeling accuracy concerning mobility's spatial and temporal characteristics, the application of optimization techniques remains hindered by various factors, including data quality, operational policies, and computational demands.

6.1. Data Quality Considerations

A primary challenge in leveraging ITS data is ensuring data quality. Many studies have indicated that the integrity and reliability of the data used in modeling can greatly influence the outcomes of research and implementation. Inadequate data quality can stem from privacy concerns and restrictive data-sharing policies adopted by certain transportation operators, limiting the availability of comprehensive datasets for analysis [11, 12]. Moreover, although simulation is frequently used to verify control strategies, such environments often overlook the technical challenges associated with real-time implementations, which can skew results and hinder practical applications [11].

6.2. Operator Data-Sharing Policies

The variation in operators' data-sharing policies poses another barrier. While some entities have adopted a more open approach, fostering ITS-related research, many others maintain restrictive practices that further complicate access to crucial data. This lack of data availability not only stifles research opportunities but also complicates the design and operational planning of public transportation systems, as identified by various studies [11].

6.3. Computational Burden

The computational burden associated with processing and analyzing ITS data is significant. The integration of optimization techniques with large datasets requires considerable computational resources and advanced algorithms, which may not be readily available to all researchers or agencies. This limitation can impede the timely and effective deployment of ITS applications in transportation and logistics, preventing optimal decision-making and strategic planning [11].

6.4. Interoperability Issues

Interoperability among different systems and agencies remains a critical concern. The lack of common standards and specifications for data inputs can lead to inefficiencies and hinder collaboration among various stakeholders in the transportation sector. A cohesive approach to establishing interoperability standards could enhance research and practical applications in ITS, yet achieving this remains a complex task due to existing disparities in technological adoption and infrastructure capabilities [11].

6.5. Overall Research Gaps

Lastly, there is a notable gap in the literature addressing design-related and operational-level problems in the context of ITS. While tactical and real-time issues have been explored extensively, the strategic level planning that incorporates

broader research insights is still underrepresented [11, 12]. Addressing these research gaps is essential for developing more robust decision-support tools and improving the overall effectiveness of ITS in the transportation and logistics industry.

7. Future Trends

The future of the transportation and logistics industry in Iran is poised for significant transformation, largely driven by advancements in soft computing technologies. As digitalization continues to reshape various sectors, a deeper understanding of how these technologies can be leveraged may provide insights into emerging opportunities and challenges in logistics management [2].

7.1. Digitalization and Sustainability

One of the key trends is the integration of soft computing methods to enhance sustainability in inventory management. Recent studies highlight the growing interest in applying IoT technologies to achieve sustainable practices within the green supply chain. This trend indicates a shift towards more eco-friendly approaches in logistics, driven by the necessity of reducing carbon footprints and improving operational efficiency [9, 13].

7.2. Intelligent Transportation Systems

Moreover, the railway sector in Iran, similar to global trends, is expected to adopt intelligent digital transformation strategies to improve efficiency and safety. This includes optimizing maintenance through AI applications and robotics, which can facilitate a significant reduction in operational costs and increase service reliability. As the industry moves towards more automated systems, the role of human operators may evolve, leading to a workforce that closely collaborates with advanced technologies [6, 18]. Figure 3 shows the components of Intelligent Transportation Systems.

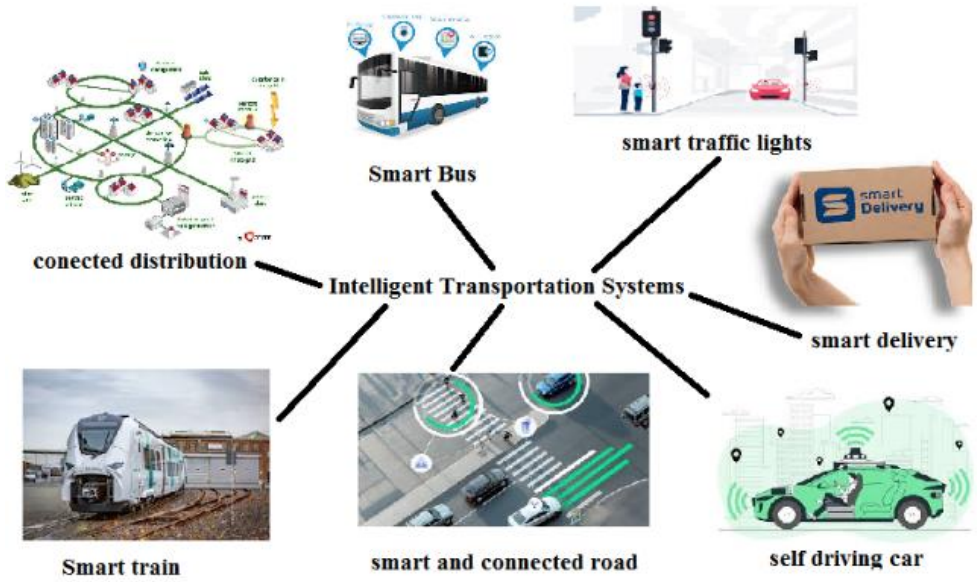


Figure 3. Components of Intelligent Transportation Systems

7.3. Data Utilization and Analytics

The logistics sector is also set to benefit from enhanced data analytics capabilities, fueled by the massive volumes of data generated daily. Transportation companies will increasingly harness big data and machine learning algorithms to optimize routing, manage fleets, and improve overall service delivery. Effective integration of these technologies is crucial for achieving high performance and return on investment, particularly as the complexity of logistics operations continues to grow [19, 20].

7.4. Research and Development Focus

Future research is expected to concentrate on bridging existing gaps in soft computing applications within the logistics field. Areas such as the coordination of product flows, green logistics strategies, and the development of adaptive supply chain systems are likely to gain prominence. By focusing on these aspects, companies can enhance their responsiveness to market demands and environmental considerations [3-5].

8. Conclusion

The exploration of the relationship between advancements in soft computing technology and the transportation and logistics industry in Iran reveals a transformative potential that can significantly enhance operational efficiency and cost-effectiveness. As outlined in the historical context, the evolution of computing in Iran has set the stage for the adoption of innovative methodologies such as fuzzy logic, genetic algorithms, and neural computing. These techniques have proven instrumental in addressing complex logistical challenges, optimizing resource allocation, and improving decision-making processes.

The case studies presented illustrate successful implementations of soft computing technologies, particularly in the context of digital transformation within the railway sector. These examples underscore the capacity of these technologies to not only streamline operations but also to respond dynamically to the evolving demands of the industry. However, while the implications are promising, challenges such as data quality, interoperability, and computational burdens remain critical hurdles that must be addressed to fully realize the benefits of these advancements.

Looking ahead, the integration of intelligent transportation systems and enhanced data utilization will be paramount for driving future innovations in the sector. By embracing digitalization and fostering a culture of data-sharing among operators, the Iranian transportation and logistics industry can position itself for sustainable growth.

In conclusion, this research highlights a clear pathway for leveraging soft computing technologies to overcome existing limitations and seize opportunities for improvement. Continued investment in these areas will not only bolster the efficiency and effectiveness of transportation and logistics operations in Iran but will also contribute to the broader economic landscape, paving the way for a more resilient and competitive industry.

It is suggested that for the continuation of this work, a transportation problem be analyzed in detail using soft computing techniques, and the impact of these methods on performance, execution time, cost optimization, and so on be examined.

References

[1] Heidari, A., Jafari Navimipour, N. and Unal, M., 2022. The history of computing in Iran (Persia)-since the Achaemenid Empire. Technologies, 10(4), p. 94. https://doi.org/10.3390/technologies10040094

[2] Herold, D.M., Ćwiklicki, M., Pilch, K. and Mikl, J., 2021. The emergence and adoption of digitalization in the logistics and supply chain industry: an institutional perspective. Journal of Enterprise Information Management, 34(6), pp.1917-1938. https://doi.org/10.1108/JEIM-09-2020-0382

[3] Wu, Z., Liao, H., Lu, K. and Zavadskas, E.K., 2021. Soft computing techniques and their applications in intelligent industrial control systems: A survey. International Journal of Computers Communications & Control, 16(1), p. 4142. https://etalpykla.vilniustech.lt/handle/123456789/151666

[4] Kharola, S., Ram, M., Mangla, S.K. and Kazancoglu, Y., 2023. Advances in soft computing applications, 1st Edition, New York. River Publishers. https://api.taylorfrancis.com/content/books/mono/download?identifierName=doi&identifierValue=10.1201/9781003425885&type=googlepdf

[5] Jauha, S.K. and Pant, M., 2013. Recent trends in supply chain management: A soft computing approach. In Proceedings of Seventh International Conference on Bio-Inspired Computing: Theories and Applications (BIC-TA 2012), 2, pp. 465-478. https://doi.org/10.1007/978-81-322-1041-2_40

[6] Li, P., Xue, R., Shao, S., Zhu, Y. and Liu, Y., 2023. Current state and predicted technological trends in global railway intelligent digital transformation. Railway Sciences, 2(4), pp.397-412. https://doi.org/10.1108/RS-10-2023-0036

[7] Oladimeji, D., Gupta, K., Kose, N.A., Gundogan, K., Ge, L. and Liang, F., 2023. Smart transportation: an overview of technologies and applications. Sensors, 23(8), p.3880. https://doi.org/10.3390/s23083880

[8] Kalsoom, T., Ahmed, S., Rafi-ul-Shan, P.M., Azmat, M., Akhtar, P., Pervez, Z., Imran, M.A. and Ur-Rehman, M., 2021. Impact of IOT on Manufacturing Industry 4.0: A new triangular systematic review. Sustainability, 13(22), p.12506. https://doi.org/10.3390/su132212506

[9] Mashayekhy, Y., Babaei, A., Yuan, X.M. and Xue, A., 2022. Impact of Internet of Things (IoT) on inventory management: A literature survey. Logistics, 6(2), p.33. https://doi.org/10.3390/logistics6020033

[10] Hatami, A., Sasanpour, F., Asadzadeh, H. and van Bodegom, P.M., 2023. Scenario analyses to reach smart sustainability in Tehran. Journal of Urban Management, 12(4), pp. 385-397. https://doi.org/10.1016/j.jum.2023.09.002

[11] Iliopoulou, C. and Kepaptsoglou, K., 2019. Combining ITS and optimization in public transportation planning: state of the art and future research paths. European Transport Research Review, 11, p. 27. https://doi.org/10.1186/s12544-019-0365-5

[12] Najafi, M., Keymanesh, M., Moshrefi, R. and Zefreh, M.M., 2019. Impact of Economic Inputs on Production of Transportation Services in Iran. International Journal of Transportation Engineering,

7(1), pp.77-89. https://doi.org/10.22119/ijte.2018.101785.1371
[13] Sumbal, M.S., Ahmed, W., Shahzeb, H. and Chan, F., 2023. Sustainable Technology Strategies for Transportation and Logistics Challenges: An Implementation Feasibility Study. Sustainability, 15(21), p.15224. https://doi.org/10.3390/su152115224
[14] Expósito-Izquierdo, C., Melián-Batista, B. and Moreno-Vega, J.M., 2018. A review of soft computing techniques in maritime logistics and its related fields. Soft Computing Based Optimization and Decision Models: To Commemorate the 65th Birthday of Professor José Luis" Curro" Verdegay, pp. 1-23. https://doi.org/10.1007/978-3-319-64286-4_1
[15] Brito, J., Castellanos-Nieves, D., Expósito, A. and Moreno, J.A., 2018. Soft computing methods in transport and logistics. Soft Computing Based Optimization and Decision Models: To Commemorate the 65th Birthday of Professor José Luis" Curro" Verdegay, pp. 45-61. https://doi.org/10.1007/978-3-319-64286-4_3
[16] Dote, Y. and Ovaska, S.J., 2001. Industrial applications of soft computing: a review. Proceedings of the IEEE, 89(9), pp. 1243-1265. https://doi.org/10.1109/5.949483
[17] Ko, M., Tiwari, A. and Mehnen, J., 2010. A review of soft computing applications in supply chain management. Applied Soft Computing, 10(3), pp. 661-674. https://doi.org/10.1016/j.asoc.2009.09.004
[18] Li, X., 2024. Optimization of logistics flow management through big data analytics for sustainable development and environmental cycles. Soft Computing, 28(3), pp.2701-2717. https://doi.org/10.1007/s00500-023-09591-x
[19] Yaiprasert, C. and Hidayanto, A.N., 2024. AI-powered ensemble machine learning to optimize cost strategies in logistics business. International Journal of Information Management Data Insights, 4(1), p. 100209. https://doi.org/10.1016/j.jjimei.2023.100209
[20] Torre-Bastida, A.I., Del Ser, J., Laña, I., Ilardia, M., Bilbao, M.N. and Campos-Cordobés, S., 2018. Big Data for transportation and mobility: recent advances, trends and challenges. IET Intelligent Transport Systems, 12(8), pp. 742-755. https://doi.org/10.1049/iet-its.2018.5188

Nomenclature	
GAs	Genetic Algorithms
ANNs	Artificial Neural Networks
IoT	Internet of Things
ITS	Intelligent Transportation Systems

Journal of Optimization of Soft Computing (JOSC)

Vol. 2, Issue 2, pp: (44-50), Summer-2024

Journal homepage: http://josc.rimag.ir/en

Research paper

A Brief Review of Different Methods of Building Energy Optimization in Hot & Humid Malaysian Climate

Seyed Mohammad Noorbakhsh [1], Heidar Ali Raeisi [*, 1], Behrang Moradi [2], Mohd Hamdan Ahmad [3, 4]

1. Energy and Environment Research Center, Shahrekord Branch, Islamic Azad University, Shahrekord, Iran.
2.Department of Architecture, South Tehran Branch, Islamic Azad University, Tehran, Iran
3. Institute of High Voltage and High Current, Faculty of Electrical Engineering, Universiti Teknologi Malaysia, Johor Bahru, Johor, Malaysia
4. Faculty of Built Environment and Surveying, Universiti Teknologi Malaysia

Article Info

Article History:
Received: 2024/08/27
Revised: 2024/09/25
Accepted: 2024/09/27

DOI:

Keywords:
Energy Optimization; Malaysian Climate; Passive Design Strategies; Active Design Technologies; Sustainability.

**Corresponding Author's Email Address: h.raisi@gmail.com*

Abstract

The increasing demand for energy efficiency in the built environment has prompted a critical examination of energy optimization methods, particularly in building such as those found in Malaysia. This article addresses the urgent necessity of investigating energy optimization strategies tailored to the unique climatic conditions of Malaysia, where temperature and humidity variations significantly influence energy consumption patterns in buildings. Despite the growing body of research on energy efficiency, there remains a notable gap in comprehensive studies that systematically review and compare the effectiveness of different optimization techniques in Malaysian contexts. Our investigation highlights innovative approaches, including passive and active design strategies, advanced building materials, and smart technology integration, which have shown promise in enhancing energy performance. Key results indicate that region-specific optimization methods can lead to substantial reductions in energy usage, with potential savings in certain climates. This review not only contributes to the existing literature by filling a critical gap but also provides practical insights for architects, engineers, and policymakers aiming to implement sustainable building practices in Malaysia's diverse climatic landscape.

1. Introduction

Energy optimization in buildings across hot & humid Malaysian climate is a critical area of study, given the nation's predominantly tropical environment characterized by high temperatures and humidity year-round. Malaysia's unique climate, influenced by both monsoon seasons and regional variations, necessitates tailored energy strategies to maintain indoor comfort and reduce energy consumption. As Malaysia undergoes rapid urbanization and development, understanding and implementing effective energy optimization methods in building design and operation become increasingly crucial to achieving sustainability and reducing environmental impact. In Malaysian climates, energy optimization methods span both passive and active design strategies. Passive strategies focus on leveraging natural resources to enhance building performance, including optimal building orientation, natural ventilation, and daylight capture. These methods are vital in minimizing heat ingress and reducing reliance on artificial lighting, which can contribute to substantial energy savings.

Humidity management is also a critical component, given the tropical climate's challenges, which necessitate innovative solutions like dehumidifiers

and vegetation to maintain indoor air quality and comfort.

Active design strategies in Malaysia incorporate advanced technologies and systems, such as hybrid heat pump systems and building-integrated photovoltaic/thermal (BIPV/T) systems, to improve energy efficiency. These technologies enable buildings to generate renewable energy, significantly reducing their carbon footprint. Additionally, energy optimization strategies involve the use of simulation tools and optimization techniques, such as the Taguchi Orthogonal Arrays and Response Surface Methodology, to design nearly zero-energy buildings (NZERBs) tailored to tropical climates. Smart operation systems further enhance building performance by dynamically adjusting energy use based on real-time conditions.

Case studies of energy optimization in Malaysian buildings provide valuable insights into effective strategies and their impacts. For instance, studies on government office buildings in Kuala Lumpur highlight the inefficiencies in energy consumption and recommend retrofits based on non-design and passive design factors. Comparative analyses of different architectural forms underscore the importance of design in achieving thermal comfort and energy efficiency. Furthermore, the integration of energy simulation tools in evaluating retrofit measures demonstrates their effectiveness in reducing energy consumption and enhancing sustainability.

The future of energy optimization in Malaysian buildings is poised to benefit from advanced optimization techniques, hybrid renewable energy systems, and supportive policy frameworks. The integration of smart technologies and continued investment in renewable energy infrastructure are expected to drive significant improvements in energy efficiency and sustainability. Long-term projections indicate a substantial increase in the global share of renewable energy, emphasizing the need for Malaysia to adopt innovative energy optimization strategies to meet future energy demands sustainably.

2. Overview of Malaysian Climates

Malaysia's climate is predominantly characterized by its tropical nature, marked by high temperatures and humidity year-round. The main variable influencing the climate is not temperature or air pressure but rainfall. Coastal plains in Malaysia average temperatures around 28°C, while inland and mountainous areas average around 26°C, and higher mountain regions around 23°C. Relative humidity typically ranges between 70% and 90% [1].

The climate can be divided into two main seasons influenced by monsoons: the dry season, which spans from June to September, and the rainy season, from December to March. Western and northern parts of Malaysia receive the most precipitation due to the moisture-laden monsoon clouds that move north- and westward [1]. Typhoons can also impact Malaysia from July to mid-November, causing heavy damage, flooding, and erosion [1].

The climate in Malaysia can further be broken down into regional specifics. West Malaysia, which includes major cities like Kuala Lumpur, Ipoh, and Georgetown, experiences relatively constant temperatures and humidity levels throughout the year. The summer months (May to August) see slightly less precipitation compared to other periods. Conversely, the east coast of West Malaysia exhibits a more pronounced dry and rainy season cycle, with heavy rainfall from December to February, leading to closures of many hotels and tourist attractions, while the period from March to October remains relatively dry and attracts more visitors [2].

Kuala Lumpur is noted as the warmest and rainiest part of Malaysia, whereas Kelantan is considered the coldest [3]. The best time for travel is generally in February, due to less rainfall. Most precipitation occurs from October to December [3]. Over the years, consistent data from weather stations have revealed notable temperature trends; for instance, the hottest month recorded was May 1998 at 29.2°C, and October 2023 was one of the driest months in 13 years [3].

East Malaysia, located on the island of Borneo, is influenced by a tropical monsoon climate, with a distinct short dry season and a longer rainy season with substantial rainfall. This climatic distinction results from its proximity to the equator and monsoon influences, creating significant variability in weather patterns across different regions of the country [2].

3. Energy Optimization Methods

Energy optimization in buildings, particularly in the context of Malaysia's tropical climate, involves various strategies aimed at reducing energy consumption while maintaining comfort and functionality. These methods incorporate both passive and active design elements to achieve optimal energy efficiency.

3.1. Passive Design Strategies

Passive design strategies focus on using natural resources to enhance building performance and minimize energy consumption. These methods emphasize the importance of building orientation, natural ventilation, and daylight capture.

3.1.1. Building Orientation and Natural Ventilation

Proper orientation of buildings is crucial for minimizing heat ingress, especially through the east and west façades due to the consistent path of the sun throughout the year in Malaysia [4]. Natural ventilation can be optimized by designing buildings with larger openings for inlets and outlets, enabling free wind passage, and incorporating multiple floors to enhance airflow [5].

3.1.2. Daylight Capture

Daylight capture is a significant component of passive design, reducing the need for artificial lighting and associated thermal gains. Indirect natural light can be harnessed through windows, skylights, and light reflectors, which should be carefully designed to balance illumination and glare while restricting unwanted heat penetration [5].

However, excessive daylight can lead to glare and overheating, necessitating the use of metrics to optimize these competing demands [4].

3.1.3. Humidity Management

Humidity poses a challenge in tropical climates, affecting the feasibility of natural ventilation. Solutions include the use of dehumidifiers or incorporating vegetation to manage indoor humidity levels [4]. Additionally, rain and tropical weather conditions must be considered, which can be addressed through wind-driven rain simulations and the design of appropriate shading devices [4].

3.2. Active Design Strategies

Active design strategies involve the use of technologies and systems to further enhance energy efficiency in buildings. These methods include hybrid heat pump systems, BIPV/T systems, and energy optimization strategies.

3.2.1. Renewable Energy Integration

Hybrid heat pump systems that integrate renewable energy sources, such as solar and thermal energy, are essential for optimizing energy consumption in buildings [6]. BIPV/T systems combine photovoltaic panels with thermal collectors to generate both electricity and heat, improving overall energy efficiency [6].

3.2.2. Energy Optimization and Smart Operation

Energy optimization strategies involve the use of simulation tools and optimization techniques to design energy-efficient buildings. For instance, the use of Taguchi Orthogonal Arrays and RSM-Box-Behnken methods can identify significant factors influencing energy performance and develop prescriptive requirements for NZERBs in tropical climates [7]. Smart operation and demand energy response systems can further enhance building performance by dynamically adjusting energy use based on real-time conditions [6].

3.2.3. Multi-Objective Optimization

Multi-objective optimization approaches are used for retrofitting buildings to achieve net-zero energy status. These approaches consider various factors such as energy efficiency, cost, and environmental impact to make informed decisions [8]. For example, optimizing air conditioning operational modes in residential buildings with the assistance of solar energy and thermal energy storage can significantly reduce energy consumption [8].

4. Case Studies

The analysis of energy optimization methods in Malaysian buildings has been explored through various case studies, providing valuable insights into effective strategies and their impacts.

4.1. Kuala Lumpur Government Office Buildings

A comprehensive study focused on six government office building blocks located in Kuala Lumpur, Malaysia's capital city [9]. The study conducted a literature review and collected energy consumption data to calculate the Building Energy Index (BEI), comparing it against the MS1525:2019 and GBI benchmarks to evaluate energy performance. SketchUp software was utilized to illustrate solar radiation and sun path diagrams. Recommendations were derived based on non-design factors and passive design strategies [9].

4.1.1. Findings and Recommendations

The study found that Malaysian government office buildings tend to consume energy inefficiently due to a lack of optimization measures. The research highlighted the necessity of studying the energy performance of existing buildings constructed before the implementation of energy-efficient standards to mitigate wastage [9]. Retrofits based on non-design and passive design factors were recommended, emphasizing the need for effective

energy conservation measures to improve the energy performance of these buildings [9].

4.2. Daylight and Thermal Comfort in Architectural Modelling

Another case study compared different architectural forms regarding thermal comfort, Energy Efficiency Index, and energy consumption. Case 3 provided more sufficient natural daylight than Case 4, indicating that geometrical design plays a crucial role in energy optimization [10]. The study also considered simplicity, popularity, and solar potential in developing architectural alternatives [10].

4.2.1. Energy-Efficient Retrofitting

The retrofitting of existing government high-rise office buildings in Malaysia showed that various levels of intervention could achieve between 4% to 7% savings in annual energy consumption. These interventions demonstrated compliance with BEI benchmark margins of the GBI and EPU standards [11]. This validated model highlighted the effectiveness of energy retrofit measures in reducing energy consumption in high-rise office buildings [11].

4.3. Energy-Efficient Building Construction

The evaluation of energy-efficient building construction and embodied energy emphasized passive strategies as more effective in reducing energy consumption. A combined approach with active strategies, however, yielded optimal results. The challenges of retrofitting, such as initial costs and regulatory barriers, were noted. Tools like EnergyPlus, eQUEST, and IES VE were identified as essential for evaluating and identifying cost-effective retrofit measures in building performance [12].

4.3.1. Key Insights

The importance of energy simulation software in assessing retrofit measures was underscored. These tools provide comprehensive insights into various strategies available for achieving energy efficiency and sustainability goals. The review serves as an authoritative resource for building owners, managers, and professionals, offering guidance for informed decision-making in retrofitting practices [12].

5. Comparative Analysis

5.1. Sensitivity Analysis

Sensitivity analysis plays a crucial role in examining the impact of varying specific variables on the output performance of an energy system [13]. This analysis is especially significant for optimizing energy use in rural areas where grid power is inaccessible. For example, in Kuala Terengganu, sensitivity analysis was conducted after simulation and optimization processes. The findings indicate that for a hybrid PV-wind-grid system to achieve a payback period comparable to that of a PV-grid system, the feed-in tariff rate needs to be set at a minimum of RM1.80. Therefore, it can be concluded that the grid-connected hybrid PV-wind system represents the most economically viable option [14].

5.2. Statistical Validation

Regression analysis is another critical tool used for optimizing energy efficiency, particularly in models where environmental and electrical parameters are examined. High correlation coefficients (r values ranging from 0.8168 to 0.9803) indicate a significant relationship between these parameters. The predictive power of these models is further validated through statistical indicators such as R^2, Mean Bias Error, Root Mean Square Error, Mean Absolute Percentage Error (MAPE), and Symmetric Mean Absolute Percentage Error (SMAPE). Among these, SMAPE is often considered more reliable than MAPE, particularly when zero or near-zero data points are present [15].

5.3. Climatic Considerations

Malaysia's hot & humid climate, characterized by high temperatures and significant humidity, presents unique challenges and opportunities for energy optimization. The country's climate, influenced by both the Southwest (May to September) and Northeast (November to March) monsoons, sees consistent annual rainfall without a defined dry season [16, 17]. These climatic conditions necessitate specialized strategies for energy optimization, particularly in building designs and landscape planning.

5.4. Energy Efficiency in Tropical Buildings

The form and design of buildings play a crucial role in energy efficiency, especially in tropical climates like Malaysia. Research from Penang, Malaysia, underscores the significance of building form in optimizing energy use. For instance, buildings designed to harness natural ventilation and shade can significantly reduce energy consumption by mitigating the heat load [10]. Furthermore, planting trees with high shade qualities around buildings can serve as natural air conditioners, reducing the temperature and improving microclimates through the process of evapotranspiration [18].

5.5. Environmental Impact

Malaysia's biodiversity and forest ecosystems have faced degradation due to factors like agricultural expansion. This environmental impact directly influences the strategies for energy optimization in buildings. For example, deforestation and land use changes can affect local climates, thereby impacting the energy needs and efficiency measures required for buildings [19, 20]. Efforts to integrate climate resilience into national planning have been evident in Malaysia's Tenth and Eleventh National Plans, aimed at enhancing climate resilience and sustainable development [19].

6. Challenges and Considerations

Green buildings have emerged as a solution to reduce the negative impacts of development while providing positive impacts on the environment and residents. Designed to minimize water and energy use and to improve air and water quality, green buildings offer a host of benefits. Although their upfront costs are higher, the long-term benefits include reduced energy and maintenance costs, as well as improved quality of life for users and residents [21]. This aligns with Malaysia's commitment to reducing carbon emissions and improving energy efficiency, contributing to a healthier environment with cleaner air and reduced reliance on natural resources [22].

7. Future Trends in Energy Optimization

7.1. Advanced Optimization Techniques

In the pursuit of designing NZERBs in tropical climates, innovative optimization techniques are increasingly employed. Two notable approaches include the Taguchi Orthogonal Arrays and the Response Surface Methodology (RSM) with Box-Behnken design. The Taguchi Orthogonal approach is particularly favored for its reduced computing time, aiding in the selection of the most influential factors. Subsequently, the RSM-Box-Behnken technique, integrated with the Design-Expert tool, is utilized to develop prescriptive path requirements for NZERB design. These methods have demonstrated substantial accuracy and reliability, as evidenced by energy simulation tool HAP predicting a BEI of 67.085 kWh/m²/year - a 54% reduction compared to the base-case building design - and the PVWatts calculator predicting a 13 kW system, which is 53.57% smaller than the base case [7].

7.2. Hybrid Renewable Energy Systems

Hybrid Renewable Energy Systems are poised to play a critical role in energy optimization for remote and rural areas. Studies have explored various combinations of hybrid systems, such as photovoltaic-wind turbine-diesel-battery configurations, using the Hybrid Optimization Model for Electric Renewable (HOMER) simulation. These configurations have shown promise in providing sustainable and cost-efficient solutions for rural electrification. For instance, an analysis in Kuala Terengganu, Malaysia, indicated that PV-wind-grid system has the lowest NPC which is approximately 16% lower than the NPC of PV-grid system configuration [14].

7.3. Policy and Regulatory Support

The future of energy optimization is also heavily dependent on policy and regulatory frameworks. The introduction of carbon pricing and improved carbon accounting practices are essential steps to incentivize renewable energy adoption and reduce carbon emissions. Governmental support through explicit and predictable policies can facilitate the transition towards low-carbon futures in sectors such as property and construction. Regulations, incentives, and improved governance are critical to reducing energy consumption, waste, and carbon emissions from construction materials. Furthermore, initiatives like the science-based targets initiative help differentiate genuine decarbonization efforts from corporate greenwashing, ensuring that real decarbonization is achieved across the economy [22].

7.4. Integration of Smart Technologies

The integration of smart technologies, such as BIPV/T systems and hybrid heat pump systems, is another promising trend.

These technologies enable efficient energy optimization strategies and support smart operation and demand energy response, making buildings more energy-efficient and reducing their overall carbon footprint [6].

7.5. Long-term Projections and Goals

Long-term projections indicate that the global share of renewable energy in total energy consumption is expected to rise significantly, from 16.6% in 2010 to 21% by 2030, with renewable energy sources accounting for 45% of global electricity generation by 2040. This growth is likely to be driven primarily by solar, wind, and hydropower technologies [13]. These trends underscore the importance of continued investment in renewable energy infrastructure and the adoption

of advanced energy optimization techniques to meet future energy demands sustainably.

8. Conclusions

In summary, this investigation into energy optimization methods in building design within the Malaysian climate underscores the critical necessity of addressing energy efficiency in the face of rapid urbanization and environmental challenges. The unique climatic conditions of Malaysia, characterized by high humidity and temperature fluctuations, necessitate innovative approaches to building design that can effectively balance occupant comfort with energy consumption.

The exploration of both passive and active design strategies reveals significant opportunities for enhancing energy performance. Passive strategies such as optimal building orientation, natural ventilation, daylight capture, and humidity management provide foundational benefits, while active strategies—including renewable energy integration, smart operational techniques, and multi-objective optimization—offer advanced solutions for maximizing energy efficiency. The case studies conducted, particularly those focused on Kuala Lumpur's government office buildings and energy-efficient retrofitting practices, highlight practical applications of these strategies and yield actionable recommendations for future projects.

Despite the wealth of existing literature on energy optimization, a notable scientific gap remains in the application of these principles specifically tailored to Malaysia's tropical climate. This research uses a strong hybrid approach, combining the expression of different optimization methods and the review of qualitative case studies to fill this gap and provide a comprehensive understanding of effective strategies.

The results of this study demonstrate the potential for significant energy savings and improved environmental impact through the integration of innovative design practices. The comparative analysis reveals the importance of considering climatic factors in energy-efficient design while also addressing challenges such as regulatory support and technological integration.

Looking ahead, future trends in energy optimization must focus on advanced optimization techniques, hybrid renewable energy systems, and the incorporation of smart technologies to achieve long-term sustainability goals. As policymakers and industry stakeholders continue to navigate the complexities of energy efficiency in building design, this review research serves as a vital resource for fostering a more sustainable built environment in Malaysia. Ultimately, this review research advocates for a holistic approach that prioritizes both innovation and practicality, ensuring that the architectural solutions of today contribute meaningfully to the ecological and economic resilience of tomorrow.

References

[1] Malaysia, Weather Online, 2024. https://www.weatheronline.co.uk/reports/climate/Malaysia.htm [Accessed 07 August 2024]

[2] Climate Malaysia, Travel Climate, 2024. https://www.travelclimate.net/Climate/Malaysia [Accessed 07 August 2024]

[3] The climate in Malaysia, World Data, 2024. https://www.worlddata.info/asia/malaysia/climate.php [Accessed 07 August 2024]

[4] Green Moves: At the forefront of passive design, The Edge Malaysia, 2024. https://theedgemalaysia.com/article/green-moves-forefront-passive-design [Accessed 07 August 2024]

[5] Bulbaai, R. and Halman, J.I., 2021. Energy-efficient building design for a tropical climate: A field study on the caribbean island curaçao. Sustainability, 13(23), p. 13274.

[6] Montagud, C., 2023. Special Issues: Energy Efficiency and Optimization Strategies in Buildings for a Sustainable Future. Energies, https://www.mdpi.com/journal/energies/special_issues/Energy_Efficiency_and_Optimization_Strategies_in_Buildings_for_a_Sustainable_Future [Accessed 07 August 2024]

[7] Sharif, H.Z., 2021. Optimal design of net zero energy residential buildings in tropical climate, Doctoral dissertation, Universiti Tun Hussein Onn Malaysia. http://eprints.uthm.edu.my/8419 [Accessed 07 August 2024]

[8] Chung-Camargo, K., González, J., Chen Austin, M., Carpino, C., Mora, D. and Arcuri, N., 2024. Advances in Retrofitting Strategies for Energy Efficiency in Tropical Climates: A Systematic Review and Analysis. Buildings, 14(6), p. 1633.

[9] Tan, X.Y., Mahyuddin, N., Kamaruzzaman, S.N., Mat Wajid, N. and Zainal Abidin, A.M., 2024. Investigation into energy performance of a multi-building complex in a hot and humid climate: efficacy of energy saving measures. Open House International, 49(3), pp. 489-513.

[10] Mohsenzadeh, M., Marzbali, M.H., Tilaki, M.J.M. and Abdullah, A., 2021. Building form and energy efficiency in tropical climates: A case study of Penang, Malaysia. urbe. Revista Brasileira de Gestão Urbana, 13, p. e20200280.

[11] Shari, Z., Mohamad, N.L. and Dahlan, N.D., 2023. Building Envelope Retrofit For Energy Savings in Malaysian Government High-Rise Offices: A Calibrated Energy Simulation. Jurnal Teknologi, 85(4), pp. 1-15.

[12] Lakhiar, M.T., Sanmargaraja, S., Olanrewaju, A., Lim, C.H., Ponniah, V. and Mathalamuthu, A.D., 2024. Energy retrofitting strategies for existing buildings in Malaysia: A systematic review and bibliometric analysis. Environmental Science and Pollution Research, 31(9), pp. 12780-12814.
[13] Thirunavukkarasu, M. and Sawle, Y., 2021. A comparative study of the optimal sizing and management of off-grid solar/wind/diesel and battery energy systems for remote areas. Frontiers in Energy Research, 9, p. 752043.
[14] Mukhtaruddin, R.N.S.R., Rahman, H.A. and Hassan, M.Y., 2013, June. Economic analysis of grid-connected hybrid photovoltaic-wind system in Malaysia. In 2013 International Conference on Clean Electrical Power (ICCEP), pp. 577-583.
[15] Islam, S.Z., Othman, M.L., Saufi, M., Omar, R., Toudeshki, A. and Islam, S.Z., 2020. Photovoltaic modules evaluation and dry-season energy yield prediction model for NEM in Malaysia. Plos one, 15(11), p. e0241927.
[16] Climate and monthly weather forecast, Malaysia, Weather Atlas, 2024. https://www.weather-atlas.com/en/malaysia-climate [Accessed 07 August 2024]
[17] The Climate of Malaysia, Blue Green Atlas, 2024. https://www.bluegreenatlas.com/climate/malaysia_climate.html [Accessed 07 August 2024]
[18] Shahidan, M.F., 2023. Landscape Energy Efficient Design: The Understanding, Faculty of Design and Architecture. Universiti Putra Malaysia. https://frsb.upm.edu.my/article/landscape_energy_eeficient_design_the_understanding-71909 [Accessed 07 August 2024]
[19] Malaysia – Country Summary | Climate Change Knowledge Portal, 2024. https://climateknowledgeportal.worldbank.org/country/malaysia [Accessed 07 August 2024]
[20] Cover Story: Charting a net-zero pathway for Malaysia, The Edge Malaysia, 2024. https://theedgemalaysia.com/article/cover-story-charting-netzero-pathway-malaysia [Accessed 07 August 2024]
[21] Malaysia's Green Building Index Paves the Way to Sustainable Development, Energy Watch, 2024. https://www.energywatch.com.my/blog/2022/08/22/malaysias-green-building-index-paves-the-way-to-sustainable-development [Accessed 07 August 2024]
[22] Malik, M.Z.A., 2023. Malaysia's Green Plan: Paving the Way for a Sustainable Future, South East Asia. https://south-east-asia.bureauveritas.com/magazine/malaysias-green-plan-paving-way-sustainable-future [Accessed 07 August 2024]
[23] Momade, M.H. and Hainin, M.R., 2018. Review of sustainable construction practices in Malaysian construction industry. Int J Eng Technol, 7(4), pp. 5018-5021.