Using Practical Swarm Optimization Algorithm For Dna Profile Matching

Nawal S. Jabir[1], Zainab A. Kahlaf[2*]
[1]University of Basrah, College of Education Pure Sciences, Iraq
[2]University of Basrah, College of Sciences, Iraq

Abstract:
In forensic science, Deoxyribonucleic Acid (DNA) is essential for identifying individuals and exonerating innocents, and convicting guilty persons. Processing and analysing biological evidence transmitted to a crime scene and identifying the genetic material in DNA enables the identification of criminals, paternity, and unknown bodies. DNA technology plays an important role in individual recognition. However, the traditional methodology for personnel identification using DNA in diagnosis requires medical experts' intervention. This is not a simple task due to the large searching dimension in the utilized database. This paper automates DNA profiling identification using the Particle Swarm Optimization algorithm by locating the best matching data for a query on an unknown person (e.g., corpse, determination of proportions, identification of the suspect, etc). The data used in this research was collected from the Iraqi interior ministry at the Basra Investigation Centre. The proposed system has demonstrated significant improvement in the obtained results by matching DNA sequences for an unknown person with the collected database with 100% accuracy.
Keywords: Particle swarm optimization, DNA profiling, Bioinformatics, identification, DNA forensic.
DOI: 10.24297/j.cims.2023.1.6

1. Introduction

Even though biologists have gathered a massive size of DNA data, however, the specifics of those sequence data is mostly still not understood [1]. Biologists can now convert huge amounts of biological data into usable data due to improved techniques developed in recent years. Genomics functional is the creation and implementation of global experimental techniques to estimate gene function using structural genomics' knowledge and reagents. High-throughput or large-scale experimental approaches distinguish it, as well as statistical and computer analysis of the outcomes [2]. It is used to track the expression of a large number of genes simultaneously. The process of turning a gene's DNA sequence into RNA, which acts as a template for manufacturing proteins, is known as gene expression. Gene expression level describes how active a gene is in a particular tissue, at a certain time or under specific experimental

conditions[3] . This technology entails several steps. The complementary DNA (cDNA) molecules, also known as oligos, are initially printed on slides as dots. Following that, sample and control dye-labeled samples are hybridised. The term comparable here refers to the fact that any difference in a gene's measured expression value between two trials should reflect the genuine expression levels of that gene [4]. Most gene expression assessments are carried out manually using scant experimental data. Currently, there is a great need for automatic analysis of the overall link underlying a large number of genes from their expression.

The study of algorithms that can learn from experience and forecast is known as optimization algorithms. The optimization algorithm's theoretical parts are founded in informatics and statistics, but computational concerns are also necessary. Optimization algorithms could play a crucial part in the analysis process due to the complicated nature of biological data [5][6]. Due to distinctive DNA sequences that can be used as biomarkers for profile identification, forensic scientists all over the world have used DNA profiling-based techniques. Any DNA-based techniques that distinguish the DNA from a specific person or group of people within a community of organisms are referred to as DNA fingerprinting or profiling. Nevertheless, DNA profiling is considered one of the hardest problems in the forensic science domain, and this considers an active area of research. Many papers addressed this area to develop and improve DNA profiling matching.

The contributions of the current study are to identify typical gene expression analysis problems by using an optimisation technique that is used to find the best common gene expression for known (e.g. criminal or suspect) or unknown person (e.g. Victims of war or terrorism) through the match the gene expression within the database. The designed system can be used to assist investigators in revealing the identity of accused or deceased persons and other cases.

This paper is organised into three main sections. Related works will be detailed in section 2. Next, Particle swarm optimisation will be discussed in section 3. The proposed system will be shown in section 4. In section 5, experimental results will be presented. Finally, conclusions and future works will be highlighted in section 6.

2. Related Work

The employment of intelligent computing models has resulted in numerous advancements in the field of DNA categorisation. Some of these works are briefly described below:

A technique based on frequency patterns and entropy to build representative vectors of DNA sequences for efficiently computing the similarity between DNA sequences is presented in [7]. The proposed method is tested through experiments and contrasted with two recently proposed alignment-free methods and the BLASTN tool. Moreover, the authors claimed that the proposed method outperforms the two alignment-free methods and the BLASTN tool when tested on the b-globin genes of 11 species and using the MEGA results as the baseline. To solve the DNA sequence assembly problem, [8] proposed a solution for the DNA sequence assembly problem using Particle Swarm Optimization (PSO) with Naïve Crossover and Shortest Position Value (SPV) rule. The results show the high robustness of the proposed model in finding DNA sequences.

A technique for autonomously learning high-performance deep networks called DNA computing-inspired networks design is proposed in [9], where authors claimed that the recommended model could categorize comet photos into four categories with an overall accuracy rate of 96.1 percent (DNAND). They also describe the termination approach, in which "poor" models are stopped from being trained if they fail to meet a certain accuracy level on the validation set, reducing computational costs and speeding up the learning process. As for
A study on similarity kernels based on different similarity schemes, proposes a hybrid one. In this study[10], the researchers combine different similarity schemes; each scheme is deduced based on alignment. The authors demonstrate that combining different similarity schemes does, in fact, generalize well in machine learning. The scoring scheme also turned out to have an impact on generalization. Neural network, have used for DNA damage quantifying where a Convolutional Neural Network (CNN) was compared to other approaches in the literature for using comet assay images in which they tagged 796 single comet grayscale photos with resolution and divided them into four classes with approximately 200 samples each: G0 (healthy), G1 (poorly defective), G2 (defective), and G3 (extremely defective). The result shows that CNN has overperformed the other methods[11]. A strategy for dealing with high-dimensional low-sample-size (HDLSS) DNA methylation datasets based on the usage of AEs and survival analysis is presented [12]. This approach was utilized to obtain useful information about breast cancer recurrence based on key genes. The discovery of multiple enriched words and

related linkages between the genes using functional annotation enrichment analysis was another outcome that validated the methodology.

3. The Proposed Model

The adopted identification system consists of four main phases, as shown in Fig. 1. Data collection, pre-processing data phase, identification phase, and evaluation phase. The pre-processing data phase is an important stage in preparing the DNA data for the next phase. Then, the DNA sequence will be passed to the matching algorithm (PSO) to find the best matching results. Finally, the evaluation is applied to compute the system performance.

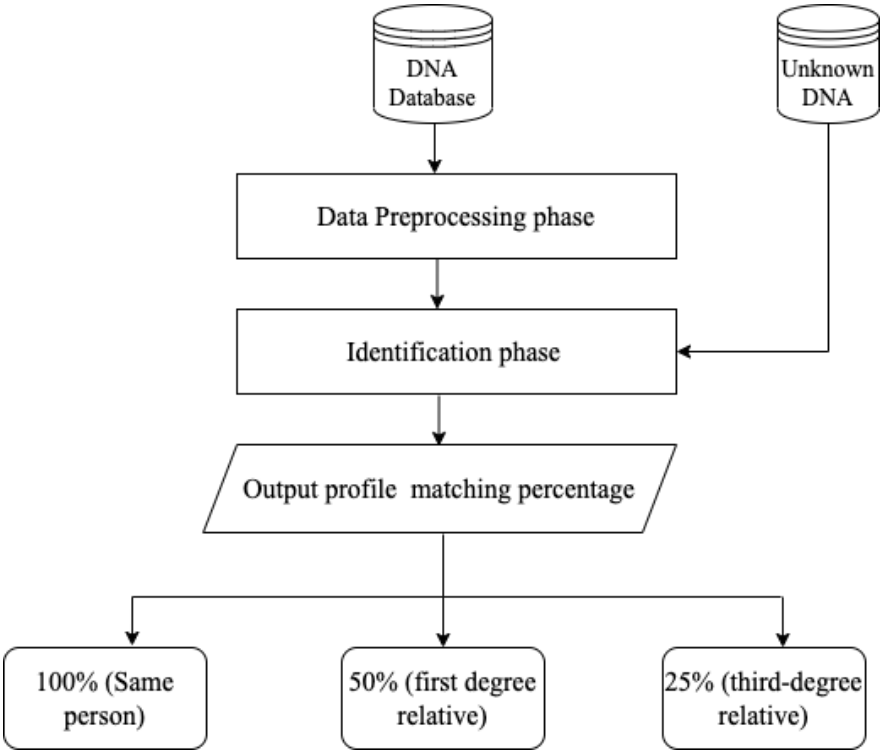

Fig.1: The proposed model phases

Data Collection

The DNA data used in this study was gathered from the Basra Investigation Center belonging to the Iraqi Ministry of Interior. This data contains 1000 DNA profiles for victims of mass graves, unknown bodies, civilians and law enforcement victims who fell due to a terrorist bombing, suspect individuals, and paternity tests that comprise Y-chromosomal STRs and mtDNA sequencing.

Data Pre-processing Phase

Data quality is essential for accurate results. Thus, pre-processing considers the first step in any designed system that helps to obtain high-quality data for data analysis. In this study, the removal of DNA sequences (or some regions of the DNA sequence) containing missing data from the DNA profile was used to ensure reliable results.

Identification Phase

In this study, the PSO algorithm is used to find the best alignment between two DNA sequences and identify the family relationship between them. The PSO algorithm is behaved as animal social behavior, including that of insects, herds, birds, and fish. It is a stochastic optimisation technique based on the swarm. These swarms adhere to a cooperative approach to food acquisition, and each swarm member continuously modifies the search pattern in response to its own and other members' learning experiences. The PSO algorithm's basic concept is strongly connected to two categories of research: One employs an evolutionary algorithm, while PSO also uses an evolutionary algorithm, but it searches a much larger portion of the solution space for the optimised objective function at once in a swarm mode. Particles in PSO can modify their movement mode to react to environmental changes while maintaining steady travel in the search space. Particle swarm systems, therefore, adhere to the aforementioned five criteria [12][13][14].

Eberhart and Kennedy, two American academics, proposed the PSO algorithm in 1995. During operation, the PSO algorithm tracks individual optimal particle $pbest_k^i$ and group optimal particle $gbest_k^y$, updating the particle velocity and position using the following formulas:

$$v_{k+1}^i = wv_k^i + c_1 \times r_1 \times \left(pbest_k^i - x_k^i\right) + c_2 \times r_2 \times \left(gbest_k^y - x_k^i\right) \quad (1)$$

$$x_{k+1}^i = x_k^i + v_{k+1}^i \quad (2)$$

Where k =1, 2, …, M is the dimension of search space and population size; r1, r2 are random numbers in the range of (0,1), The learning factors c1 and c2 (usually between (0,2)) indicate that particles can, respectively, learn from one another and themselves .v_k^i and x_k^i represent the particle's current location and velocity respectively, where $v \in [v_{min} v_{max}]$, and w is a constant inertia weight used to control the variety of particles; i is the algebra of the current population. The positions of the current individual and group optimal particles, $pbest_k^i$ and $gbest_k^y$respectively, are shown in Fig.2, which also shows the process of updating the particle's velocity and location[15].

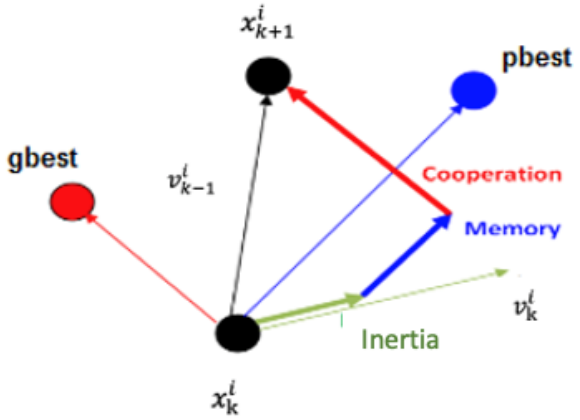

Fig.2: The updating velocity and position for PSO

3.4 Evaluation phase

The Match confidence scores between two DNA profiles depend on the inheritance of common traits from a common ancestor, and this score decreases with each successive generation. In this paper, a matching score is used as a fitness function for the PSO algorithm to evaluate the proposed system.

Let's suppose (*Newp, DBp*) is database of unknown persons and predefined DNA sequences, respectively. The length of *Newp*, and *DBp* act as the length of DNA sequence. Also, let N be the length of DNA sequence in Newp, and DBp that equals to 16 where $Newp = \{s_1 s_2 s_3 \_\_\_\_ s_n\}$, and $DBp = \{p_1 p_2 p_3 \_\_\_\_ p_n\}$ as shown in Fig.3

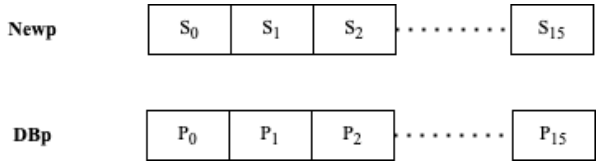

Fig.3 : New$_p$ and DBP DNA sequences during matching process

Suppose seq is the length of specific segments of DNA that *New*, and *DB* share as relatives. The red box in Fig.4 defines the matching between two DNA sub profiles i.e. the Query DNA and the Target DNA. From Fig.3, it can be seen that, out of 6 numbers, 4 are matched.

So, alignment $L_{s,p}$ between New_s, and DB_p can be represented by $V_{s,p} = (New_s^{query}, DB_p^{target})$ as follow:

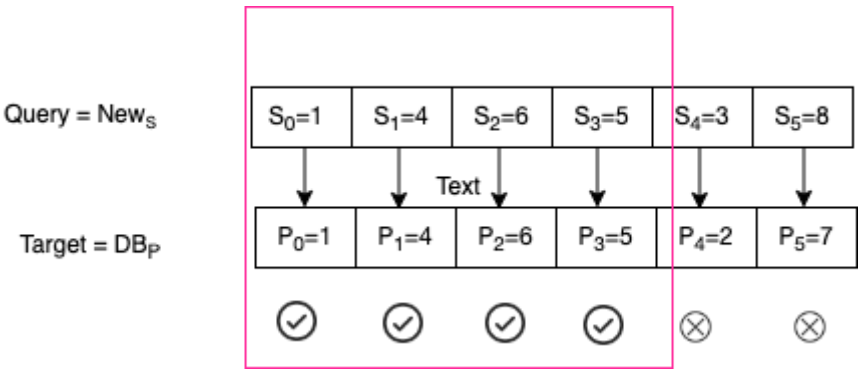

Fig.4: Matching and mismatching indications for sequence = 6

To compute matching scores between Newp and DBp, Alignment accuracy can be quantified by the mismatching score. To compare two sequences and measure their similarity, a mismatching score measures the correspondence between sequences by checking if the two patterns are aligned in the same columns. The higher mismatching score indicates less similarity between profiles, meaning the profiles are considered misaligned.

The $Mismatching_{score}$ is computed using the following formula:

$$Mismatching_{score} = Number\ of\ mismatching\ locations / N \quad (3)$$

Number of mismatching locations = Substitutions + Insertions + Deletions

A substitution occurs when one sequence location gets replaced.

An insertion is when a extra sequence is added.

A deletion happens when a location is left out of the sequence

Suppose $1 \le i, j \le N$ and New_i , $DB_j \in \Sigma$ a column $(New_i\ DB_j)$ of an alignment L is called a match if $New_i = DB_j$ and mismatch (or substitution) if $New_i \neq DB_j$.

$$Matching_{score} = (1 - Mismatching_{score}) * 100 \quad (4)$$

Depends on the $Mismatching_{score}$, the proposed system print the final decision as following :

$$\text{Matching}_{\text{score}}(\text{New}, \text{DB}_i) \begin{cases} 100\% \text{ if } \forall j, \text{New}_j = \text{DB}_j, \text{seq}=16 \\ 50\% \text{ if } \forall j, \text{New}_j/2 = \text{DB}_j/2, \text{seq}=8 \\ 25\% \text{ if } \forall j, \text{New}_j/4 = \text{DB}_j/4, \text{seq}=4 \\ 0\% \text{ if } \forall j \text{ New} \neq \text{DB}_i, \text{seq}=0 \end{cases} \quad (5)$$

Where 100% means $New = DB_i$ and the DNA sequence belongs to the same person, while 50% means the first degree relative (parents and siblings). Whereas 25% denotes second-degree relative (grandparents, half-siblings, aunts/uncles, or double first cousins), and 0% means there is no matching at all.

In eq. (5) denotes the fitness value for unknown DNA profiling New_i of PSO. DB_i who has the best matching value, is considered the optimum solution. The fitness function is the best matching score calculated for each unknown DNA profiling. Fig. 5 summarize the possible outcome

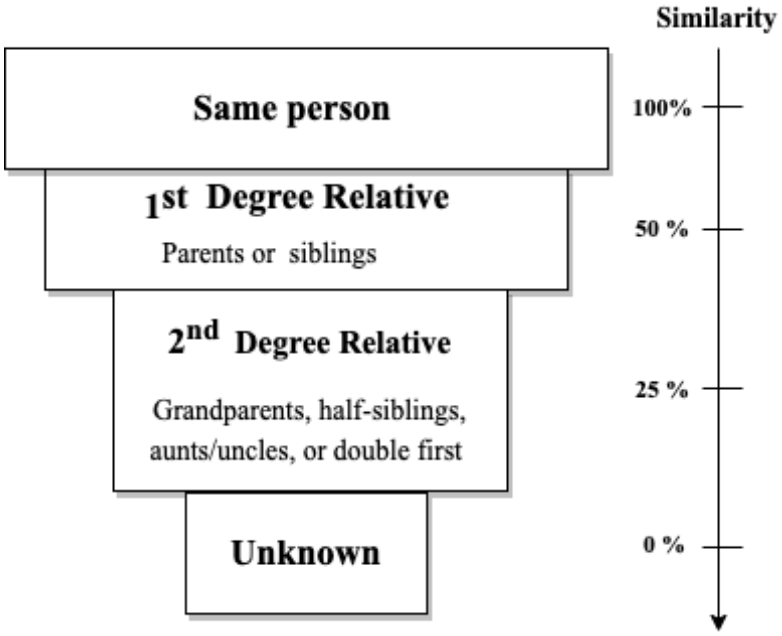

Fig.5: Similarity percentage and the degree of kinship

4. Results and Discussion

In this section, the proposed system will be validated the effectiveness to identify the persons' profile of an unknown DNA sequence and also needs to define the parameters that used in proposed system phases as discussed in the following subsections:

4.1 Data Collection Phase

In order to assess the allele frequencies and gene diversity of each Y-STR locus of the Y filer TM PCR amplification kit, we examined seventeen Y-chromosomal STR loci. These seventeen loci include DYS19, DYS391, DYS438, DYS390, DYS439, DYS392, DYS393, DYS385a, and DYS385b.

They are also known as DYS635, DYS437, DYS448, DYS456, DYS458, GATA Y H4, DYS389I, and DYS389II. But in our project, 16 genetic loci were used only according to the type of kit used, as we have different types of kits, some of which give us 16 and some of them give us 24 genetic loci. Each genetic locus consists of two alleles, the first half of which comes from the father and the other half from the mother, and through these Genetic loci We can distinguish and reveal the identity of the unknown or suspected person through the matching of these allels with each other[16].

4.2 Pre-processing Phase

In this study, the removal of DNA sequences containing missing data from the DNA profile was used to ensure reliable results. The total collected data was roughly 1500 profile, and the number of eliminated profiles were 325. The remaining was 1175 distributed into 1000 profiles defined as a searchable database. While 250 DNA sequences for unknown persons were used for testing as summarized in Table 1 :

Table 1: DNA database distribution based on the similarity percentage

Degree of Relative	Number of profiles
Same Person	75
First-degree relative	75
Second-degree relative	50
Not Found	50
Total	250

4.3 Identification Phase

To optimize the parameters and to find the best matching profile for a given DNA sequence, the match probabilities and mismatch probabilities must be calculated based on the fitness function using the PSO algorithm. Several experiments have been performed to gain the best configuration for the PSO algorithm as shown in Table 2 :

Table 2 : PSO algorithm parameters

Parameter	Value
r1	0.5
r2	0.5

c1	1
c2	1
w	0.7
N	16

Fitness value or matching score value using to retrieve the best matching person's profile. The matching score for unknown DNA is based on the amount of shared location of the DNA with that

Table 2: DNA matching results stored database. It is almost clear that a person's DNA sequence belongs to this family since it was passed down from a common ancestor, according to the proportion of matching scores that show the degrees of family links. The number of births that divide relatives from one another determines the degree of kinship for direct blood relations. So, four cases can be resulted from the proposed system, as shown in the Table 3 below:

Table 3: Proposed system possible results

Degree of Relative	Matching Score
Same Person	100%
First-degree relative	50%
Second-degree relative	25%
Not Found	0

4.4 Discussion

In this study, the unknown DNA sequence was used as the query and search about in stored database for known profiling persons using generated pairwise alignments. Using a dynamic programming algorithm, the suggested system's accuracy was compared to that of pairwise sequence alignment (biopython package). There are two methods in this package: global and local, that conduct alignments between two sequences. The best concordance between all characters in two sequences is discovered via a global alignment. A local alignment identifies the exact subsequence that best aligns. Global alignment was utilized, and it produced good results for 100% matching for the same individual, but it was insufficiently reliable to produce good

results for kinship degree since the degree of kinship decreases with the distance between the parties. For instance, the majority of autosomal(non-sex chromosome gene) DNA ancestry tests indicate a 45–50% likelihood of correctly identifying a match with a fourth cousin, but an accuracy rate of 90–98% for discovering a match with a first kinship. The suggested approach, however, provided 100% accuracy up to the third degree of familial relationships.

5. Conclusion

DNA plays an important role in forensic science by exonerating the innocent and convicting the guilty and also for identification. By processing and analysing biological evidence transmitted to the crime scene and identification, the genetic material in DNA enables the identification of the criminal. These organic elements can be found in many different forms, including bodily fluids and human remains. Therefore, DNA technology is crucial for identifying people. In this work, the PSO algorithm is used for DNA matching for issues related to law enforcement and other issues related to DNA tests. The used database is real data collected from the Iraqi interior ministry.The results using PSO algorithm show 100% accuracy. Furthermore, besides the simulation results, the proposed work has been tested in a real-world case scenario in the Iraqi ministry of interior at the Basra Investigation Center, in which the robustness and accuracy have been confirmed. Finally, the proposed method can replace the manual method that is currently used there (Basra Investigation Center).In the foreseen future, our method will be occupied with a user-friendly interface to be handled by law enforcement that needs such technology.

References

1. P. Scherz, "The Displacement of Human Judgment in Science: The Problems of Biomedical Research in an Age of Big Data," Social Research: An International Quarterly Johns Hopkins University Press, vol. 86, no. 4, pp. 957–976, 2019, doi: 10.1353/sor.2019.0048
2. N. Nagarajan, E. K. Y. Yapp, N. Q. K. Le, B. Kamaraj, A. M. Al-Subaie, and H.-Y. Yeh, "Application of Computational Biology and Artificial Intelligence Technologies in Cancer Precision Drug Discovery.," BioMed research international, vol. 2019, p. 8427042, 2019, doi: 10.1155/2019/8427042.
3. P. J. Thul and C. Lindskog, "The human protein atlas: A spatial map of the human proteome.," Protein science : a publication of the Protein Society, vol. 27, no. 1, pp. 233–244, Jan. 2018, doi: 10.1002/pro.3307.

4. N. Islam et al., "RNA Biomarkers: Diagnostic and Prognostic Potentials and Recent Developments of Electrochemical Biosensors," Wiley Online Library, vol. 1, no. 7, pp. 122–131, 2017.
5. N. Peng, H. Poon, C. Quirk, K. Toutanova, and W. Yih, "Cross-Sentence N-ary Relation Extraction with Graph LSTMs," Transactions of the Association for Computational Linguistics, vol. 5, pp. 101–115, Apr. 2017, doi: 10.1162/tacl_a_00049.
6. Ö. A. Aslan and R. Samet, "A Comprehensive Review on Malware Detection Approaches," IEEE Access, vol. 8, pp. 6249–6271, 2020, doi: 10.1109/ACCESS.2019.2963724.
7. S. V. Ravi, "DNA Sequence Assembly using Particle Swarm Optimization," International Journal of Computer Applications, vol. 28, no. 10, pp. 33–38, 2011, doi: 10.5120/3425-4777.
8. Q. Wang, F. Wang, and X. Zhou, "High-efficiency and integrable DNA arithmetic and logic system based on strand displacement synthesis," Nature Communications, vol. 10, no. 1, p. 5390, 2019, doi: 10.1038/s41467-019-13310-2.
9. M. Awad and L. Khan, "Hybrid DNA Sequence Similarity Scheme for Training Support Vector Machines *," University of Texas at Dallas, pp. 1–5, 2009.
10. S. Ahmed et al., "ACP-MHCNN: an accurate multi-headed deep-convolutional neural network to predict anticancer peptides," Scientific Reports, vol. 11, no. 1, p. 23676, 2021, doi: 10.1038/s41598-021-02703-3.
11. R. Chanpa, M. Ali, J. Jamali, A. Hatamlou, and B. Anari, "An optimized swarm intelligence algorithm based on the mass defence of bees," International Journal of Nonlinear Analysis and Applications, vol. 13, no. October 2021, pp. 3451–3462, 2022.
12. Mahdi AjdaniHamidreza Ghaffary, "Introduced a new method for enhancement of intrusion detection with random forest and PSO algorithm," Wiley -Security and Privacy, vol. 4, no. 1, 2021.
13. R. Kadry and O. Ismael, "A New Hybrid KNN Classification Approach based on Particle Swarm Optimization," International Journal of Advanced Computer Science and Applications, vol. 11, no. 11, pp. 291–296, 2020, doi: 10.14569/IJACSA.2020.0111137.
14. V. Gupta, S. Sachdeva, and N. Dohare, "Cosine Similarity an Overivew," Trends in Deep Learning Methodologies: Algorithms, Applications, and Systems, pp. 183–206, Jan. 2020, doi: 10.1016/B978-0-12-822226-3.00008-8.

15. A. Al-Mayyahi, W. Wang, and P. Birch, "Path tracking of autonomous ground vehicle based on fractional order PID controller optimized by PSO," in 2015 IEEE 13th International Symposium on Applied Machine Intelligence and Informatics (SAMI), 2015, pp. 109–114. doi: 10.1109/SAMI.2015.7061857.
16. B. M. Auhied, "Short Tandem Repeats (STRs) Alleles Frequencies for Basrah Population," Master Thesis, University of Basrah, Basrah, Iraq, 2013.