Hindawi
Journal of Electrical and Computer Engineering
Volume 2023, Article ID 6678837, 10 pages
https://doi.org/10.1155/2023/6678837

*Research Article*

# An Artificial Intelligence Approach for Verifying Persons by Employing the Deoxyribonucleic Acid (DNA) Nucleotides

**Raid Rafi Omar Al-Nima, Marwa Mawfaq Mohamedsheet Al-Hatab, and Maysaloon Abed Qasim**

*Technical Engineering College of Mosul, Northern Technical University (NTU), Mosul, Iraq*

Correspondence should be addressed to Raid Rafi Omar Al-Nima; raidrafi5@gmail.com

Received 27 June 2023; Revised 12 October 2023; Accepted 26 October 2023; Published 17 November 2023

Academic Editor: Jit S. Mandeep

Copyright © 2023 Raid Rafi Omar Al-Nima et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Deoxyribonucleic acid (DNA) can be considered as one of the most useful biometrics. It has effectively been used for recognizing persons. However, it seems that there is still a need to propose a new approach for verifying humans, especially after the recent big wars, where too many people lost and die. This approach should have the capability to provide high personal verification performance. In this paper, a personal recognition approach based on artificial intelligence is proposed. This approach is called the artificial DNA algorithm for recognition (ADAR). It utilizes a unique identity for each person acquired from DNA nucleotides, and it can verify individuals efficiently with high performance. The ADAR has been designed and applied to multiple datasets, namely, the DNA classification (DC), sample DNA sequence (SDS), human DNA sequences (HDS), and DNA sequences (DS). For all datasets, a low value of 0% is achieved for each of the false acceptance rate (FAR) and false rejection rate (FRR).

## 1. Introduction

With advanced science and technology, it is now possible to authenticate people in order to achieve high levels of security. Maintaining private data and meeting the increased demands for security have become important matters. There are several methods that use biometrics to approve the identity such as fingerprint [1], palm print [2], iris print [3], and voice print [4]. Biometrics include measuring an individual's distinctive physical or behavioral biometric trait [5]. The Greek word "bio" means life and "metric" means measuring; both words are combined to form the phrase "biometric" [6]. In fact, there are different terminologies that are associated with the word "biometrics" such as verification, identification, classification, authentication, and recognition. It seems hard to distinguish between each one of them. However, such terminologies are clarified over years of working. Verification utilizes the one-to-one policy, where a user declares his/her identity in order to compare with specific related information belonging to the same user. Then, a decision about accepting or rejecting the personal identity claim is provided [7]. Identification exploits the one-to-many policy. Here, it is necessary to apply matching between the provided information by a user and all the stored information of all users. So, there is no need to provide a user's identity, and the decision can either assign or refuse to declare the identity [8]. Classification refers to categorizing information into a certain group or set [9]. Authentication refers to the process of proving an actual action. In computer science, this term is typically associated with approving a user's identity [10]. Recognition is a general terminology, and it can be used to mention any of the previous biometric styles (verification, identification, classification, or authentication).

Deoxyribonucleic acid (DNA) can offer trustworthy personal verification. It is inherently digital and remains unchanged during the person's lifetime and even after death [11]. The form of DNA known as a double helix; it is comprised of two connected strands that twist around one another to resemble a spiral ladder. Deoxyribose and phosphate are the main components of the backbone of each strand. Each sugar molecule in the DNA has one of four

bases (or nucleotides): adenine (A), cytosine (C), guanine (G), or thymine (T) [12]. The A, T, C, and G refer to the chemical elements that connect the two strands together. Figure 1 demonstrates a sample of the DNA with the chemical components.

DNA bases pair up with each other, A with T and C with G, to form units called base pairs. Each base is also attached to a sugar molecule and a phosphate molecule. The sequence of these pairs differs from one person to another, making the DNA unique for each individual and therefore this can be used for personal verification or any other recognition style.

It is known that using the DNA is so valuable for personal verification. However, a more effective DNA system is still required. This has been exposed in Iraq because of the big issues and wars, where too many humans died and were lost. Such DNA verification system should have the ability to deal with huge number of samples and provide precise outcomes. This work presents a new system based on the artificial intelligence by employing a unique DNA pattern of the nucleotides (A, T, C, and G) for verifying persons. The proposed approach here is called the artificial DNA algorithm for recognition (ADAR). It can provide high performance, it facilitates searching for DNA verification samples, and its efficiency is proven with four utilized datasets.

The next sections are architectured as follows: Section 2 presents the literature review. Section 3 describes the ADAR theory. Section 4 discusses the experimental work and Section 5 provides the conclusion.

## 2. Literature Review

There are many prior DNA studies that can be highlighted. In 2005, Mitra presented a survey about the roles of different soft computing techniques such as fuzzy sets, artificial neural networks (ANNs), evolutionary computation (EC), and support vector machines (SVMs) to classify and recognize the major pattern for DNA genomic sequence and protein architecture. The SVM classifier recorded the highest accuracy and least error compared to other applied methods [14]. In 2009, Wei proposed a system for categorizing the DNA sequence of four types of bacteria. It consists of the following steps: extracting DNA sequence features, constructing the ANN model, and classifying data. The accuracies of classifying the four types of bacteria for lengthy and repetitive DNA sequences in the utilized dataset was 92.9%, 90.2%, 80.4%, and 41.7% after learning the ANN model [15]. In 2012, Khashei et al. presented a novel hybrid model integrating AI and fuzzy logic for the analysis of gene data. Comparative evaluations against conventional approaches such as artificial neural networks (ANN), linear discriminant analysis (LDA), quadratic discriminant analysis (QDA), *K*-nearest neighbor (KNN), and support vector machines (SVM) demonstrated that the proposed model achieved enhanced classification accuracy. This suggests that the suggested hybrid model holds promise as a viable alternative technique, particularly in scenarios where data scarcity is a concern [16]. In 2017, Pashaei et al. concentrated on the human genome by considering splice site identification with random forest. The effectiveness of the employed classifiers

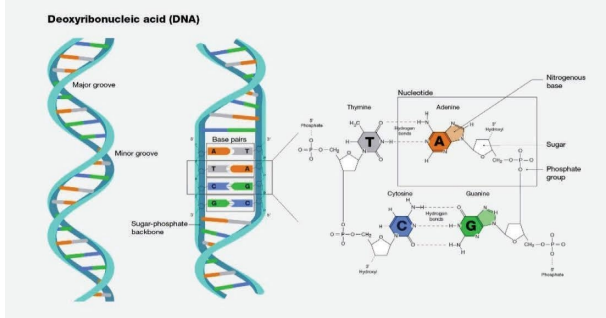



Figure 1: Part of a DNA with its chemical components [13].

was mostly influenced by feature extraction and feature selection techniques used in DNA encoding. The feature selection methods removed the extraneous information, whereas the feature extraction methods attempted to extract as much information from the DNA sequences as possible. The applied random forest was examined as a means of feature selection and classification in the splice site domain [17]. In 2018, Pashaei and Aydin worked on Markovian encoding models. Recognition of splice sites for persons was considered. A third order Markov model with SVM (MM3-SVM) was proposed. It outperformed the best-known state-of-the-art methods [18]. In the same year, Kaniwa and Phuthego explained how genetics is affected by next-generation sequencing to rapidly generate the DNA, and ribonucleic acid (RNA) sequences. This is for swiftly constructing the DNA and RNA sequences. Madrid, Spain, was the site of this study. It was based on the fundamental notion that DNA sequence information was expanded, which made simple and affordable analysis possible [19]. In 2020, Sun et al. a novel multilayer deep neural network (DNN) was devised and implemented for survival prediction in a genome-wide association study. This DNN survival model exhibited superior predictive accuracy compared to several existing models, while also successfully identifying clinically significant risk subgroups. The model employed an effective approach for capturing complex architectures among genetic variants. The evaluation of the model was conducted on genome-wide association studies (GWAS) data from two large-scale randomized clinical trials involving over 7800 participants with age-related macular degeneration (AMD) [20]. In 2021, Alatrany et al. proposed a hybrid machine learning (ML) technique for the prediction of Alzheimer's disease using genome sequence. The most important single-nucleotide polymorphisms linked to Alzheimer's disease were chosen. Using data from a random forest, a DL model for the illness prediction was then provided. Utilizing a convolutional neural network (CNN) and multilayer perceptron (MLP), the simulation results showed that the hybrid model was effective in predicting people who had Alzheimer's disease [21]. In 2022, Manhal investigated the use of DNA to identify individuals. An efficient algorithm was used to find the distinctive DNA patterns. The unique personal DNA pattern (UPDP) was approached for personal identification. Four databases were employed, and they all yielded low reported errors [22]. In the same year, Rukhsar

et al. introduced DL analysis of RNA sequence gene expression data for cancer classification. Five different kinds of cancer data from the Mendeley archive were examined. The appropriate characteristics were retrieved and chosen using the DL. Eight DL algorithms were employed to accomplish classification in the final phase. The evaluation of DL classifiers was performed using k-fold cross-validation and four different data splitting techniques. Among the evaluated classifiers, the CNN exhibited the highest overall performance [23]. Also in the same year, Hamed et al. provided a review on enhancing algorithms for pattern matching. This survey concentrated on biological sequences. It presented analyses of techniques, efficiency, and complexity. Furthermore, it offered comparisons between various algorithms for matching [24]. In 2023, Ibrahim et al. proposed a novel fast technique. It was for pattern matching. It is determined by biological sequences. This work was constructed to increase speed up the search for DNA sequence pattern [25]. In the same year, Hamed et al. investigated the efficiency of optimizing classification. It considered machine learning. It focused on pattern matching. This study suggested a new DNA sequence classification model. It fused between a pattern-matching procedure and machine learning techniques [26].

This paper adds a significant contribution to previous work by approaching an artificial intelligence algorithm named the ADAR. This algorithm is employed for verifying persons according to their DNA sequences of nucleotides.

## 3. Proposed Approach

The proposed approach is called the ADAR. Its construction starts with substantial numbers of DNA sequences. Each DNA sequence has a unique nucleotide pattern code. Each strand of DNA is viewed as a fundamental sequence of nucleotides (or bases). Figure 2 depicts a DNA sequencing sample of two strands with nucleotide arrangements.

Any sequencing arrangement in a single DNA strand consists of A, G, T, and C nucleotides. In this work, determining the identity of a person is considered after counting the number of repeated patterns of four nucleotides (quaternary nucleotides).

The ADAR algorithm considers counting all numbers of repeated four-nucleotide patterns. Then, the maximum repeated pattern is determined. An identification claim is applied to a specific person. Therefore, comparisons for (pattern index, maximum repetition and identity claim) are employed in the case of verification.

The full system of the ADAR works in two main phases: enrolment and verification. In the enrolment phase, DNA samples are received and processed for storage in the system. In the verification phase, an identity claim and a DNA sample are provided for testing. A flowchart for the proposed ADAR with the two phases is given in Figure 3.

For the enrolment phase, the system of the ADAR consists of the following layers: input layer, search layer, max layer, identity layer, and comparison layer, which will be used for comparison. The verification phase of the ADAR system involves the same stages as the enrolment phase and the output layer, which is added at the end and provides the verification decision. The proposed ADAR layers for the two phases of enrolment and verification are demonstrated in Figure 4. They can be illustrated as follows:

Input Layer: It is required for receiving DNA sample **D** as a string of sequences of nucleotides (or bases).

Search Layer: It is employed for counting the numbers of repeated quaternary patterns **X** of nucleotides (frequencies of quaternary patterns of nucleotides). It considers all possible probabilities $P(\mathbf{X})$, starting from "AAAA" and ending with "CCCC" (this covers 256 probabilities).

Max Layer: This layer collects the maximum frequencies of the most repeated quaternary patterns of nucleotides for all **D** samples. The following equation expresses a maximum operation:

$$Y = \max(\mathbf{X}_i), \tag{1}$$

where $Y$ is the maximum frequency of the most repeated quaternary pattern of nucleotides, max is the maximum operation between all frequencies of $\mathbf{X}_i$ patterns, and $i = 1, 2, \ldots, 256$ possibilities.

Identity Layer: This layer during the enrolment phase stores the identity of $n$ persons who provide their DNA sequences. Whereas, this layer during the verification phase matches between the identity claim for a person who requires to be verified and his/her stored information.

Comparison Layer: It assigns three factors for each DNA sequence provided by any person in order to be used for verification comparisons. These factors are [pattern index ($i$), maximum repetition ($Y$), and identity claim]. This layer is crucial for ensuring reliable and accurate verification of individuals.

Output Layer: It provides the output verification decision according to all processing layers and identity claim.

The ADAR verification algorithm can be illustrated as follows:

Step 1: Receiving the DNA sample as a string of sequence of nucleotides.

Step 2: Counting the numbers of repeated quaternary patterns of nucleotides.

Step 3: Collecting the maximum frequencies of the most repeated quaternary patterns of nucleotides for all the DNA sample.

Step 4: Matching between the identity claim for a person who requires verification and his/her stored information.

Step 5: Comparing with the three factors of (pattern index ($i$), maximum repetition ($Y$) and identity claim).

Step 6: Providing the output verification decision according to all processing layers and identity claim.

Parameters used for the ADAR analysis are given in Table 1.

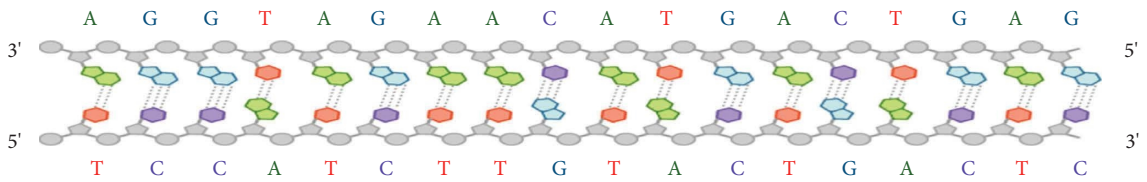



Figure 2: Nucleotide sequences for a DNA sample of two strands [27].

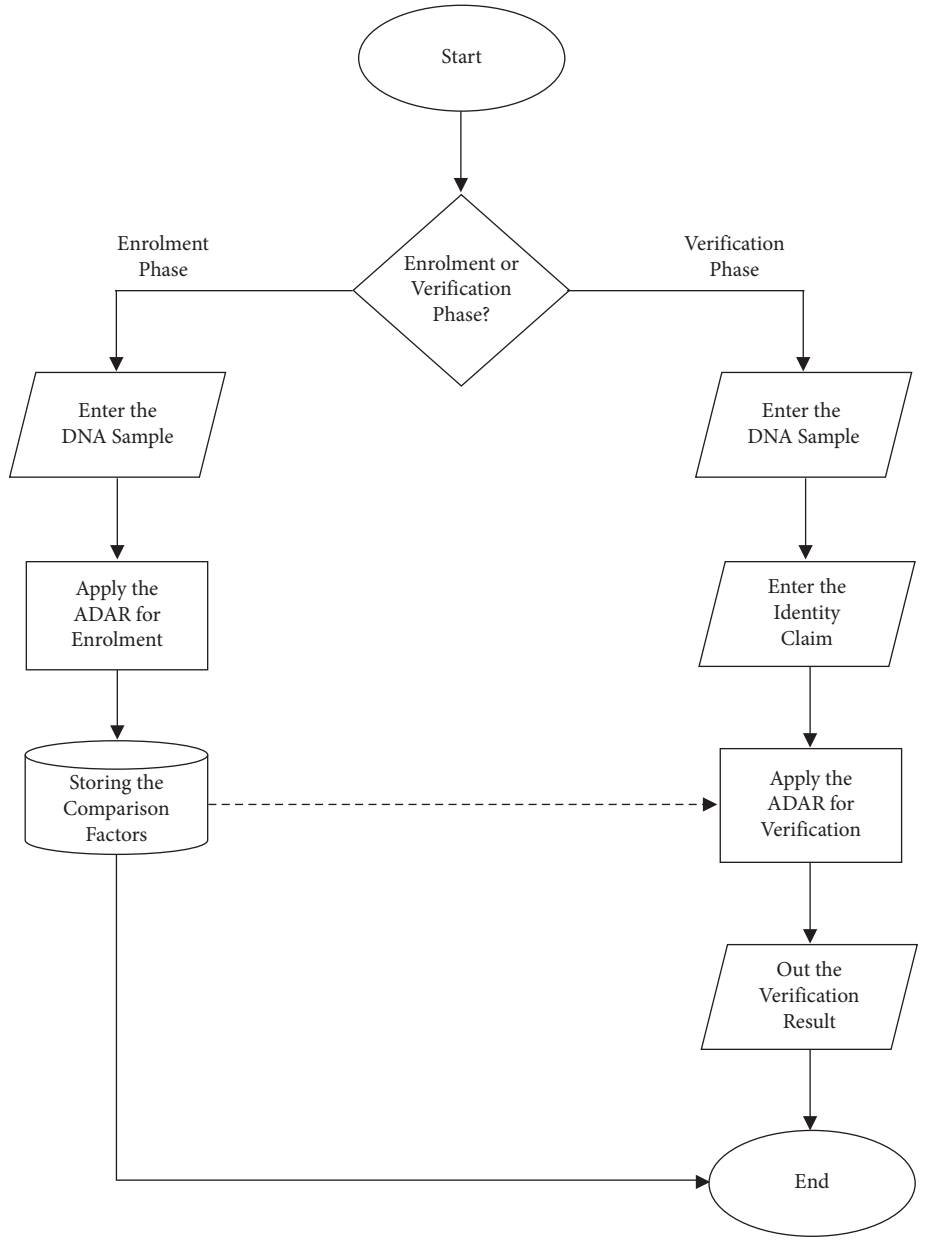



Figure 3: Flowchart for the proposed ADAR with the two phases of enrolment and verification.

## 4. Results and Discussion

*4.1. Datasets Descriptions.* Four datasets are employed in this paper: these are the DNA classification (DC) [28], sample DNA sequence (SDS) [29], human DNA sequences (HDS) [30], and DNA sequences (DS) [31]. Each one of these datasets consists of many DNA sequences of nucleotides (A, G, T, and C). The DC database involves 106 samples, the SDS dataset includes 426 samples, the HDS dataset contains 4380 samples, and the DS dataset has 11738 samples. All samples are used as strings of DNA sequences for nucleotides.

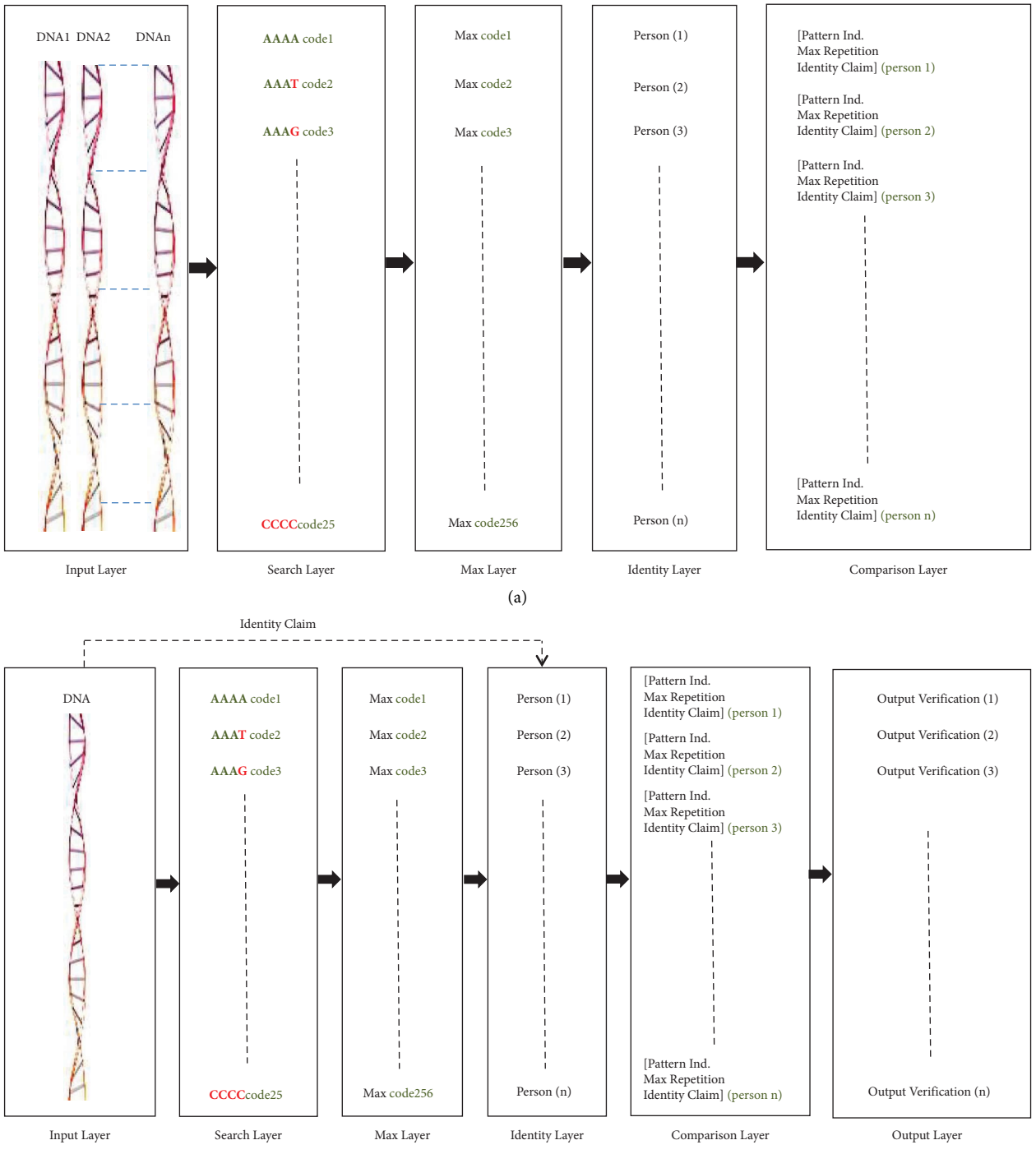



Figure 4: Proposed ADAR architecture for the two phases of (a) enrolment phase and (b) verification phase.

In more details, such datasets with their total numbers provide huge numbers of probabilities for clients and imposters, as shown in Table 2.

*4.2. ADAR System.* The proposed ADAR approach is constructed within a system. It is applied four times, each for an employed dataset. The ADAR is implemented in both phases of enrolment and verification. Simple yet effective graphical unit interfaces (GUIs) are designed and provided. Figure 5 shows first GUI, which has 5 essential buttons:

(1) Load dataset: it is responsible for loading the dataset and applying the enrolment phase.

(2) Input DNA pattern: it allows entering a DNA sequence for the verification phase, as demonstrated in

Table 1: Parameters used for the ADAR analysis.

| Parameter | Values |
|---|---|
| DNA sequence | Changeable |
| Nucleotide pattern | Quaternary |
| Nucleotide pattern probability | 256 |
| Overall ADAR layers | 6 |

Table 2: Used datasets and their total numbers with the probabilities for clients and imposters.

| Dataset | No. of utilized samples | No. of probabilities for clients | No. of probabilities for imposters |
|---|---|---|---|
| DC | 106 | 106 | 11130 |
| SDS | 426 | 426 | 181050 |
| HDS | 4380 | 4380 | 19180020 |
| DS | 11738 | 11738 | 137768906 |

Figure 6, where the requesting window to enter a DNA sequence and an example of providing a DNA sequence are shown.

(3) Input identity claim: it facilitates entering an identity claim for the verification phase, as illustrated in Figure 7, where a request window to enter an identity claim and an example of providing an identity claim are given.

(4) Result: it is for performing the verification process and displaying the result of accepting or rejecting the identity claim.

(5) End: It is for stopping and closing the ADAR system. Otherwise, the system stays working and can be used for other information.

As mentioned, the verification result should include accepting or rejecting the identity claim. Figure 8 shows both expected verification results in the ADAR system, where the output of rejecting the identity claim and the output of accepting the identity claim are displayed. Rejecting the identity claim is reported as incorrect identity with a red colored icon and accepting the identity claim is reported as correct identity with a blue-colored icon.

*4.3. Results Discussion.* For evaluating the generalization of any ADAR system, holding out separate testing samples with effective loop instructions can be used. This causes intensive evaluations as: 106 clients and 11130 imposters for the DC datasets; 426 clients and 181050 imposters for the SDS dataset; 4380 clients and 19180020 for the HDS dataset; and 11738 clients and 137768906 imposters for the DS dataset.

It can be concluded that the ADAR system was successfully constructed. Furthermore, very high verification performance can be attained for each of the four datasets, as false acceptance rate (FAR) equals to 0%, and false rejection rate (FRR) equals to 0%. It can also be highlighted that the artificial intelligence system in ADAR is user-friendly and easy to implement.

Additional metrics are also considered, these are the precision, recall, loss, and $F1$-score. In addition, receiver operating characteristic (ROC) curve and confusion

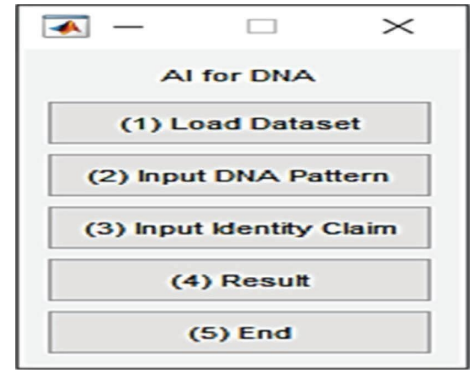



Figure 5: First GUI window in any constructed ADAR system.

matrices are also provided, as given in Figures 9 and 10, respectively. For all the employed datasets, the following values are computed: Precision = 1, Recall = 1, Loss = 0, $F1$-score = 1, and area under the curve (AUC) = 1. This is expected as all false positive verifications and all false negative verifications have 0 values, as they are demonstrated in the confusion matrices.

Time spent for an ADAR verification has been measured, and it attained an interesting outcome of around 0.23 second. This measurement was carried out on a computer with the following specifications: a hp laptop, an Intel Core i7 processor, 2.70 GHz processor speed, and 8 GB main memory.

*4.4. ADAR Limitations.* The proposed ADAR approach still has limitations and challenges to be considered. Examples of these are as follows:

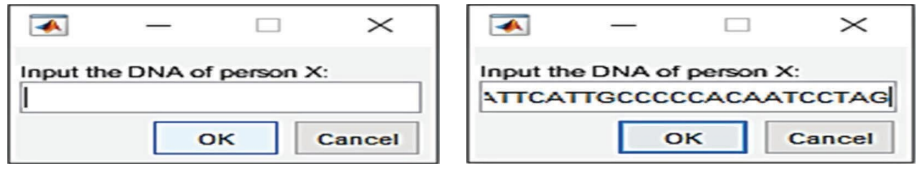
(i) It cannot be utilized for DNA samples that have no nucleotides (having different values instead).

(ii) It is assigned for the verification, so, it requires adaptation for the identification too.

(iii) It is not a machine learning technique; therefore, it is suggested to be developed in this direction.

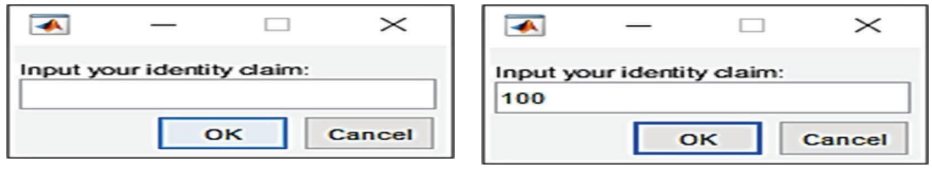
*4.5. Comparisons.* Comparisons between the proposed ADAR approach and state-of-the-art studies are considered, as given in Table 3.

This table shows performance of state-of-the-art studies, which are conducted with the Unique Personal DNA Pattern (UPDP) method. They use the same employed datasets but with the numbers of samples as: 106 samples for the DC, 426 samples for the SDS, 500 samples for the HDS, and 1000 samples for the DS. Manhal et al. [22, 32] focus on identification and have reported the FAR achievements as: 2.07%, 1.41%, 0.26%, and 0.75% for the DC, SDS, HDS, and DS, respectively. Ahmad et al. [32, 33] work on verification and have recorded the FAR results as: 0.32%, 0.31%, 0%, and 0.16% for the DC, SDS, HDS, and DS, respectively. The verification tasks using our ADAR can achieve even better performances. Significantly, it accepts full numbers of samples for all employed datasets: 106 samples for the DC, 426 samples for the SDS, 4380 samples



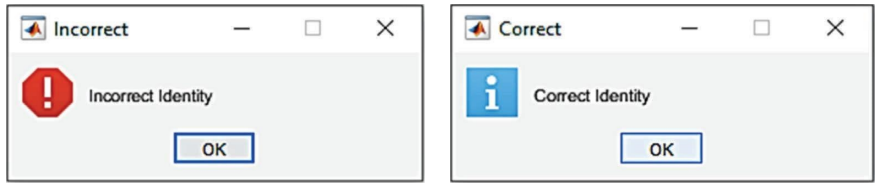

(a) (b)

Figure 6: Windows to input a DNA sequence for the verification phase: (a) requesting window to enter a DNA sequence, and (b) example of providing a DNA sequence.




(a) (b)

Figure 7: Windows to input an identity claim for the verification phase: (a) requesting window to enter an identity claim and (b) example of providing an identity claim.




(a) (b)

Figure 8: Verification results of accepting or rejecting the identity claim in the ADAR system: (a) the output of rejecting the identity claim, and (b) the output of accepting the identity claim.

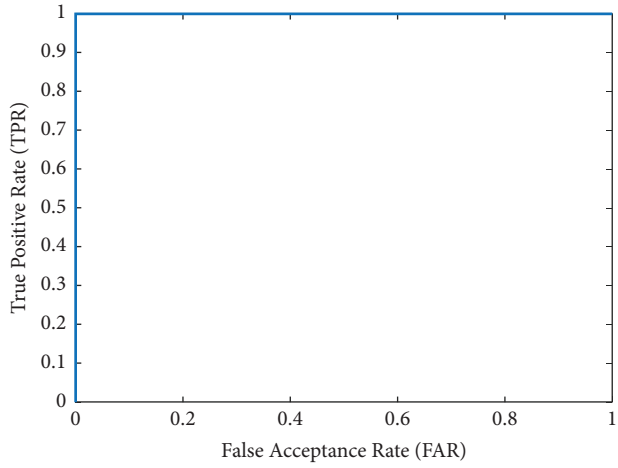



Figure 9: ROC curve of each ADAR verification.

for the HDS, and 11783 samples for the DS. Each one of the datasets can benchmark a remarkable FAR performance of 0% by using the proposed system. The FRR can be reported as 0% for any method, recognition (verification or identification), and dataset.

As a summary, the proposed system which uses the ADAR approach for verification has the capability to provide superior performance compared to previous state-of-the-art studies. It also accepts the full numbers of samples for all employed datasets. It can provide high reliabilities and performance.

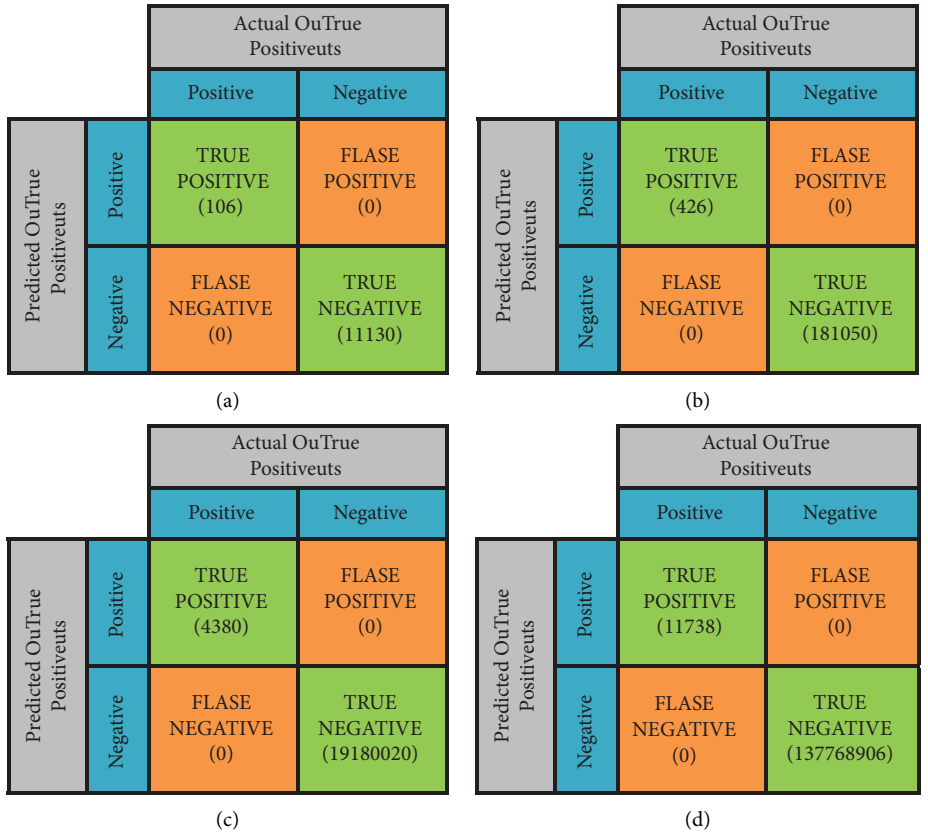



Figure 10: Confusion matrices of: (a) ADAR verification for the DC dataset, (b) ADAR verification for the SDS dataset, (c) ADAR verification for the HDS dataset, and (d) ADAR verification for the DS dataset.

Table 3: Comparisons between the proposed ADAR approach and state-of-the-art studies.

| Reference | Approach or method | Recognition | Dataset | No. of utilized samples | FAR (%) | FRR (%) |
|---|---|---|---|---|---|---|
| Manhal et al. [22, 32] | Unique personal DNA pattern (UPDP) | Identification | DC | 106 | 2.07 | 0 |
| | | | SDS | 426 | 1.41 | 0 |
| | | | HDS | 500 | 0.26 | 0 |
| | | | DS | 1000 | 0.75 | 0 |
| Ahmad et al. [32, 33] | UPDP | Verification | DC | 106 | 0.32 | 0 |
| | | | SDS | 426 | 0.31 | 0 |
| | | | HDS | 500 | 0 | 0 |
| | | | DS | 1000 | 0.16 | 0 |
| Proposed system | ADAR | Verification | DC | 106 | 0 | 0 |
| | | | SDS | 426 | 0 | 0 |
| | | | HDS | 4380 | 0 | 0 |
| | | | DS | 11738 | 0 | 0 |

## 5. Conclusion

This paper provides a new artificial intelligence approach called the ADAR. It has been proposed for person verification by DNA nucleotides. ADAR works on two main phases: enrolment and verification. During the enrolment phase, DNA samples are received, processed, and stored for their unique information. In the verification phase, a DNA sample and identity claim are provided and processed, and their unique information is compared with the stored ones to make a verification decision. The ADAR approach involves multiple layers: input layer for receiving a DNA sample, search layer for counting the frequencies of repeated quaternary patterns of nucleotides, max layer for specifying the maximum frequency among the repeated patterns, identity layer for storing or matching the identity claims, comparison layer for assigning comparison factors, and the output layer for providing the verification decision in the verification phase.

A system is also presented in this study; it implements the proposed ADAR. Moreover, four datasets, namely, the DC, SDS, HDS, and DS are employed. Remarkable

performances can be achieved as 0% FAR and 0% FRR for applying the ADAR in a system of any employed dataset. Comparisons with state-of-the-art studies are also illustrated. The ADAR approach can overcome previous proposed methods or approaches. In addition to its ability to accept the full number of DNA samples for any employed dataset. It can be revealed that the ADAR can deal with a huge number of DNA samples.

In the future, multiple considerations can be suggested such as developing the ADAR to be used for identification and adapting it for machine learning.

## Data Availability

The (DNA classification (DC), sample DNA sequence (SDS), human DNA sequences (HDS), and DNA sequences (DS)) data used to support the findings of this study are available from the corresponding author upon request.

## Disclosure

The research was performed as a part of the employment of authors where the employer's name is Northern Technical University.

## Conflicts of Interest

The authors declare that they have no conflicts of interest.

## References


[1] A. K. Jain, H. Lin, S. Pankanti, and R. Bolle, "An identity-authentication system using fingerprints," *Proceedings of the IEEE*, vol. 85, no. 9, pp. 1365–1388, 1997.

[2] L. H. Albak, R. R. O. Al-Nima, and A. H. Salih, "Palm print verification based deep learning"," *TELKOMNIKA (Telecommunication Computing Electronics and Control)*, vol. 19, no. 3, p. 851, 2021.

[3] R. Rafi Omar Al-Nima, "Personal identification based on Iris patterns," M.Sc. thesis, Technical Engineering College of Mosul, Mosul, Iraq, 2006.

[4] M. T. S. Al-Kaltakchi, R. R. O. Al-Nima, M. Alfathe, and M. A. M. Abdullah, "Speaker verification using cosine distance scoring with i-vector approach," in *Proceedings of the 2020 International Conference on Computer Science and Software Engineering (CSASE)*, pp. 157–161, Hoboken, NJ, USA, April 2020.

[5] V. Akriti, V. Moghaddam, and A. Anwar, "Data-driven behavioural biometrics for continuous and adaptive user verification using Smartphone and Smartwatch," *Sustainability*, vol. 12, no. 14, 2022.

[6] L. Devaraj and K. Modi, "Advancements in biometric technology with artificial intelligence," 2022, https://arxiv.org/abs/2212.13187.

[7] A. Jain, A. Ross, and S. Prabhakar, "An introduction to biometric recognition," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 14, no. 1, pp. 4–20, 2004.

[8] S. O. Ali, R. R. Al-Nima, and E. A. Mohammed, "Individual recognition with deep earprint learning," in *Proceedings of the 2021 International Conference on Communication & Information Technology (ICICT)*, pp. 304–309, London, UK, December 2021.

[9] C. C. Aggarwal, *Data Classification*, Springer International Publishing, Berlin, Germany, 2015.

[10] S. Ilankumaran and C. Deisy, "Multi-biometric authentication system using finger vein and iris in cloud computing," *Cluster Computing*, vol. 22, no. 1, pp. 103–117, 2019.

[11] N. H. Faleh and K. H. Al-Saedi, "Forensic biometrics identification system for DNA profile human based on association rules," *Journal Port Science Research*, vol. 150, 143 pages, 2022.

[12] N. D. Esiobu, I. M. Ezeonu, and F. Nwaokorie, "Principles and techniques for deoxyribonucleic acid (DNA) manipulation," *Medical Biotechnology, Biopharmaceutics, Forensic Science and Bioinformatics*, vol. 3, p. 32, 2022.

[13] A. Harbola, D. Negi, M. Manchanda, and R. K. Kesharwani, "Bioinformatics and biological data mining," in *Bioinformatics*, pp. 457–471, Academic Press, Cambridge, MA, USA, 2022.

[14] S. Mitra, "Computational intelligence in bioinformatics," *Transactions on Rough Sets*, vol. 3, pp. 134–152, 2005.

[15] W. You, *Classification of DNA Sequences Basing on the Dinucleotide Compositions*, ISCID, Washington, DC, USA, 2009.

[16] M. Khashei, A. Zeinal Hamadani, and M. Bijari, "A fuzzy intelligent approach to the classification problem in gene expression data analysis," *Knowledge-Based Systems*, vol. 27, pp. 465–474, 2012.

[17] E. Pashaei, M. Ozen, and N. Aydin, "Splice site identification in human genome using random forest," *Health Technology*, vol. 7, no. 1, pp. 141–152, 2017.

[18] E. Pashaei and N. Aydin, "Markovian encoding models in human splice site recognition using SVM," *Computational Biology and Chemistry*, vol. 73, pp. 159–170, 2018.

[19] F. Kaniwa and M. Phuthego, "A practical algorithm for DNA pattern searching using DatabaseBased approach," in *Proceedings of the 2018 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, vol. 1, pp. 1484–1489, Madrid, Spain, December 2018.

[20] T. Sun, Y. Wei, W. Chen, and Y. Ding, "Genome-wide association study-based deep learning for survival prediction," *Statistics in Medicine*, vol. 39, no. 30, pp. 4605–4620, 2020.

[21] A. Alatrany, A. Hussain, J. Mustafina, and D. Al-Jumeily, "A novel hybrid machine learning approach using deep learning for the prediction of Alzheimer disease using genome data Intelligent Computing Theories and Application," in *Proceedings of the 17th International Conference, ICIC 2021*, Shenzhen, China, August 2021.

[22] M. A. Saleh Al-Hussein, R. R. Omar Al-Nima, and W. L. Woo, "Applying the deoxyribonucleic acid (DNA) for people identification," *Harbin Gongye Daxue Xuebao/Journal of Harbin Institute of Technology*, vol. 54, no. No. 8, pp. 112–122, 2022.

[23] L. Rukhsar, W. H. Bangyal, M. S. Ali Khan, A. A. Ag Ibrahim, K. Nisar, and D. B. Rawat, "Analyzing RNA-seq gene expression data using deep learning approaches for cancer classification," *Applied Sciences*, vol. 12, no. 4, p. 1850, 2022.

[24] B. A. Hamed, O. A. S. Ibrahim, and T. Abd El-Hafeez, "A survey on improving pattern matching algorithms for biological sequences," *Concurrency and Computation: Practice and Experience*, vol. 34, no. 26, 2022.

[25] O. A. S. Ibrahim, B. A. Hamed, and T. A. El-Hafeez, "A new fast technique for pattern matching in biological sequences," *The Journal of Supercomputing*, vol. 79, no. 1, pp. 367–388, 2023.

[26] B. A. Hamed, O. A. S. Ibrahim, and T. Abd El-Hafeez, "Optimizing classification efficiency with machine learning

techniques for pattern matching," *Journal of Big Data*, vol. 10, no. 1, p. 124, 2023.

[27] A. Bruce, B. Alberts, A. Johnson et al., "Chromosomal DNA and its packaging in the chromatin fiber," *Molecular Biology of the Cell*, Garland Science, New York, NY, USA, 4th edition, 2002.

[28] P. Ashish, "DNA-classification," 2022, https://www.kaggle.com/code/ashishpawar511/dna-classification/data.

[29] S. Putchala, "Sample DNA sequence," 2020, https://www.kaggle.com/sreshta140/covid19-genome-sequence?select=sequence.fasta.

[30] S. Prakash, "Human DNA sequences," 2021, https://www.kaggle.com/datasets/sooryaprakash12/human-dna-sequences?select=human_data.txt.

[31] N. Singh Chauhan, "DNA sequence," 2020, https://www.kaggle.com/datasets/nageshsingh/dna-sequence-dataset?select=human.txt.

[32] M. A. Saleh Al-Hussein, R. Rafi Omar Al-Nima, and W. L. Woo, *Deoxyribonucleic Acid (DNA) for Individual Recognition*, Noor Publishing, Saarbrücken, Germany, 2022, https://www.noorpublishing.com/catalogue/details/ae/978-620-4-72413-3/deoxyribonucleic-acid-dna-forindividual-recognition.

[33] M. A. Saleh Al-Hussein, R. R. O. Al-Nima, and W. L. Woo, "Employing deoxyribonucleic acid (DNA) for personal verification," *International Journal of Health Sciences*, vol. 6, no. S9, pp. 126–140, 2022.