Bulletin of Electrical Engineering and Informatics
Vol. 13, No. 2, April 2024, pp. 1344~1350
ISSN: 2302-9285, DOI: 10.11591/eei.v13i2.6198

Identifying deoxyribonucleic acids of individuals based on their chromosomes by proposing a special deep learning model

Raid Rafi Omar Al-Nima[1], Musab Tahseen Salahaldeen Al-Kaltakchi[2], Hasan A. Abdulla[1]
[1]Technical Engineering College of Mosul, Northern Technical University, Mosul, Iraq
[2]Department of Electrical Engineering, College of Engineering, Mustansiriyah University, Baghdad, Iraq

Article Info

Article history:

Received Mar 11, 2023
Revised Jul 13, 2023
Accepted Aug 4, 2023

Keywords:

Chromosomes
Deep learning
Deoxyribonucleic acid
DNA identification
Machine learning

ABSTRACT

One of the most significant physiological biometrics is the deoxyribonucleic acid (DNA). It can be found in every human cell as in hair, blood, and skin. In this paper, a special DNA deep learning (SDDL) is proposed as a novel machine learning (ML) model to identify persons depending on their DNAs. The proposed model is designed to collect DNA chromosomes of parents for an individual. It is flexible (can be enlarged or reduced) and it can identify one or both parents of a person, based on the provided chromosomes. The SDDL is so fast in training compared to other traditional deep learning models. Two real datasets from Iraq are utilized called: Real Iraqi Dataset for Kurd (RIDK) and Real Iraqi Dataset for Arab (RIDA). The results yield that the suggested SDDL model achieves 100% testing accuracy for each of the employed datasets.

Corresponding Author:

Hasan A. Abdulla
Department of Mechanical Power Technology Engineering, Technical Engineering College
Northern Technical University
Mosul, Iraq
Email: hasan.alsarraf@ntu.edu.iq

1. INTRODUCTION

Deoxyribonucleic acid (DNA) carries the genetic information for humans and practically all other creatures. The DNA of an individual may be found in practically all of their cells. Nuclear DNA is a type of DNA that is found inside the cell nucleus. There is also a tiny quantity of DNA in mitochondria, which is called the mitochondrial DNA or mtDNA [1]–[7].

The code that stores an individual's information in the DNA is made up of four chemical bases: adenine (A), guanine (G), cytosine (C), and thymine (T). Over 99% of the 3 billion bases in each person are identical. The order or sequence of these bases determines the information that is responsible for forming the creature, much like how the letters of the alphabet appear in a specific order to create words and sentences. The remaining DNA percentage is so valuable as it differs between individuals, so, it can be used for recognition. Replicating the chemical bases in a DNA is a crucial characteristic as it is exploited as a blueprint for individual's recognition. The blueprint for replicating or sequencing the bases is found in the double helix DNA strands [1]–[7].

Two lengthy strands of nucleotides combine to form the double helix spiral that is DNA. A nucleotide is composed up of a base, a sugar and a phosphate. DNA nucleotides combine to form units of base pairs when A is combined with T and C is combined with G. Each base also has a phosphate and sugar molecule attached to it. In the double helix structure, the base pairs act as the rungs and the sugar and phosphate molecules as the vertical side rails of the ladder. Figure 1 [1]–[7] provide a DNA demonstration. Each cell contains 46 long structures called chromosomes that are distributed with DNA instructions. These

chromosomes [1]–[7] are made up of many smaller pieces of DNA, called genes. There are two sources of chromosomes, which are the parents. That is, a chromosome comes from a mother and another chromosome comes from a father.

The main aim and contribution of this paper is generating a novel deep learning model. It is named the special DNA deep learning (SDDL). This model is employed for identifying DNAs of individuals based on their chromosomes. Following the introduction, the structure of this essay is as follows: section 2 reviews prior work, section 3 demonstrates the proposed SDDL model, section 4 exhibits the findings and discusses the outcomes, and section 4 concludes the paper.

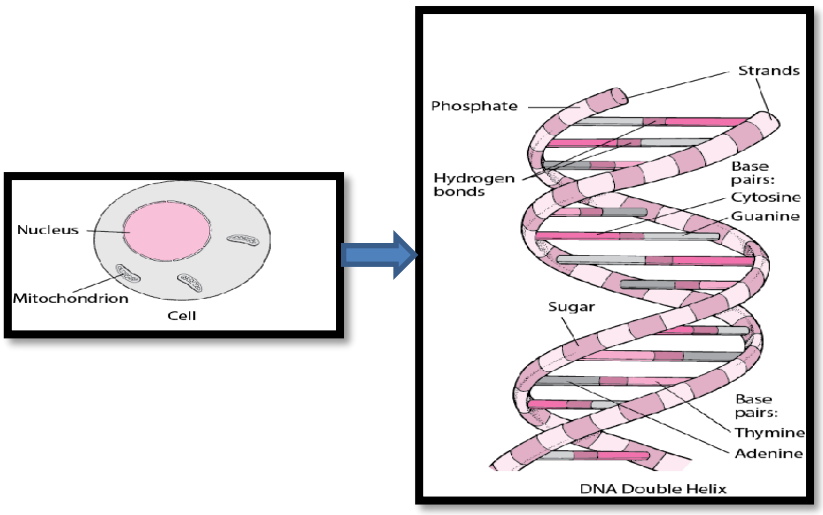

Figure 1. A DNA demonstration for the location and shape

2. PRIOR WORK

Minaee *et al*. [1] presented an overview related to the DNA as a biometric data identifier. A comprehensive survey and review of different types of biometrics are presented in [2]–[7]. The objective of the research in [8] was to develop an algorithm that could analyze the DNA sequence of an systemic lupus erythematosus (SLE) patient and predict the killer T-cell responses. Comparing a gene's DNA coding sequence from the reference genome to that gene's coding sequence in a patient's genome was the method used to identify gene variation in a patient. The threshold for significant single nucleotide polymorphisms is 0.1% of the DNA sequence that codes for a gene's length. Using the suggested approximation sequence matching method, the matching was done. The findings indicate that each of the 16 subjects will have autoimmune killer T-cells. Additionally, the algorithm's accuracy and predictive power both reached 80%. The DNA is used to identify individuals [9]. An efficient method is used to find the distinctive DNA patterns. The term unique personal DNA pattern (UPDP) is employed. Four datasets are used in this article. These are for DNA classification (DC), DNA sequences (DS), sample DNA sequence (SDS), and human DNA sequences (HDS). Identification yields outcomes that are so fascinating and amazing. False acceptance rates (FARs) were achieved for the DC, SDS, HDS, and DS are 2.07%, 1.41%, 0.26%, and 0.75%, respectively. However, for the four datasets, all false rejection rates (FRRs) are recorded as 0%. Two DNA sequencing methods are evaluated in [10], where these algorithms were the Rabin-Karp (R-K) and maximum common substream (MCS). Different code implementations and methods were used to evaluate these two approaches. Accuracy and performance were the two parameters used to assess the work. Study's goal Afolabi and Akintaro [11] was to present a summary of current technological advancements in the area of biometric security, with a focus on the effects that the usage of DNA-based biometric systems on both human lives and cyber security. Additionally, creating a biometric system based on the DNA for identifying people in order to lower the level of precision at which current technologies are insufficient for a system of universal identity (ID). Signal processing was utilized to condense DNA sequences [12]. While, the DNA typing based on forensics was discussed in [13], [14]. Further new studies are provided for different DNA applications [15]–[21]. It can be noticed that suggesting a special machine learning (ML) model which is able to identify an individual with his/her parents is valuable. This study focuses on proposing a novel deep learning approach that can provide such facility.

3. PROPOSED METHOD

In this paper, a new deep learning model or network termed the SDDL is created. This network has five layers, these are the: input layer (chromosome layer), the 1st hidden layer (distance layer), the 2nd hidden layer (impulse response (IR) layer), the 3rd layer (concatenation layer) and the output layer (decision layer). The first two hidden layers represent a feature extraction part, whereas, the last two layers represent a classifier part. The infrastructure of the novel SDDL model is given in Figure 2.

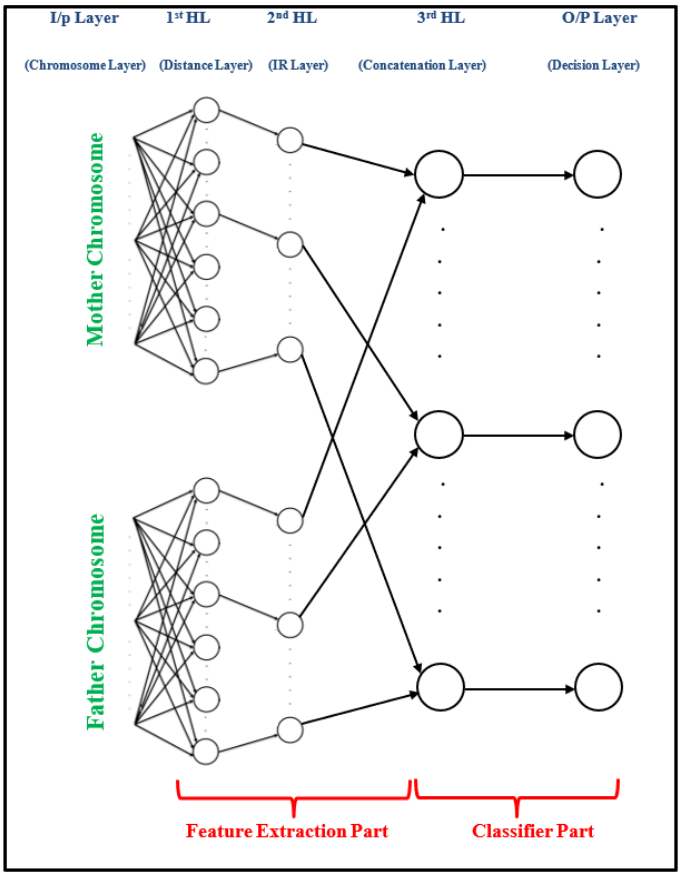

Figure 2. Infrastructure of the novel SDDL model

The input layer accepts two input vectors (X_1 and X_2). The first vector X_1 represents input values from the chromosome of a mother. The second vector X_2 represents input values from the chromosome of a father. The 1st hidden layer calculates Euclidean distance between the inputs and weights for each chromosome as in (1) and (2):

$$Z\_(i1,j1) = \|X\_1 - W\_(i1,j1)\|, i1 = 1,2,\dots,n1, j1 = 1,2,\dots,m1 \quad (1)$$

$$Z\_(i2,j2) = \|X\_2 - W\_(i2,j2)\|, i2 = 1,2,\dots,n2, j2 = 1,2,\dots,m2 \quad (2)$$

where $Z_{i1,j1}$ represents a node value of the 1st hidden layer for the chromosome of a mother, $W_{i1,j1}$ represents the weight vector of the 1st hidden layer for the chromosome of the mother, $n1$ represents the number of chromosome values for the mother, $m1$ represents the number of chromosome training vectors for the mother, $Z_{i2,j2}$ represents a node value of the 1st hidden layer for the chromosome of a father, $W_{i2,j2}$ represents the weight vector of the 1st hidden layer for the chromosome of the father, $n2$ represents the number of chromosome values for the father and $m2$ represents the number of chromosome training vectors for the father. The 2nd hidden layer computes an IR (δ_{j1} or δ_{j2}) of the calculated Euclidean distance, where δ_{j1} represents an IR for the chromosome of the mother and δ_{j2} represents an IR for the chromosome of the father. A demonstration of the employed IR function in the 2nd hidden layer is shown in Figure 3.

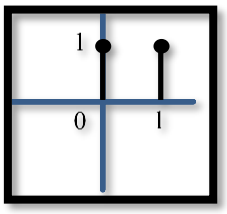

Figure 3. A demonstration of the employed IR function in the 2nd hidden layer

In fact, the outcome of the Euclidean distance is a value of either 0 or 1. These values are within an acceptable tolerance that is acceptable in the employed real Iraqi datasets. The equations of the IR function here can be calculated as (3) and (4):

$$\delta_{j1} = \begin{cases} 1 & \text{if } Z_{i1,j1} = 0 \text{ or } 1 \\ 0 & \text{otherwise} \end{cases} \quad (3)$$

$$\delta_{j2} = \begin{cases} 1 & \text{if } Z_{i2,j2} = 0 \text{ or } 1 \\ 0 & \text{otherwise} \end{cases} \quad (4)$$

Hence, the 3rd layer concatenates together the values of δ_{j1} and δ_{j2}. As mentioned, δ_{j1} is for the mother's chromosome and δ_{j2} is for father's chromosome. The 3rd layer's equation can be expressed as (5):

$$C_k = \begin{cases} 11 & \text{if } \delta_{j1} = 1 \text{ and } \delta_{j2} = 1 \\ 10 & \text{if } \delta_{j1} = 1 \text{ and } \delta_{j2} = 0 \\ 01 & \text{if } \delta_{j1} = 0 \text{ and } \delta_{j2} = 1 \end{cases}, k = 1,2,\dots,q \quad (5)$$

where C_k represents a node value of the 3rd hidden layer for both chromosomes and q represents the number of chromosome values for the mother ($n1$) or father ($n2$) as $n1 = n2$.

Consequently, the output layer produces the decision outcomes. Each decision outcome is represented by the required identification value. It can be computed according as (6):

$$D_k = \begin{cases} 2 & \text{if } C_k = 11 \\ 1 & \text{if } C_k = 10 \\ -1 & \text{if } C_k = 01 \end{cases}, k = 1,2,\dots,q \quad (6)$$

where D_k represents the identification value of the output layer.

It is worth mentioning that the training weights are directly initialized from the input training vectors. Same idea of initializing the training weights is explained in [22]. This yields important advantages for the proposed SDDL model as: i) it is so flexible, where it allows adding and removing hidden and output nodes; ii) it does not require to do iterations during the training stage, so, its train is so fast; iii) it does not deceive by the problem of local error in training as other deep learning models which use the backpropagation training algorithm; and iv) it has the ability to identify one or two parents of an individual.

4. RESULTS AND DISCUSSION

4.1. Datasets descriptions

Two real datasets from Iraq are employed in this study. The first one is named the Real Iraqi Dataset for Kurd (RIDK). The second one is called the Real Iraqi Dataset for Arab (RIDA). Any individual has a chromosome of 30 values, 15 values from a mother and 15 values from a father. The RIDK dataset has chromosome values for 52 persons. Whereas, The RIDA dataset has chromosome values for 200 persons.

As a tolerance of ±1 for each value in a chromosome is acceptable in real consideration by Iraqi forensic medicine, this has been employed to establish training data. So, the augmentation of applying ±1 for each value in a chromosome is used for both datasets. As such, 1560 training data are produced for the RIDK dataset and 6000 training data are generated for the RIDA dataset. On the other hand, the real values of both datasets are used in the testing phase. That is, 52 testing data are applied for the RIDK and 200 testing data are utilized for the RIDA dataset.

4.2. SDDL performances

To evaluate the SDDL, the accuracies and times are considered. They are applied for the two employed datasets. Actually, accuracies and times are considered for fully applying the employed RIDK and RIDA datasets. First of all, Table 1 shows the SDDL performances of accuracies and times for both datasets.

Table 1. SDDL performances of accuracies and times for the employed datasets

Employed datasets	RIDK dataset	RIDA dataset
Number of training samples	1560	6000
Number of testing samples	52	200
Training time	1.24 Sec.	6.19 Sec.
Testing time	0.10 Sec.	1.08 Sec.
Accuracy	100%	100%

The results show that the proposed system is successful with fantastic performance accuracies of 100% for both utilized datasets. In addition, the training and testing time are considered in this work and the proposed SDDL model shows very short training and testing times. It can be noticed that the SDDL is so fast in training as it reports a very short training time. This can be a remarkable outcome for this proposed approach.

To approve the SDDL validity, it is compared to other deep neural networks. It also shows outperformances compared to other models or networks for the same training and testing samples used in Table 1, the comparisons are detailed in Table 2. That is, the suggested algorithm outperformed previous deep learning models of the stacked autoencoder (SA) [23], deep autoencoder network (DAN) [24], and autoencoder deep learning (ADL) [25] in terms of its flexibility, training time, mean square error (MSE) and its ability to recognize parents. Regarding the flexibility, the SDDL can be enlarged or reduced without the requiring of re-train again. On the other hand, other compared deep learning networks have specific parameters, as numbers of hidden layers and neurons, to be determined. For training time, it is obvious that the SDDL approach has recorded the lowest training times compared to other deep learning models for both employed datasets. This can be considered as a significant advantage of the SDDL. For testing time, the proposed SDDL has taken longer time than other deep learning models. However, its testing time still small and can be acceptable especially for the RIDK dataset. If a single input is considered, the testing time will have not significant effect. The proposed SDDL outperforms previous deep neural networks as it is the only one that can provide the lowest MSE of 0 value for both employed datasets. In addition, the SDDL approach has the capability to recognize the parent or parents of identified persons. Whilst, such ability is not provided by any of the other compared deep learning models.

Table 2. Comparisons between the proposed novel SDDL and other deep neural models (for the same training and testing samples used in Table 1)

Deep learning model	Parameters	Error and time for The Ridk dataset	Error and time for The Rida dataset	Ability to recognize parents
SA [23]	NoHL=3 NoHN in the 1st HL=25 NoHN in the 2nd HL=20 NoHN in the 3rd HL=15	MSE=0.08 TRT=6.59 Sec. TET=0.04 Sec.	MSE=0.02 TRT=20.33 Sec. TET=0.01 Sec.	No
DAN [24]	NoHL=3 NoHN in the 1st HL=64 NoHN in the 2nd HL=64 NoHN in the 3rd HL=64	MSE=0.08 TRT=10.99 Sec. TET=0.01 Sec.	MSE=0.02 TRT=35.67 Sec. TET=0.02 Sec.	No
ADL [25]	NoHL= 4 NoHN in the 1st HL = 30 NoHN in the2nd HL = 30 NoHN in the 3rd HL = 30 NoHN in the 4th HL = 30	MSE=0.08 TRT=8.62 Sec. TET=0.01 Sec.	MSE=0.02 TRT=25.40 Sec. TET=0.02 Sec.	No
Proposed SDDL	NoHL=3 NoHN in the 1st HL (flexible)=2× no. of training vectors NoHN in the 2nd HL (flexible)=2× no. of training vectors NoHN in the 3rd HL (flexible)= no. of training vectors	MSE=0 TRT=1.24 Sec. TET=0.10 Sec.	MSE=0 TRT=6.19 Sec. TET=1.08 Sec.	Yes

Where NoHL is the number of hidden layers, NoHN is the number of hidden nodes, and HL is the hidden layer. TRT is the training time and TET is the testing time.

5. CONCLUSION

In order to identify people based on their DNAs, a novel SDDL model is proposed in this study. This suggested approach can identify either one parent or both parents for an individual depending on the provided chromosomes. The SDDL is flexible as it can be enlarged or reduced accordingly. During the phase of training, it does not require iterations and it does not suffer from the local error. Its training is also so quicker than other compared deep learning models. Two real datasets from Iraq are employed and termed the RIDK and RIDA. For each of both datasets, the provided approach achieves highest accuracy of 100%. It also performs better and gives the lowest MSE of 0 value compared to other deep learnings.

ACKNOWLEDGEMENTS

Author thanks to Department of Electrical Engineering, College of Engineering, Mustansiriyah University, Baghdad, Iraq and Technical Engineering College of Mosul, North Technical University, Mosul, Iraq.

REFERENCES

[1] S. Minaee, A. Abdolrashidi, H. Su, M. Bennamoun, and D. Zhang, "Biometrics recognition using deep learning: a survey," *Artificial Intelligence Review*, vol. 56, no. 8, pp. 8647–8695, Aug. 2023, doi: 10.1007/s10462-022-10237-x.

[2] J. M. Butler, "Recent advances in forensic biology and forensic DNA typing: INTERPOL review 2019–2022," *Forensic Science International: Synergy*, vol. 6, p. 100311, 2023, doi: 10.1016/j.fsisyn.2022.100311.

[3] T. Sabhanayagam, V. P. Venkatesan, and K. Senthamaraikannan, "A comprehensive survey on various biometric systems," *International Journal of Applied Engineering Research*, vol. 13, no. 5, pp. 2276–2297, 2018.

[4] M. Sharif, M. Raza, J. H. Shah, M. Yasmin, and S. L. Fernandes, "An Overview of Biometrics Methods," in *Handbook of Multimedia Information Security: Techniques and Applications*, Cham: Springer International Publishing, 2019, pp. 15–35, doi: 10.1007/978-3-030-15887-3_2.

[5] S. Dargan and M. Kumar, "A comprehensive survey on the biometric recognition systems based on physiological and behavioral modalities," *Expert Systems with Applications*, vol. 143, p. 113114, Apr. 2020, doi: 10.1016/j.eswa.2019.113114.

[6] D. H. Fernando, B. Kuhaneswaran, and B. T. G. S. Kumara, "A Systematic Literature Review on the Application of Blockchain Technology in Biometric Analysis Focusing on DNA," in *Handbook of Research on Technological Advances of Library and Information Science in Industry 5.0*, 2022, pp. 77–99, doi: 10.4018/978-1-6684-4755-0.ch005.

[7] M. Kumar and A. K. Tiwari, "Computational Intelligence Techniques for Biometric Recognition: A Review," *SSRN Electronic Journal*, 2019, doi: 10.2139/ssrn.3350260.

[8] W. L. Woo, N. O. Ephraim, F. P. Obinna, N. O. I., B. O. Jessy, and R. R. Al-Nima, "A Scalable Algorithm for Interpreting DNA Sequence and Predicting the Response of Killer T-Cells in Systemic Lupus Erythematosus Patients," *International Journal of Bioinformatics and Intelligent Computing*, vol. 1, no. 1, 2022.

[9] M. A. S. Al-Hussein, R. R. O. Al-Nima, and W. L. Woo, "Applying the deoxyribonucleic acid (DNA) for people identification," *Journal of Harbin Institute of Technology*, vol. 54, no. 8, pp. 112–122, 2022, doi: 10.11720/JHIT.54082022.13.

[10] S. Tripathi and A. K. Pandey, "Identifying DNA sequence by using stream matching techniques," in *2016 International Conference System Modeling & Advancement in Research Trends (SMART)*, 2016, pp. 334–338, doi: 10.1109/SYSMART.2016.7894545.

[11] A. Afolabi and O. Akintaro, "Design of DNA Based Biometric Security System for Examination Conduct," *Journal of Advances in Mathematics and Computer Science*, vol. 23, no. 6, pp. 1–7, Jan. 2017, doi: 10.9734/JAMCS/2017/27251.

[12] B. N. Law, "Application of signal processing for DNA sequence compression," *IET Signal Processing*, vol. 13, no. 6, pp. 569–580, Aug. 2019, doi: 10.1049/iet-spr.2018.5392.

[13] D. Primorac and M. Schanfield, "Forensic DNA Applications," in *CRC press*, Boca Raton: CRC Press, 2023.

[14] D. Palma and P. L. Montessoro, "Biometric-based human recognition systems: an overview," in *Recent Advances in Biometrics*, 2022, pp. 1–21.

[15] D. Abbott, "Finding Somerton Man: How DNA, AI Facial Reconstruction, and Sheer Grit Cracked a 75-Year-Old Cold Case," *IEEE Spectrum*, vol. 60, no. 4, pp. 22–30, Apr. 2023, doi: 10.1109/MSPEC.2023.10092395.

[16] J. Johnson, R. Chitra, and A. M. Anusha Bamini, "An Efficient Fingerprint Analysis and DNA Profiling from the Same Latent Evidence for the Forensic Applications," in *2022 Smart Technologies, Communication and Robotics (STCR)*, Dec. 2022, pp. 1–6, doi: 10.1109/STCR55312.2022.10009376.

[17] A. Michaleas, P. Fremont-Smith, C. Lennartz, and D. O. Ricke, "Parallel Computing with DNA Forensics Data," in *2022 IEEE High Performance Extreme Computing Conference (HPEC)*, Sep. 2022, pp. 1–7, doi: 10.1109/HPEC55821.2022.9926352.

[18] S. Guan, Y. Qian, T. Jiang, M. Jiang, Y. Ding, and H. Wu, "MV-H-RKM: A Multiple View-Based Hypergraph Regularized Restricted Kernel Machine for Predicting DNA-Binding Proteins," *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol. 20, no. 2, pp. 1246–1256, Mar. 2023, doi: 10.1109/TCBB.2022.3183191.

[19] Y. He *et al.*, "Predicting the sequence specificities of DNA-binding proteins by DNA Fine-tuned Language Model with decaying learning rates," *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, pp. 1–1, 2022, doi: 10.1109/TCBB.2022.3165592.

[20] Q. Yu, X. Zhang, Y. Hu, S. Chen, and L. Yang, "A Method for Predicting DNA Motif Length Based on Deep Learning," *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol. 20, no. 1, 2022, doi: 10.1109/TCBB.2022.3158471.

[21] Y. Liang *et al.*, "Weakly Supervised Deep Nuclei Segmentation with Sparsely Annotated Bounding Boxes for DNA Image Cytometry," *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol. 20, no. 1, pp. 785–795, 2022, doi: 10.1109/TCBB.2021.3138189.

[22] R. R. O. Al-Nima, "Signal Processing and Machine Learning Techniques for Human Verification Based on Finger Textures," Newcastle University, 2017.

[23] MATLAB, "Deep Learning Toolbox," in *The MathWorks Inc.*, 2020.

[24] A. S. Anaz, M. Y. Al-Ridha, and R. R. O. Al-Nima, "Signal multiple encodings by using autoencoder deep learning," *Bulletin of*

Electrical Engineering and Informatics, vol. 12, no. 1, pp. 435–440, Feb. 2023, doi: 10.11591/eei.v12i1.4229.
[25] A. S. Anaz, R. R. O. Al-Nima, and M. Y. Al-Ridha, "Multi-Encryptions System Based on Autoencoder Deep Learning Network," *Solid State Technology*, vol. 63, no. 6, pp. 3632–3645, 2020.

BIOGRAPHIES OF AUTHORS

Raid Rafi Omar Al-Nima received the B.Sc. and M.Sc. degrees in Technical Computer Engineering in 2000 and 2006, respectively. During 2006, he worked as an assistant lecturer in the Technical College of Mosul, Iraq. In 2011, he obtained the lecturer scientific title in the same college. In 2017, he accomplished his Ph.D. in the School of Electrical and Electronic Engineering at Newcastle University, UK. In 2020, he achieved the title of assistant professor in the Northern Technical University. His research interests are in the fields of pattern recognition, security, artificial intelligence, and image processing. He can be contacted at email: raidrafi3@ntu.edu.iq.

Musab Tahseen Salahaldeen Al-Kaltakchi is a lecturer in the Electrical Engineering Department, Mustansiriyah University, Baghdad-Iraq. He obtained his B.Sc. in Electrical Engineering in 1996 and obtained his M.Sc. in Communication and Electronics in 2004 from Mustansiriyah University. He was awarded a Ph.D. degree in Electrical Engineering/Digital Signal Processing from Newcastle University, UK in 2018. He is a member of the Institute of Electrical and Electronic Engineering (IEEE) and also in the Institute of Engineering and Technology (IET). His research interests include speaker identification and verification, speech and audio signal processing, machine learning, artificial intelligence, pattern recognition, and biometrics. He can be contacted at email: musab.tahseen@gmail.com and m.t.s.al_kaltakchi@uomustansiriyah.edu.iq.

Hasan A. Abdulla is an academic assistant lecturer of Computer Engineering at Northern Technical University, Mosul, Iraq. He received his bachelor's degree in Computer Technology Engineering in 2007 and master's degree in Computer Technology Engineering in 2021 from Technical Engineering College of Mosul. His current research interests include CNC software control system, deep learning, image processing, artificial intelligent, brain computer interface (BCI), EEG signal processing, and machine learning technique. He can be contacted at email: hasan.alsarraf@ntu.edu.iq.