ISSN (Print) 2794-7629
ISSN (Online) 2794- 4549
Received 15/03/2025
Accepted 20/05/2025

FULL PAPER

DNA BASED FEDERATED LEARNING

Prepared by

A.Prof.Dr.Raid Rafi Omar Al-Nima
Technical College of Engineering for Computer and Artificial Intelligence Mosul, Northern Technical University,Mosul, Iraq
raidrafi1@gmail.com

Maan A. Yahya
Northern Technical University, Mosul, Iraq
dr.maan@ntu.edu.iq

Azzah Tareq Salim Qaba
Northern Technical University, Mosul, Iraq
mti.lec317.azzah@ntu.edu.iq

Abstract

The Deoxyribonucleic Acid (DNA) can be considered as one of the most effective biometrics. It can be used in different fields such as personal identification, parental verification and relativity. In this study, federated learning is suggested. It is a type of machine learning which recently attracts researchers' attentions. Here it has been suggested to be employed for controlling the DNA samples over large areas. That is, the DNA centers have DNA samples, each DNA center can deal with its samples by using a single machine learning method. Then, according to the federated learning machine learnings information are combined by a main machine learning. In this case, all DNA centers can effectively share their information. So, there is no need to waste long correspondence time between the DNA centers to share required data. Also, there may no need to re-train any machine learning again for new DNA samples.

Keywords: DNA, Machine Learning, Pattern Recognition

The Scholar Journal for Sciences & Technology sjst

1. INTRODUCTION

Deoxyribonucleic Acid (DNA) is one of the most famous biometrics. It is utilized in various topics. Examples of such topics are personal verification (Al-Hussein, Al-Nima, et al., 2022c) and parental identification (Al-Nima et al., 2024). In addition, it has a strong capability to recognize relativity (Al-Nima et al., 2024) and to be utilized in forensics (Roewer, 2013).

The DNA is a pair of twisted strips. It is found in a chromosome inside nucleus of a cell. It consists of genes, where a gene is a DNA segment. It composes of nitrogenous bases namely: adenine, guanine, thymine and cytosine (Professional, 2022). Figure 1 shows a demonstration of principles for the DNA, genes and chromosomes.

Each DNA is constructed from a combination between two chromosomes. A chromosome comes from a father and another chromosome obtains from a mother. So, this is the key idea of recognizing relatives for an individual. Moreover, patterns of nitrogenous bases are utilized for personal recognition (Al-Hussein, Al-Nima, et al., 2022a, 2022b; Al-Nima et al., 2023).

The aim of this work is to recognize persons by using their DNAs. It has been noticed that the DNA centers are far from each other and they require a main DNA center to correspond with. This inspires our contribution here of proposing a federated learning which has the ability to let each DNA center works independently, then, combines the whole data in the main DNA center. This can provide fast correspondence of data usability between the different DNA centers.

The sections of this paper are arranged as follows: Section 1 determines for the introduction, Section 2 assigns for the literature review, Section 3 reveals the theory idea of the proposed DNA federated learning, Section 4 demonstrates the results and discussions, and Section 5 provides the conclusion.

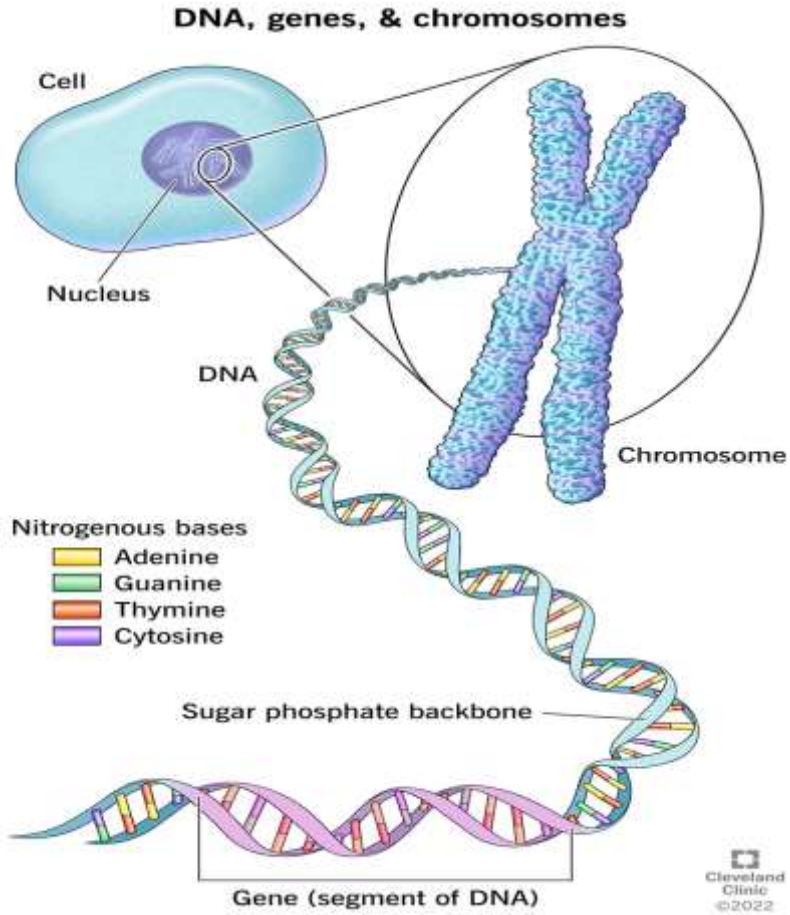


Figure 1: Demonstration of principles for the DNA, genes and chromosomes as in (Professional, 2022)

2. LITERATURE REVIEW

In the literature, there are many studies consider the DNA. Examples of such studies are as follows:

In 2021, Jacob *et al.* presented a DNA study. It was for biometric template security. It utilized DNA codec and it concentrated on transformation (Jacob et al., 2021). In the same year, Katsanis *et al.* explained United States (US) adult perspectives. The DNA was considered as a part of the study. Facial images were also utilized. Other biometrics were also investigated (Katsanis et al., 2021). In 2022, Sakr *et al.* proposed a cancelable Electrocardiogram (ECG) biometric. It was based on the fusion of deep transfer learning with amino acid and DNA approaches. It considered human authentication (Sakr et al., 2022). Also in the same year, Pandkar *et al.* worked on Three-Dimensional (3D) facial biometric. The DNA Sample was provided for applications of law enforcement. Verification was concentrated here (Pandkar et al., 2022). Further in the same year, Ephraim *et al.* suggested a scalable algorithm. It was for describing DNA sequence and expecting the killer T-cells response. Patients of systemic lupus erythematosus were considered (Ephraim et al., 2022). In 2023, Salama *et al.* illustrated secure biometric systems. DNA and bio-signals encryption was provided. Furthermore, optical spectrograms were focused (Salama et al., 2023). In the same year, Sepers *et al.* investigated developmental stress. It did not induce changes of methylation of genome-wide DNA. Wild great tit nestlings were explored (Sepers et al., 2023). In 2024, Ou *et al.* exploited impedance labeling. Free sensor of electrochemical aptamer were focused. Tetrahedral DNA nanostructures were highlighted. Doxorubicin determination was considered (Ou et al., 2024).

According to the literature and to the best of our knowledge, there is no study that worked on the federated learning for the biometric of DNA. So, this has been concentrated here.

3. PROPOSED FEDERATED LEARNING THEORY

First of all, federated learning is a machine learning method. It has the ability to combine resources and data of various organizations or areas. It can also provide direct shares with ensuring privacy and security (Liu et al., 2022). A demonstration of federated learning work is given in figure 2.

As mentioned, it has been found that the DNA centers require a main DNA center to communicate with and they are far from each other. Any DNA center has its own data and resources that cannot be shared, except with the main DNA center. This causes a big problem of consuming long correspondences time for requesting and receiving information. In addition of not allowing any sharing resources between the DNA centers. Such problems and challenges can be realized in Iraq, which is considered as a case study here.

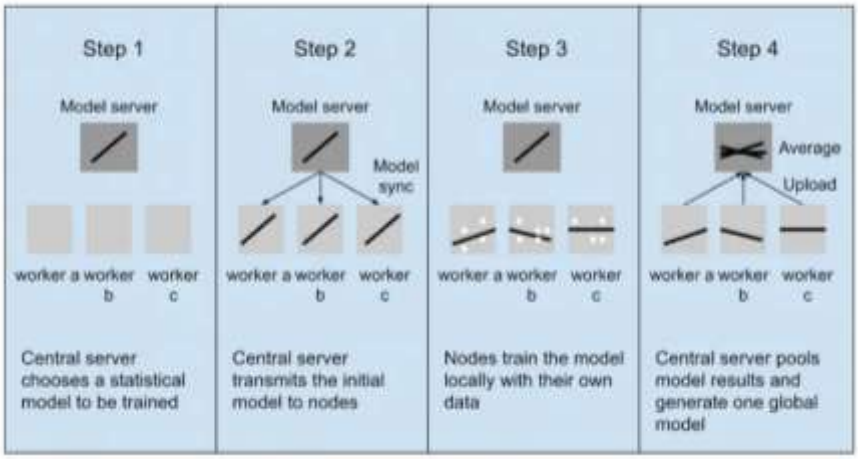


Figure 2: Demonstration of federated learning work (Zhang, 2022)

To overcome the aforementioned issues, this study proposes an efficient Federated DNA Learning

(FDL) approach. It employes multiple Radial Basis Function-Neural Networks (RBF-NNs). One RBF-NN for each DNA center and a global (or centralized) RBF-NN for the main DNA center. These

networks are utilized for individuals recognition by using their DNAs data. Figure 3 shows the general
architecture of the proposed FDL approach.

For the RBF-NN, the calculations from the input neurons to the hidden neurons can be computed according to the following equation (Fausett, 1994; Kou et al., 2010):

$$Z_j = \exp\left[-\frac{\|\mathbf{x}-\mathbf{v}_j\|^2}{2\,\alpha^2}\right] \quad (1)$$

where: Z_j represents a hidden neuron outcome value, $j = 1,2,\dots,p$, p represents the number of hidden neurons, $\mathbf{X}$ represents an input vector, $\mathbf{V}_j$ represents a hidden vector and α represents the smoothing parameter.

The calculations from the hidden neurons to the output neurons can be computed according to the following equation (Fausett, 1994):

$$Y_k = \sum_{j=1}^{p} Z_j \times V_{j,k} \quad (2)$$

Where: Y_k represents an output neuron outcome value, $k = 1,2,\dots,m$, m represents the number of output neurons and $V_{j,k}$ represents a weight between the hidden and output layers.

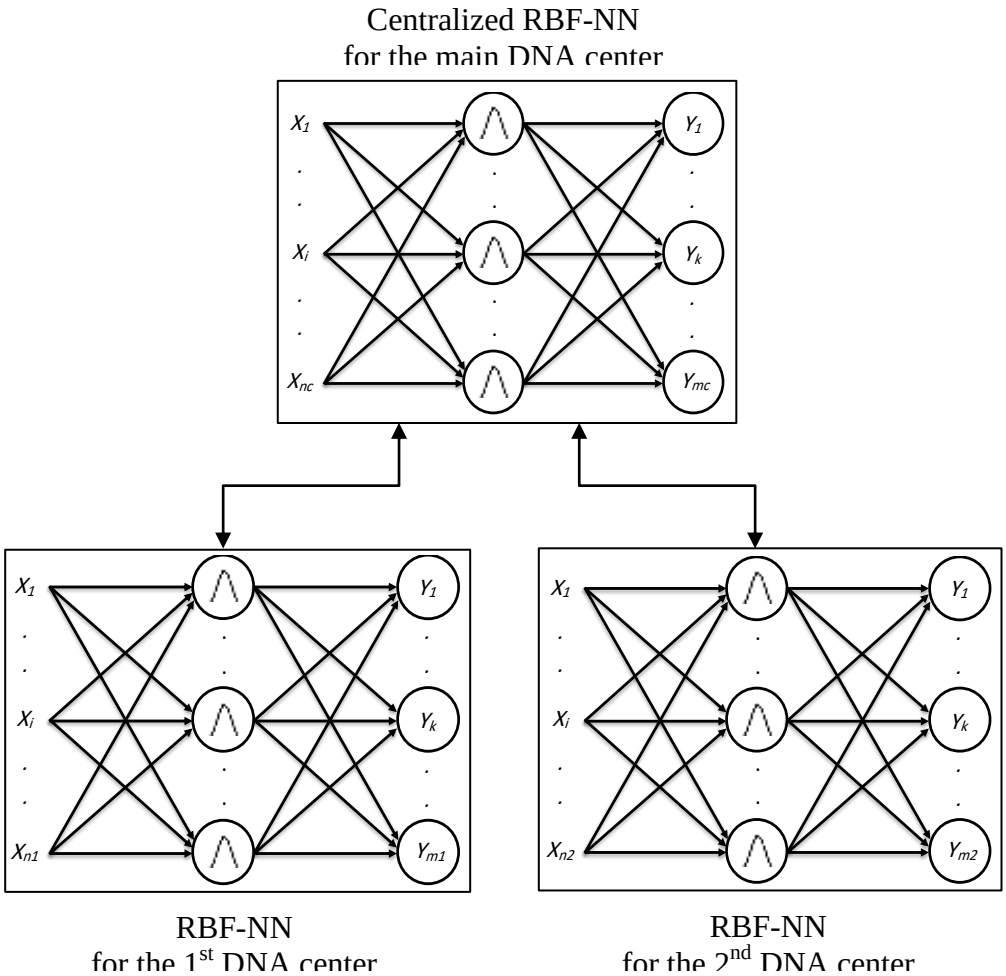


Figure 3: General architecture of the proposed FDL approach

It is worth mentioning that in the input layer $i = 1,2, \dots, n$, where: n represents the number of input neurons. Furthermore, n can be $n1$, $n2$, …, or nc, where: $n1$ and $n2$ (and so on) refer to determined DNA centers, and nc refers to the main DNA center. Similarly, p can be $p1$, $p2$, …, or pc, where: $p1$ and $p2$ (and so on) refer to determined DNA centers, and pc refers to the main DNA center. Likewise, m can be $m1$, $m2$, …, or mc, where: $m1$ and $m2$ (and so on) refer to determined DNA centers, and mc refers to the main DNA center.

Here, two RBF-NNs are suggested for two DNA centers and one global or centralized RBF-NN is suggested for the main DNA center. Hence, for all input vectors, let: $f_1(\mathbf{X})$ represents the function of the 1st RBF-NN for the 1st DNA center and $f_2(\mathbf{X})$ represents the function of the 2nd RBF-NN for the 2nd DNA center. So:

$$f_c(\mathbf{X}_T) = \{f_1(\mathbf{X}) \cup f_2(\mathbf{X})\} \quad (3)$$

where: $f_c(\mathbf{X}_T)$ represents the function of the centralized RBF-NN for the main DNA center and T represents the total number of input vectors (DNA samples).

Actually, there are significant advantages of employing the RBF-NN, some have been noticed in (Al-Nima, 2017) as:

1- It has a very short implementation time.
2- It has a flexible architecture where it can be enlarged to accept more information.
3- It stores the input training vectors in its initialized weights.

All of such advantages encourage exploiting RBF-NNs with the FDL as have been explained.

4. RESULTS AND DISCUSSIONS

First of all, it is important to highlight that two real databases in the case study of Iraq are used here. One for Kurdish people and another one for Arab people. They involve values of two chromosomes for each person. Specifically, 15 chromosome values from a father and 15 chromosome values from a mother. So, total of 30 values are obtained for each person. Moreover, the total number of Arab individuals are 200 and the total number of Kurdish individuals are 52. The total number of samples is duplicated in each database by using incremental augmentation. The augmented samples are used for the training stages, whereas, the original samples are utilized for the testing stages.

Such databases are suitable for the federated learning work in this paper, where the 1st DNA center is employed for the Arab persons and the 2nd DNA center is applied for the Kurdish persons. On the other hand, the main DNA center is exploited for both kinds of people. Figures 4, 5 and 6 show the training curves for the 1st RBF-NN of the 1st DNA center, 2nd RBF-NN of the 2nd DNA center and centralized RBF-NN of the main DNA center, respectively.

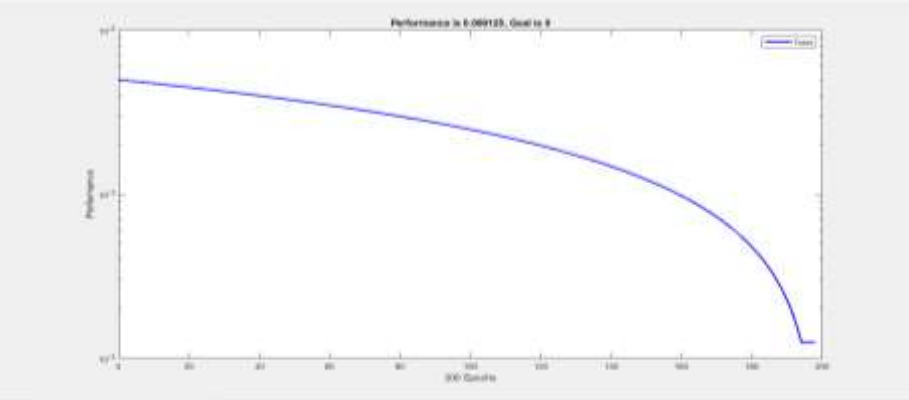


Figure 4: Training curve for the 1st RBF-NN of the 1st DNA center

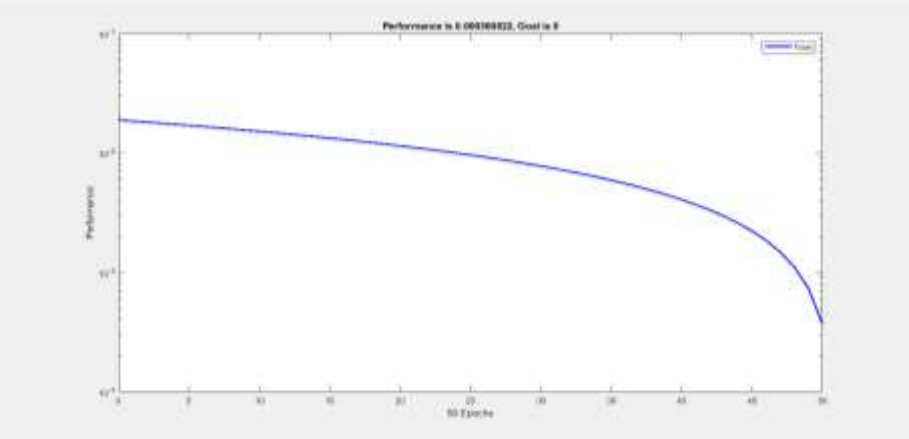


Figure 5: Training curve for the 2nd RBF-NN of the 2nd DNA center

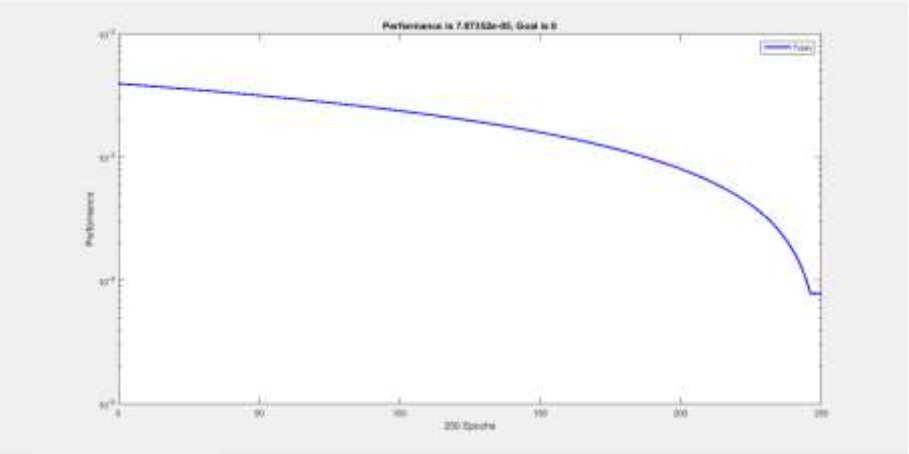


Figure 6: Training curve for the centralized RBF-NN of the main DNA center

These figures demonstrate the successfulness of all training stages where high performances are revealed. The 1st training yields that the Mean Squared Error (MSE) performance attains 0.000125 after 200 epochs. The 2nd training reaches the MSE performance equals to 0.000369822 after 52 epochs. The 3rd training obtains the MSE performance of 7.87352×10^{-05} after 250 epochs. All of these performances are significant and reported very high achievements.

For the testing stages, multiple metrics are taken into accounts. These are the MSE, Mean Absolute Error (MAE) and Mean Absolute Percentage Error (MAPE). Such metrics are provided in (SAĞIROĞLU & ÖZKAYA, 2009) and explained in (Al-Nima et al., 2014). Table 1 shows the values of these metrics for all tests of the employed federated learning networks.

Table 1: Values of error metrics of testing all employed federated learning networks

Federated Learning Network	MSE	MAE	MAPE
1st RBF-NN	0.003	0.005	0.034
2nd RBF-NN	0.009	0.019	0.033
Global or centralized RBF-NN	0.002	0.004	0.034

This table obviously reveals the obtained high performances in all tests, where low error values are attained. These values approve the efficiency and successfulness of our suggested FDL approach. It is worth highlighting that the Centralized RBF-NN in the main DNA center can share DNA information between the two DNA centers. So, there is no need to re-consider repeating the processes of providing and adding new DNA samples in the DNA center which has no any information about such new samples.

Training and testing elapsed times are also computed, by using a laptop computer of the following specifications: processor type Core-i7, memory size 8 GB, processor speed 2.70 GHz and embedded monitor card with total memory size approximately 1.5 GB. Training and testing times for all federated learning networks are given in table 2.

The times in this table can be considered as reasonable and acceptable. That is, all testing elapsed times are further less than the training elapsed times. This is reasonable as the networks require iterations (epochs) during the training stages in order to learn the applied DNA information, whereas, there are no iterations during the testing stages. Furthermore, the 1st RBF-NN elapsed times are higher than the 2nd RBF-NN elapsed times. This is due to the number of DNA samples that are applied to each network.

Table 2: Training and testing elapsed times for all federated learning networks

Federated Learning Network	Training time (sec.)	Testing time (sec.)
1st RBF-NN	2.172	0.016
2nd RBF-NN	0.544	0.013
Global or centralized RBF-NN	3.484	0.019

To illustrate, the 1st RBF-NN is assigned for the Arab database and the 2nd RBF-NN is assigned for the Kurdish database, and the number of DNA samples in the Arab database is further higher than the number of DNA samples in the Kurdish database. Noticeably, training and testing elapsed times for the centralized RBF-NN are higher than other networks. This is acceptable as the applied DNA samples here represent combination between all samples of the two databases.

5. CONCLUSION

In this paper, a novel federated learning is proposed, it is called the FDL. It is suggested to well manage DNA biometric information between DNA centers and the main DNA center. This can further reduce the correspondences between the DNA centers. Multiple RBF-NNs are considered here, where a single RBF-NN is exploited in each DNA center and a global RBF-NN is employed in the main DNA center.

Real databases in Iraq, the case study, are exploited and used. The results yield high performances. That is, low error values of the MSE, MAE and MAPE are obtained for each adopted RBF-NN. Moreover, the elapsed times of training and testing are also evaluated. It can be revealed that very fast implementations are reported for the adopted RBF-NNs. The training times are varied between less than a second to few seconds. Whilst, the testing elapsed times are very remarkable where testing any RBF-NN shows a significantly short time of less than 0.02 second.

References:

Al-Hussein, M. A. S., et al. (2022a). Applying the deoxyribonucleic acid (DNA) for people identification. *Journal of Harbin Institute of Technology*, *54*(8). https://doi.org/10.11720/JHIT.54082022.13

Al-Hussein, M. A. S., et al. (2022b). *Deoxyribonucleic Acid (DNA) for Individual Recognition*. Noor Publishing.

Al-Hussein, M. A. S., et al. (2022c). Employing deoxyribonucleic acid (DNA) for personal verification. *International Journal of Health Sciences*, *6*(S9).

Al-Nima, R. R. O. (2017). *Signal Processing and Machine Learning Techniques for Human Verification Based on Finger Textures*. Newcastle University.

Al-Nima, R. R. O., et al. (2023). An Artificial Intelligence Approach for Verifying Persons by Employing the Deoxyribonucleic Acid (DNA) Nucleotides. *Journal of Electrical and Computer Engineering*, *2023*.

Al-Nima, R. R. O., et al. (2024). Identifying deoxyribonucleic acids of individuals based on their chromosomes by proposing a special deep learning model. *Bulletin of Electrical Engineering and Informatics*, *13*(2).

Al-Nima, R. R. O., et al. (2014). A new approach to predicting physical biometrics from behavioural biometrics. *International Journal of Computer and Information Engineering*, *8*(11), 2001–2006.

Ephraim, N. O., et al. (2022). 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*, *1*(1), 48–71.

Fausett, L. V. (1994). *Fundamentals of neural networks: architectures, algorithms and applications*. Prentice-Hall, Inc.

Jacob, I. J., et al. (2021). Biometric template security using DNA codec based transformation. *Multimedia Tools and Applications*, *80*, 7547–7566.

Katsanis, S. H., et al. (2021). US adult perspectives on facial images, DNA, and other biometrics. *IEEE Transactions on Technology and Society*, *3*(1), 9–15.

Kou, J., et al. (2010). The incremental probabilistic neural network. *2010 Sixth International Conference on Natural Computation*, *3*, 1330–1333.

Liu, J., et al. (2022). From distributed machine learning to federated learning: A survey. *Knowledge and Information Systems*, *64*(4), 885–917.

Ou, D., et al. (2024). An impedance labeling free electrochemical aptamer sensor based on tetrahedral DNA nanostructures for doxorubicin determination. *Microchimica Acta*, *191*(2), 94.

Pandkar, N., et al. (2022). 3D Facial Biometric Verification Using a DNA Sample for Law Enforcement Applications. *2022 IEEE/WIC/ACM International Joint Conference on Web Intelligence and Intelligent*

Agent Technology (WI-IAT), 715–722.

Professional, Cleveland Clinic medical. (2022). *DNA, Genes & Chromosomes*. https://my.clevelandclinic.org/health/body/23064-dna-genes--chromosomes

Roewer, L. (2013). DNA fingerprinting in forensics: past, present, future. *Investigative Genetics*, *4*.

SAĞIROĞLU, Ş., & ÖZKAYA, N. (2009). An intelligent face features generation system from fingerprints. *Turkish Journal of Electrical Engineering and Computer Sciences*, *17*(2), 183–203.

Sakr, A. S., et al. (2022). Cancelable ECG biometric based on combination of deep transfer learning with DNA and amino acid approaches for human authentication. *Information Sciences*, *585*, 127–143.

Salama, G. M., et al. (2023). Secure biometric systems based on bio-signals and DNA encryption of optical spectrograms. *Optics Express*, *31*(3), 3927–3944.

Sepers, B., et al. (2023). Developmental stress does not induce genome-wide DNA methylation changes in wild great tit (Parus major) nestlings. *Molecular Ecology*, *32*(14), 3960–3974.

Zhang, X. (2022). *Next is Now: Takeaways from Grace Hopper Celebration 2022*. Retrieved June 02, 2024, from https://security.salesforce.com/pt/blog/next-is-now-takeaways-from-grace-hopper-celebration-2022