Research Article

Donya A. Khalid* and Nasser Nafea

A deep neural network model for paternity testing based on 15-loci STR for Iraqi families

https://doi.org/10.1515/jisys-2023-0041
received March 26, 2023; accepted June 04, 2023

Abstract: Paternity testing using a deoxyribose nucleic acid (DNA) profile is an essential branch of forensic science, and DNA short tandem repeat (STR) is usually used for this purpose. Nowadays, in third-world countries, conventional kinship analysis techniques used in forensic investigations result in inadequate accuracy measurements, especially when dealing with large human STR datasets; they compare human profiles manually so that the number of samples is limited due to the required human efforts and time consumption. By utilizing automation made possible by AI, forensic investigations are conducted more efficiently, saving both time conception and cost. In this article, we propose a new algorithm for predicting paternity based on the 15-loci STR-DNA datasets using a deep neural network (DNN), where comparisons among many human profiles are held regardless of the limitation of the number of samples. For the purpose of paternity testing, familial data are artificially created based on the real data of individual Iraqi people from Al-Najaf province. Such action helps to overcome the shortage of Iraqi data due to restricted policies and the secrecy of familial datasets. About 53,530 datasets are used in the proposed DNN model for the purpose of training and testing. The Keras library based on Python is used to implement and test the proposed system, as well as the confusion matrix and receiver operating characteristic curve for system evaluation. The system shows excellent accuracy of 99.6% in paternity tests, which is the highest accuracy compared to the existing works. This system shows a good attempt at testing paternity based on a technique of artificial intelligence.

Keywords: DNA, STR, Paternity testing, artificial intelligence, DNN

1 Introduction

Deoxyribose nucleic acid (DNA) is the source code of all life on earth; it contains all the information needed to create and maintain it. DNA is the fundamental component that forms a whole section of a human, and it contains information that is particular to each person and is passed down through the generations [1]. Its main building blocks are molecules called nucleotides. They are composed of a nucleobase attached to a sugar called deoxyribose, which is attached to a phosphate group. Four bases are used in a nucleotide, namely adenine (A), guanine (G), cytosine (C), or thymine (T). These can be regarded as the four-letter alphabet in which the DNA source code is written [2,3].

The human genome consists of repeating DNA strands with different pattern sizes [4,5]. DNA with small repetition units (approximately between 2 and 6 base pairs) called short tandem repeats (STRs) is incredibly useful for identifying the father in paternity testing, finding the missing person's identity, and identifying victims in mass disasters [6–9].

*** Corresponding author: Donya A. Khalid,** College of Information Engineering, Al-Nahrain University, Baghdad 10011, Iraq, e-mail: donyaabbass5@gmail.com
Nasser Nafea: College of Information Engineering, Al-Nahrain University, Baghdad 10011, Iraq, e-mail: nassrnafea@gmail.com

Numerous organizations throughout the world have conducted medical and ethnogenesis searches that have publicly confirmed and documented the polymorphic nature of STRs and their usage for human identification [10–12].

A child in a family has a chromosome where one allele is passed down from the father and the other allele is received from the mother, so the child's DNA profile combines the DNA profiles of both parents. A person's DNA can be used to determine whether they are a father's or mother's child if they are identical, accounting for around 50% of the total DNA because 50% of the DNA is directly passed down from one parent to the child [13].

Recently, deep learning has seen widespread application across many different areas of study [14–19]; it has demonstrated significantly enhanced performance in complex regression and classification problems, where the complicated structure in the high-dimensional data is hard to discover with traditional machine learning algorithms [20]. In biology, the use of deep learning to predict the function and structure of genomic features is becoming more popular and has also played an important role in forensic science and criminal investigation, like identifying individuals and matching traces found at crime scenes with a suspect's profile [21].

Busia et al. [22] introduced a deep learning approach to pattern recognition for short DNA sequences. Miyake et al. [23] presented a method to classify alleles of human leukocyte antigen-A DNAs. Anggreainy et al. [24] used an approach to find the similarity of human DNA profiles based on fuzzy, while Sino and Sears [25] showed how to use AI to extend the Elston–Stewart algorithm and arrive at a system that can link people to their family trees based on their likelihood of being related to each other.

This study introduces a deep learning technique for identifying an unknown person by finding his or her correct parents. This will be done at first by artificially creating a familial dataset and increasing the number of samples, and then applying a deep neural network (DNN) to the generated dataset to determine paternity. There are no studies that employed the same target using deep learning techniques, but based on the survey there are two studies that have the idea for DNA matching using artificial intelligence. The first one was proposed by Anggreainy et al. [24] by using fuzzy inference for STR-DNA matching for 13 loci. Second, Siino and Sears [25] used gradient descent logistic regression for kinship analysis based on 13 loci.

The research proposes paternity testing based on artificial intelligence instead of manual matching. The conventional method uses a genetic analyzer to extract the human profile for the father, mother, and child from real DNA tissue; then, a comparison will be made manually for 15 loci. A maximum of three to four profiles can be compared at the same time due to the required human efforts in comparing 15 loci with two alleles at the same locus for one profile in the manual matching. Introducing an artificially intelligent system in paternity testing reduces both effort and time consumption while increasing efficiency. The proposed system can also handle the paternity test regardless of the number of samples; this article assumes paternity testing using DNN, which facilitates paternity testing for 53 families with 10 children, in which 53,530 samples are tested at the same time.

This article has many contributions, which can be summarized as follows:

1. To overcome the restriction on gating familial datasets, familial datasets were built artificially based on real datasets obtained from Namaa et al. [26] in order to increase the data size for the learning process.
2. Overcome the number of sample limitations (up to three samples) in conventional paternity testing. The proposed system can handle paternity testing regardless of the number of tested samples.
3. The proposed system works independently of the STR loci.
4. The system accuracy reaches 0.996.
5. The proposed system opens the way for paternity testing for insufficient DNA datasets (missing STR) in disaster victim identification. We are now working on developing the algorithm to handle this task.
6. The system offers a good opportunity to create a familial database that may be useful in cases of missing person identification.

This article is divided into six sections. Section 1 provides a general overview. Section 2 provides a brief description of forensic identification and the various DNA techniques that are used in forensic genetics. Section 3 gives an overview of paternity testing using STR. Section 4 describes deep learning. Section 5 explains

the proposed system model in terms of data generation and deep learning, and Section 6 illustrates the main results and their discussions.

2 Forensic human identification

In forensic identification, the characteristics of an evidentiary specimen are compared to those of a reference sample to determine the specimen's origin. DNA markers detected in biological samples like blood and saliva are frequently used in human identification [1]. Standard DNA markers used in forensics include the following.

Restriction fragment length polymorphism (RFLP): it is a DNA fingerprinting technique based on detecting DNA fragments of different lengths. It was the first technology introduced for forensics [27].

Mitochondrial DNA (mtDNA) analysis: each person in the same maternal line has the same mtDNA. Comparing the mtDNA of a victim's brother's maternal lineage can help identify a missing person or determine a person's identity [28].

Y-Chromosome analysis: Y-chromosome analysis is conducted for forensic purposes and in the study of human evolution [29].

Single-nucleotide polymorphism (SNP) typing: DNA sequence variations called SNPs occur whenever a single base (A, T, C, or G) in the genome is changed [30].

STR: STR is an alternative to the RFLP technique that has gained widespread use. The Federal Bureau of Investigation has advocated this technique as the gold standard for DNA profiling [31].

The gold standard among forensic science approaches is human identification using genetic profiles acquired from DNA polymorphisms (STR). A series of duplicates of short DNA segments produce polymorphisms called STRs (ranging from 2 to 6 base pairs). The biological samples taken from crime scenes are processed to create a DNA profile that will be compared to the profile of the suspect, helping the investigation establish a link between the criminal and the crime scene or even ruling out suspects [32]. Police crime laboratories can be used to verify paternity and match evidence with suspects to assist determine guilt or innocence [33].

3 Paternity testing

Due to STR's variability, heritability, and occurrence in combinations that are exclusive to each individual, this marker is ideal for use in forensic science. When it comes to forensic genetics, paternity testing is among the most popular problems. Their significance stems from the fact that they are valuable genetic markers, providing significant statistical capacity for individualization and discrimination throughout many forensic and judicial regulations [34].

Paternity tests are the most widely used of all the genetic tests that have been developed with new technologies for the analysis of DNA. These tests involve comparing the STR of the DNA samples collected from two individuals, usually a child and a putative father, and can confirm a paternal relationship with a very high degree of accuracy [26].

The loci are the regions on chromosomes that differ in length between persons and are utilized as the subject of analysis in a paternity DNA test; the value of STR employed in the proposed forensic test is the Combined DNA Index System. The sequences of these loci are presented in Table 1.

Table 1: Sequences of 15-loci STR used [26]

D8S1179	D21S11	D7S820	CSF1PO	D3S1358	TH01	D13S317	D16S539	D2S1338	D19S433	vWA	TPOX	D18S51	D5S818	FGA

As mentioned previously, every human inherits half of his or her genetic material from the father and the other half from the mother, in each locus. There are two separate numbers represented as a pair, one from the mother and another from the father, and these 15-loci pairs may have the same value (homozygosity) or different values (heterozygosity). Scientifically, these numbers are referred to as alleles.

Based on the above, the following is an example to explain the idea of DNA human paternity testing, where a child's DNA result shows up at the CSF1PO locus (9, 15). The same locus for the mother will show alleles with numbers (9, 14). In other words, the child got 9 from the mother and 21 from the father; the putative father is the child's biological father, and the allele number for him must be confirmed (11, 15). Figure 1 identifies all of the possible STR allele combinations that the child could have inherited at this particular locus.

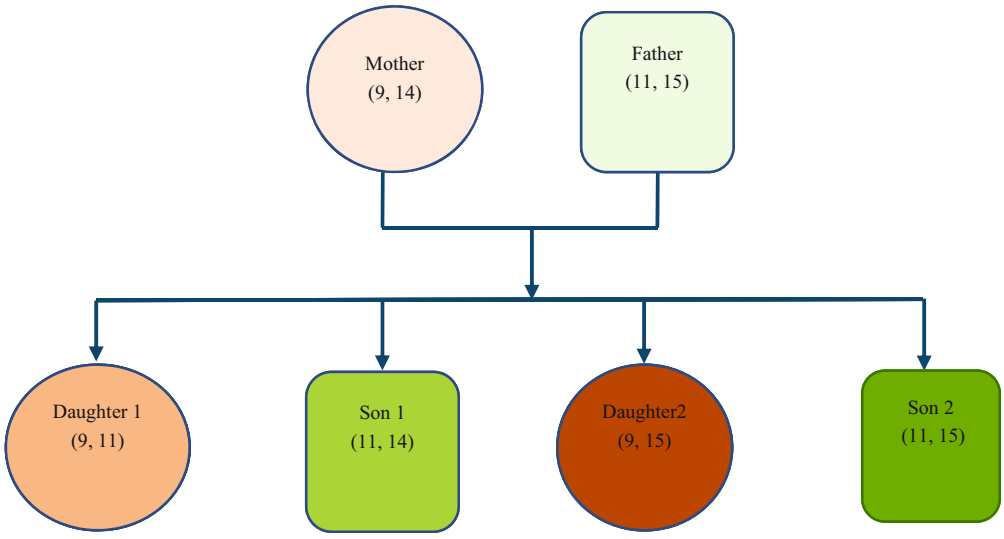

Figure 1: Allele possibilities for the current locus.

4 Deep learning for DNA

Deep learning is the emerging generation of artificial intelligence techniques, specifically in machine learning [25]. DNA fingerprints are currently utilized to validate identity-based artificial intelligence [35].

The information-processing patterns present in the human brain serve as inspiration for deep learning. It does not need any rules created by humans to work; instead, it uses a huge amount of data to map the input to particular labels. Artificial neural networks (ANNs) are used in deep learning to create multiple layers of algorithms, each of which provides a distinct interpretation of the data fed to it [36].

Due to its ability to automatically extract features from massive volumes of data and automatically learn meaningful representations, deep learning has quickly become one of the most prominent machine learning approaches today. It is used in a variety of scientific disciplines, including robotics, image processing, and voice analysis. Deep learning has also found use in the field of biology as a result of the recent growth in biological "omics" data [37]. The concept of successive layers of representations is what the term "deep" in "deep learning" refers to rather than any kind of deeper understanding that the technique can acquire. The multilayer architecture of DNN mimics the structure in visual neuroscience and can transform the data representation into an increasingly abstract form via nonlinear modules [38]. It turns out to be surprisingly successful in learning the non-linear input–output mapping with both increased selectivity and the invariance of the representation. There is a direct correlation between increasing the number of hidden layers in a network and the difficulty of its training in terms of both time and the number of resources needed to complete the training. The main benefit of deep learning is its autonomous feature extraction capability with great selectivity and invariance [39]. The feature extraction procedure is essential for reducing the

number of processing resources required without losing significant or relevant data and accelerating the learning process [38]. An input layer, several hidden layers, and an output layer comprise the fundamental architecture of DNNs.

5 The proposed methodology

This work includes two main parts: the first is data creation and the second is the implementation of the proposed deep learning model.

5.1 Data creation

Data preparation is the first step in this system; 15 loci will be used to match parents and child alleles to confirm paternity. The proposed model uses 106 unrelated Iraqi person datasets from Al-Najaf province [26]; two persons (male and female) loci are used as parents to create 53 familial datasets; each familial set includes a mother, father, and five children. Then, five false children that do not belong to the same family are appended randomly for each family, which results in 530 data rows for 53 parents. This step aims to make the proposed DNN system learn the relationship between the family (mother and father) and their corresponding children. The family number from 0 to 52 is assigned for each family. The flowchart in Figure 2 represents the data creation steps.

Tables 2 and 3 show the representation of the resultant datasets for one family (father, mother, and ten children). In all, 15 loci of STR, and each locus has two alleles (A1 and A2) for father, mother, and child.

The previous representation of the datasets is converted into a form as shown in Table 4. For each familial set, ten rows will be created. The new dataset has five columns for father alleles, mother alleles, child alleles, family numbers (0–52), and labels. The label (1) indicates whether the child truly belongs to the family or does not belong to the family (0).

To generalize the model, the system also presented a method to shuffle these STR 15-loci sequences (shown in Figure 3). Each locus in all datasets will be shuffled randomly. During shuffling the locus positions, sequencing is preserved in all family members, i.e., if the CSF1PO locus on the father dataset is shuffled to a certain position, the same locus in the mother and child datasets will be shuffled in the same manner. The purpose of this process is to make the DNN system work independently on positions at each locus during the training. The shuffling process is applied to 53 families. With 11 shuffling iterations, the final dataset size is 53,530 samples.

5.2 DNA deep learning model

Based on the problem described in Section 3 of 15 loci with 2 alleles per locus and 64 indexing for each allele, a DNN is proposed as a binary classifier for family identification; it assumes testing paternity based on the generated familial STR dataset. The system finds the correct family for a person using a dataset consisting of 53,530 individuals for training and testing purposes. The proposed model applies embedding, flattening, and dense layers for each family member's dataset for feature extraction. Then, a dense layer is applied again to the data after concatenation. Figure 4 illustrates the proposed deep learning model structure, and Table 5 lists the main specification of the model.

The proposed system is implemented using the Python-based Keras library. Colab online platform is used as an environment for the model implementation of training and testing phases.

First, a dataset of 53,530 individuals is divided into training and testing sets: 80% of the data is used for training, which equals 42,824 samples, and 20% is used for testing, with a size equal to 1,070 samples.

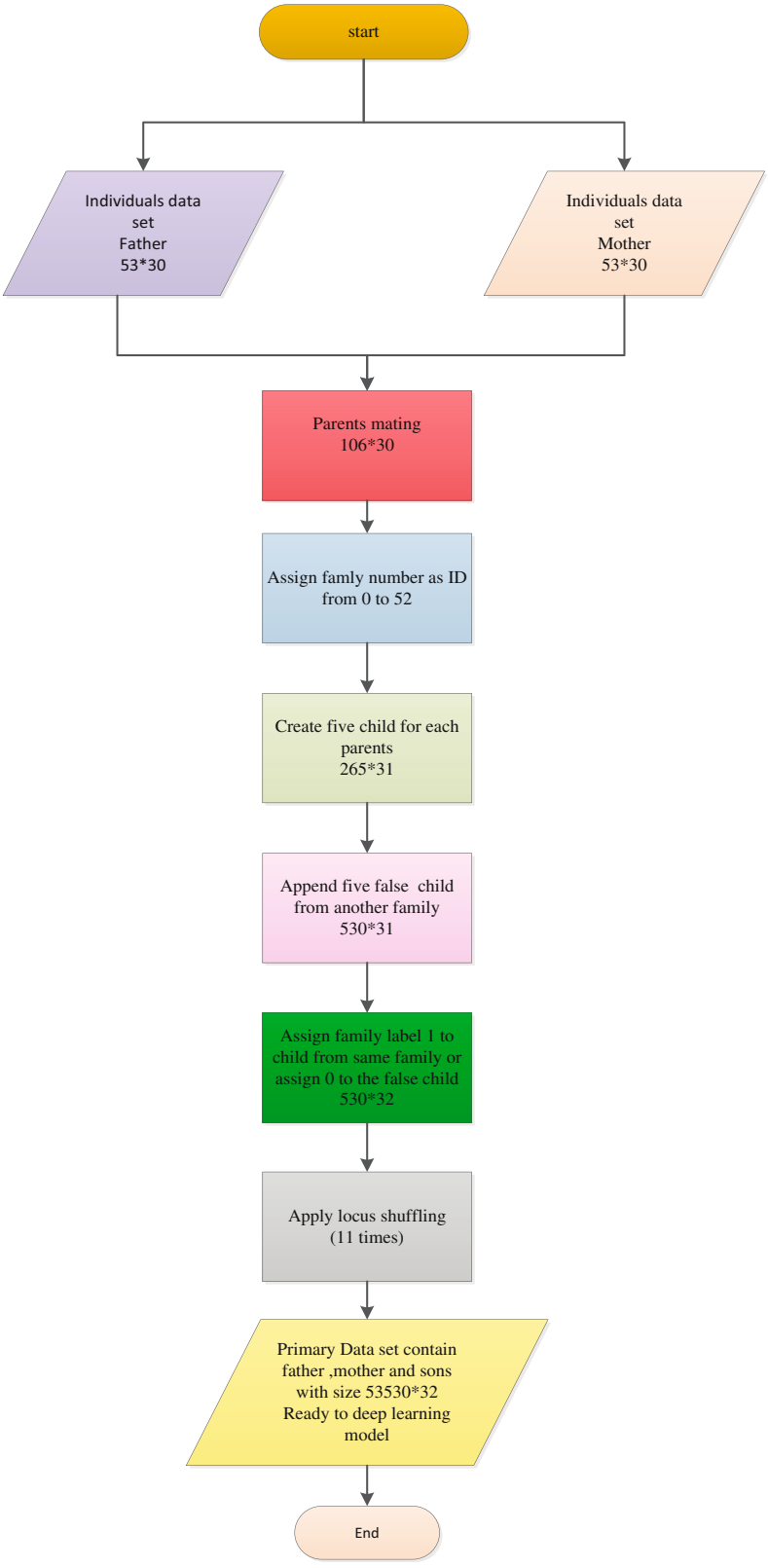

Figure 2: Flowchart for data creation.

Table 2: STR (first-7 loci) for parents and five possible children

Sample	D8S1179		D21S11		D7S820		CSF1PO		D3S1358		TH01		D13S317	
	A1	A2	A1	A2	A1	A2	A1	A1	A1	A2	A1	A2	A1	A2
Mother	12	14	29	32.2	10	11	12	12	15	18	6	6	12	13
Father	14	14	28	31	10	12	9	9	15	18	7	9.3	9	11
Son 1	14	14	28	32.2	10	12	11	11	15	18	6	9.3	11	12
Son 2	12	14	28	32.2	10	10	11	11	15	18	6	7	11	13
Son 3	12	14	28	32.2	10	11	9	9	18	18	6	7	9	12
Son 4	14	14	28	29	11	12	11	11	18	18	6	9.3	11	13
Son 5	12	14	29	31	10	10	11	11	15	18	6	7	11	13

Table 3: STR (last-8 loci) for parents and five possible children

D16S539		D2S1338		D19S433		vWA		TPOX		D18S51		D5S818		FGA		Family ID
A1	A2	A1	A2	A1	A2	A1	A2	A1	A2	A1	A2	A1	A2	A1	A2	label
9	11	18	19	13	15.2	17	17	9	11	13	14	11	12	22	25	0
12	13	18	19	14	15.2	17	19	10	11	12	18	12	13	22	24	0
9	12	18	19	15.2	15.2	17	19	10	11	14	18	12	12	22	25	0
9	13	18	18	15.2	15.2	17	17	10	11	12	13	11	13	22	25	0
11	12	19	19	13	14	17	19	9	10	12	14	11	13	22	24	0
11	13	18	18	13	15.2	17	17	9	10	13	18	11	12	22	22	0
11	13	18	19	13	14	17	17	9	10	14	18	12	12	22	22	0

Table 4: Dataset of the first family of 53

Father set	Mother set	Child set	Family number	Label
Father 1	Mother 1	Child 1	0	1
Father 1	Mother 1	Child 2	0	1
Father 1	Mother 1	Child 3	0	1
Father 1	Mother 1	Child 4	0	1
Father 1	Mother 1	Child 5	0	1
Father 1	Mother 1	Child 6	0	0
Father 1	Mother 1	Child 7	0	0
Father 1	Mother 1	Child 8	0	0
Father 1	Mother 1	Child 9	0	0
Father 1	Mother 1	Child 10	0	0

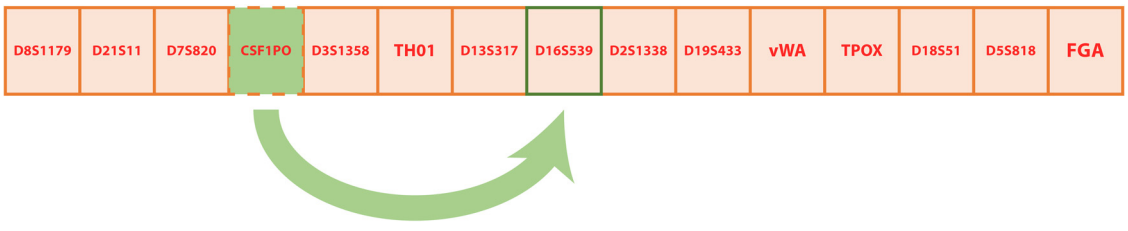

Figure 3: Locus shuffling.

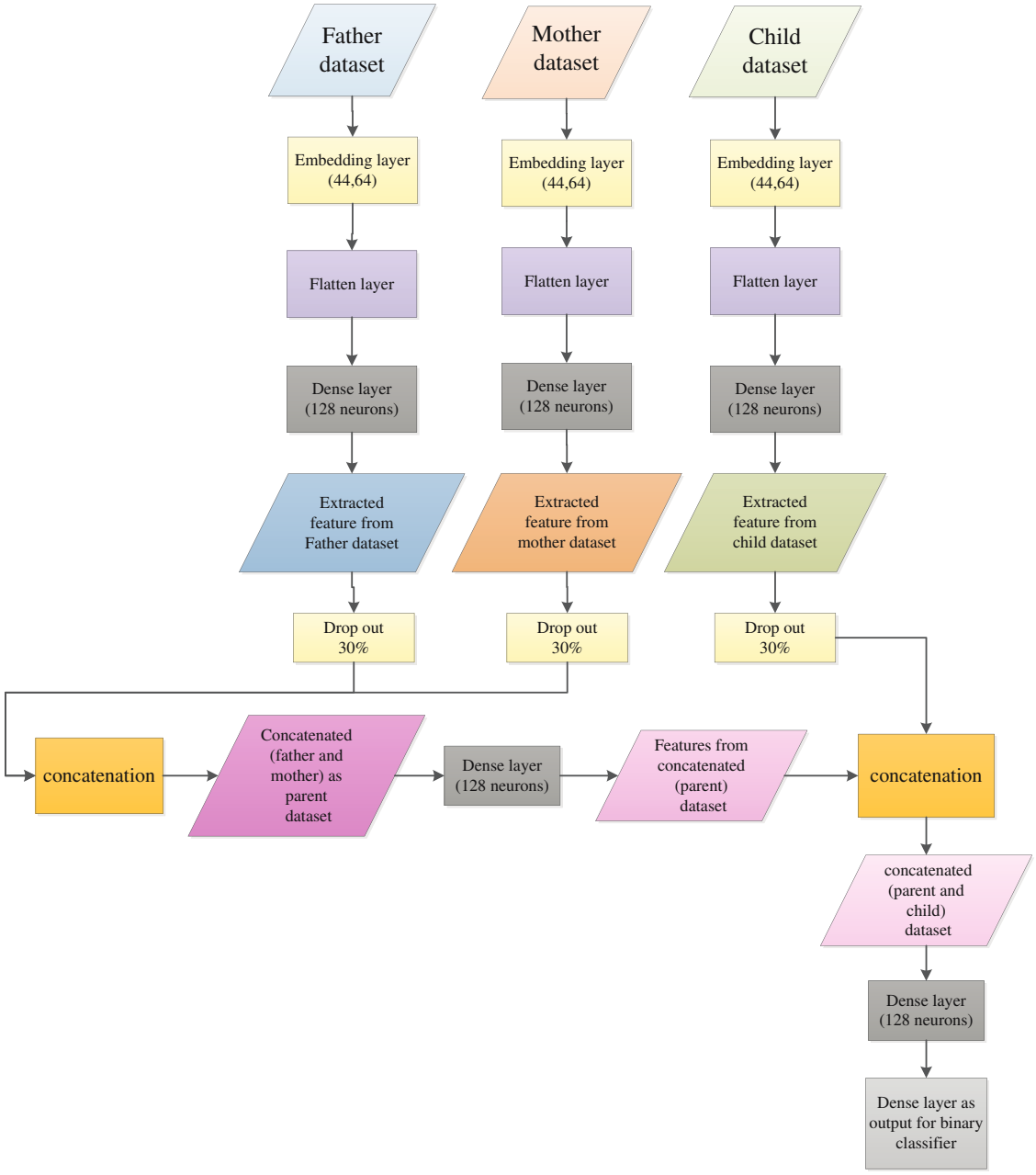

Figure 4: Proposed DNN model.

With the training dataset, the data are divided into equal labels. Half of the dataset has a label (1), which indicates true children, and the other half has a label (0), which indicates false children. Figure 5 shows the histogram of the training dataset.

The DNN model mainly includes three types of Keras layers and concatenation process layers, which are described in the following subsections.

5.2.1 Embedding layer

At first, three embedding layers are applied separately for each dataset (father, mother, and child) to convert each allele value in all datasets into a fixed-length vector of 1×64. The resultant vector is a dense one with real

Table 5: DNN model specifications

Specifications	Value
Embedding layer	3
Flatten layer	3
Number of dense layers	6
Number of neurons in each dense layer	128 neurons, except the last dense layer is the output layer with a single neuron as a binary classifier for the case of correct family or not
Number of the dropout layers	3
The percentage of dropouts	30%
The activation function for the hidden layers	Leaky-rectifier linear activation function
The activation function for the output layer	Sigmoid activation function
Dataset size	53,530 80% of data for the training phase: 42,824 20% of data for the testing phase:1,070
The number of epochs	10
Batch size	32
Applied optimizer	ADAM
The used loss function	Binary cross-entropy

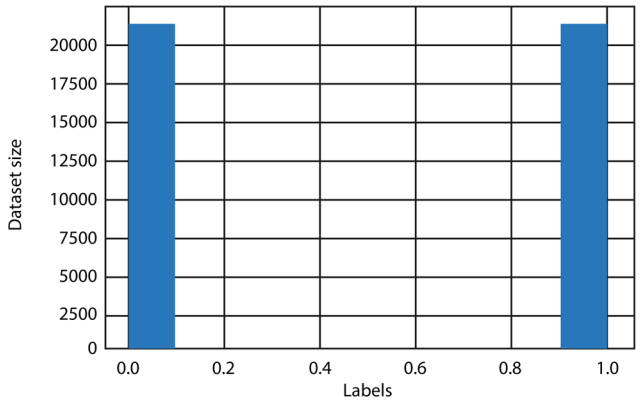

Figure 5: Histogram of the training dataset.

values instead of just 0's and 1's in the case of one-hot encoding. The fixed length of vectors helps in representing alleles in the best way, along with reduced dimensions for better performance in deep learning. The embedding layer is initialized with random weights and will learn an embedding for all of the alleles in the training dataset. The input to this layer is 44, in which there are 44 variant values in the used 15-STR dataset, and the output from this layer is 44 × 64.

5.2.2 Flatten layer

Flattening is applied three times for each dataset after embedding to convert each dataset into a one-dimensional array to prepare data for the next layer.

5.2.3 Dense layer

Six dense layers are used separately, as shown in Figure 5:

a. A dense layer with 128 nodes is applied for each embedded dataset (father, mother, and child). At each node in the dense layer, the weighted sum of inputs is calculated and fed to the leaky-rectifier linear activation function. The mathematical representation of the Leaky_ReLU function is

$$F(x) = \max(ax, x). \tag{1}$$

In addition, two dense layers are applied to the concatenated datasets.

b. One dense layer with a single node is used in the system as output that indicates a binary classifier (either 1 in the case of a child belonging to the putative family or 0 in the case of not). The sigmoid activation function is applied in the output node. The mathematical representation of the sigmoid activation function is

$$F(x) = \frac{1}{1 + e^{-x}}, \tag{2}$$

where x is the input to the neuron.

5.2.4 Dropout layer

A dropout of 30% is performed for the dense layers of the father, mother, and son datasets to overcome overfitting occurrences.

5.2.5 Concatenation

The extracted features from both the father dataset and the mother dataset are concatenated together to generate a single dataset, which will then be concatenated again with the child dataset as the final dataset to be deeply learned later for paternity.

5.3 System training

The system uses 80% of the total dataset for training and 20% for validation. The system trains in 10 epochs, with accuracy metrics and binary cross-entropy for loss. The mathematical representation of accuracy is

$$\text{Accuracy} = \frac{\text{Number of correct predictions}}{\text{Total number of predictions}}. \tag{3}$$

The mathematical representation of binary cross entropy is

$$J(z) = \frac{1}{N}\sum_{I=1}^{N} y_i \log[f(x_i, z)] + (1 - y_i)\log[1 - f(x_i, Z)], \tag{4}$$

where y_i is the predicted output and $[f(x_i, z)]$ is the actual value.

For each training phase through ten epochs, the best model with minimum loss and the highest accuracy will be saved. The Adam optimizer is used for optimum weight values.

6 Results and discussion

The results are obtained using Google-Colab, which is Python-based; all required libraries are imported to build the model. Table 6 summarizes the DNN model parameters.

Table 7 shows the model training and testing results (loss, accuracy, validation loss, and validation accuracy) over ten epochs. The model gives the best results without overfitting, with minimum loss = 0.0244, val_loss = 0.0094, and higher accuracy = 0.9917, val_acc = 0.9970.

Table 6: DNA–DNN model parameters

Layer (type)	Output shape	Parameters
Embedding (embedding)	Multiple	2,816
Dense (dense)	Multiple	245,888
Dense_1 (dense)	Multiple	245,888
Dense_2 (dense)	Multiple	32,896
Dense_3 (dense)	Multiple	245,888
Dense_4 (dense)	Multiple	32,896
Dense_5 (dense)	Multiple	129
Total parameters: 806,401		
Trainable parameters: 806,401		
Non-trainable parameters: 0		

Table 7: Model training/testing loss and accuracy over ten epochs

Epoch	Loss	Accuracy	Validation loss	Validation accuracy
	Training		Testing	
1	0.5360	0.6995	0.3853	0.8274
2	0.3178	0.8567	0.2622	0.9402
3	0.1763	0.9293	0.1419	0.9402
4	0.0881	0.9667	0.0584	0.9787
5	0.0517	0.9813	0.0333	0.9883
6	0.0389	0.9863	0.0271	0.9902
7	0.0276	0.9905	0.0247	0.9907
8	0.0271	0.9906	0.0251	0.9903
9	0.0244	0.9917	0.0094	0.9970
10	0.0213	0.9929	0.0109	0.9962

Figure 6 shows the performance of the DNN model for both the training and testing phases over ten epochs. From the figure, it seems that the DNN model performs well in terms of lower loss values for both the training and validation sets, higher accuracy, and the absence of overfitting due to the implementation of the dropout regularization technique.

To assess the performance of the binary classification model, a confusion matrix is also plotted (Figure 7) to explore the true positive (TP), true negative (TN), false positive (FP), and false negative (FN) values of model predictions.

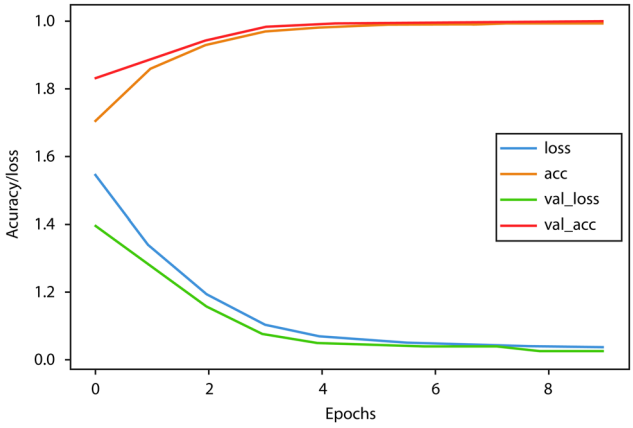

Figure 6: Accuracy and loss.

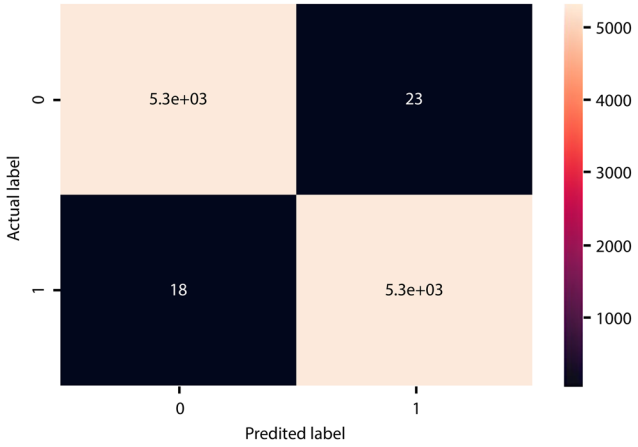

Figure 7: Confusion matrix.

The system shows 41 total false predictions for both FP (23) and FN (18), while the true predictions are around 10,600. The TP and TN values are mostly equal due to equally distributed tested datasets. The receiver operating characteristic (ROC) is illustrated to measure system performance as a binary classifier, and it visualizes the trade-off between true positive rate (TPR) and false positive rate (FPR). The TPR is plotted along the *Y*-axis, and the FPR is plotted along the *X*-axis.

TPR or sensitivity can be represented as

$$\text{TPR} = \frac{\text{TP}}{\text{TP} + \text{FN}}. \tag{5}$$

FPR can be calculated using the following equation:

$$\text{FPR} = 1 - \text{Specifity} = \frac{\text{FP}}{\text{FP} + \text{TN}}, \tag{6}$$

where specificity is the TN rate:

$$\text{TNR} = \frac{\text{TN}}{\text{TN} + \text{FP}}. \tag{7}$$

The DNN paternity testing sensitivity is equal to 0.99, specificity is equal to 0.997, and the *F*1-score is equal to 0.997.

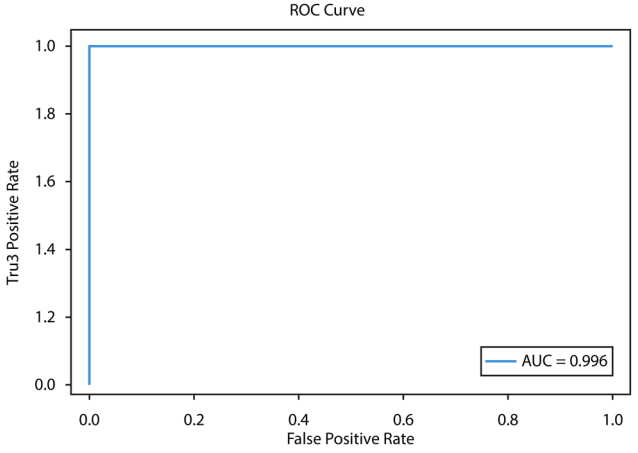

Figure 8: The ROC curve.

The higher the TPR and lower the FPR, the better. The ROC shows good results, as shown in Figure 8, with an area under the curve equal to 0.996. The ROC curve tends to be on the top-left side, which indicates the best classifier characteristics.

Table 8 shows a comparison between the related works and the proposed model in terms of the number of profiles used, number of loci, applied methods, and reported accuracy. It shows that the proposed DNN model has the highest accuracy among them, with 53,530 profiles.

Table 8: Comparison with other artificial intelligence works

Research	Number of profiles	Accuracy (%)	Number of loci	Applied method
Anggreainy et al. [24]	100	80	16	Fuzzy
Siino and Sears [25]	152,146	95	13	Gradient descent logistic regression
Proposed work	53,530	99.6	15	DNN

The proposed system has a limitation in performing paternity testing based on the availability of both the father's and mother's DNA; the system cannot work if one parent's DNA data are missing.

7 Conclusions

DNA paternity testing is a crucial area of forensic research. A paternity test often uses DNA-STR profiles, where a child's profile is compared with the parent's profiles to find out if the child correctly belongs to this family or not. This comparison is held manually and can be solved if three to four profiles are compared but the problem occurs when the number of profiles to be compared increases, which requires human effort and time. In this article, a DNN algorithm is suggested for identifying paternity based on the 15-loci STR-DNA Iraqi dataset. The proposed artificial DNA paternity testing is applied using the DNN technique as a binary classifier where many profiles can be compared at the same time. In this system, 53,530 profiles are used for the training and testing phases. The system offers a good opportunity to create a familial database that can be used as a reference in other cases, like missing person identification.

This work has been carried out using the Python library (Keras layers, Numpy, Pandas, Matplotlib, Sklearn, Pickle, Tqdm, and Seaborn) to implement and test the proposed system. The proposed algorithm is applied to an artificially extended dataset due to the insufficient real dataset available.

The overall system performance gives good results in terms of loss and accuracy, and the dropout technique is applied to avoid overfitting. A confusion matrix shows that the values of FP = 23 and FN = 18 are very low compared to the total predicted values, and the accuracy reaches 0.997. Also, the ROC chart showed that the curve tends to be on the top-left side, which indicates good classifier characteristics. The proposed system has a limitation in performing paternity testing based on the availability of both the father's and mother's DNA. In the future, DNN will be applied to the mystery DNA-STR sequence that occurred in mass causality incidents where the profile is not complete. Also, another deep learning technique can be applied, like a gated recurrent neural network.

Funding information: The authors state no funding involved.

Author contributions: Donya A. Khalid did the programs with supervision of Nasser Nafea. Donya wrote the main manuscript and figures. Nasser Nafea reviewed the manuscript.

Conflict of interest: Authors state no conflict of interest.

Data availability statement: The data that support the findings of this study are available from the corresponding suthor, [Donya A. Khalid], upon reasonable request.

References

[1] Butler JM. The future of forensic DNA analysis. Philos Trans R Soc B: Biol Sci. Aug. 2015;370(1674):20140252. doi: 10.1098/rstb.2014.0252.

[2] Phoebe Chen Y-P. Bioinformatics technologies. Springer; 2005.

[3] Clercq GDe. Deep learning for classification of DNA functional sequences. Master of Science in Bioinformatics; 2018–2019.

[4] Kimpton CP, Gill P, Walton A, Urquhart A, Millican ES, Adams M. Automated DNA profiling employing multiplex amplification of short tandem repeat loci. Genome Res. Aug. 1993;3(1):13–22. doi: 10.1101/gr.3.1.13.

[5] Keerti A, Ninave S. DNA fingerprinting: Use of autosomal short tandem repeats in forensic DNA typing. Cureus. Oct. 2022;14(10):e30210. doi: 10.7759/cureus.30210.

[6] Ruitberg CM, Reeder DJ, Butler JM. STRBase: a short tandem repeat DNA database for the human identity testing community; 2001. http://www.cstl.nist.gov/biotech/strbase/.

[7] Nwawuba Stanley U, Mohammed Khadija A, Bukola AT, Omusi Precious I, Ayevbuomwan Davidson E. Forensic DNA profiling: Autosomal short tandem repeat as a prominent marker in crime investigation. Malays J Med Sci. Jul. 2020;27(4):22–35. doi: 10.21315/mjms2020.27.4.3.

[8] Niedzwiecki E, Debus-Sherrill S, Field MB, Michael SD-S, Field B. Understanding familial DNA searching: Coming to a consensus on terminology understanding famlial DNA searching: Coming to a consensus on terminology study of familial DNA searching policies and practices. NW, Washington, D.C.: National Institute of Justice, research, development and evaluation agency of the U.S. Department of Justice. 2016.

[9] Budowle B, Bieber FR, Eisenberg AJ. Forensic aspects of mass disasters: Strategic considerations for DNA-based human identification. Leg Med. Jul. 2005;7(4):230–43. doi: 10.1016/j.legalmed.2005.01.001.

[10] López-Flores I, Garrido-Ramos MA. The repetitive DNA content of eukaryotic genomes. Repetitive DNA. 2012;7:1–28. doi: 10.1159/000337118.

[11] Yasin SR, Hamad MM, Elkarmi AZ, Jaran AS. African Jordanian population genetic database on fifteen short tandem repeat genetic loci. Croat Med J . Aug. 2005;46(4):587–92.

[12] Al-Zubaidi MM, Ibrahem HK, Ameen RS, Ameen B. Allele frequencies of 15 Autosomal STR loci in Some of Iraqi population. Iraqi J Sci. 2022;63(6):2434–43. doi: 10.24996/ijs.2022.63.6.10.

[13] Lamb ME, Sutton-Smith B, Sutton-Smith B, Lamb E. Sibling relationships their nature and significance across the lifespan. Hove, East Sussex, United Kingdom: Psychology Press; 1982.

[14] Zhang A, Lipton ZC, Li MU, Smola AJ. Dive into Deep Learning. 1st edn. United Kingdom: Cambridge University Press & Cambridge Assessment; 2022.

[15] Salman AO, Geman O. Evaluating three machine learning classification methods for effective COVID-19 diagnosis. Int J Mathematics, Statistics, Computer Sci. Jan. 2023;1:1–14. doi: 10.59543/ijmscs.v1i.7693.

[16] Yang A, Zhang W, Wang J, Yang K, Han Y, Zhang L. Review on the application of machine learning algorithms in the sequence data mining of DNA. Front Bioeng Biotechnol. Sep. 04, 2020;8:1032. doi: 10.3389/fbioe.2020.01032 Frontiers Media S.A

[17] Li Y, Huang C, Ding L, Li Z, Pan Y, Gao X. Deep learning in bioinformatics: Introduction, application, and perspective in the big data era. Methods. Aug. 2019;166:4–21. doi: 10.1016/j.ymeth.2019.04.008.

[18] Begum S, Sarkar R, Chakraborty D, Maulik U. Identification of biomarker on biological and gene expression data using fuzzy preference based rough set. J Intell Syst. Jul. 2020;30(1):130–41. doi: 10.1515/jisys-2019-0034.

[19] Arif ZH, Cengiz K. Severity classification for COVID-19 infections based on lasso-logistic regression model. Int J Mathematics, Statistics, Computer Sci. Apr. 2023;1:25–32. doi: 10.59543/ijmscs.v1i.7715.

[20] Liu J, Li J, Wang H, Yan J. Application of deep learning in genomics. Sci China Life Sci. Dec. 01, 2020;63(12):1860–78. doi: 10.1007/s11427-020-1804-5 Science in China Press.

[21] Li H, Tian S, Li Y, Fang Q, Tan R, Pan Y, et al. Modern deep learning in bioinformatics. J Mol Cell Biol. Feb. 2021;12(11):823–27. doi: 10.1093/jmcb/mjaa030.

[22] Busia A, Dahl GE, Fannjiang C, Alexander DH, Dorfman E, Poplin R, et al. A deep learning approach to pattern recognition for short DNA sequences. bioRxiv. Jun. 2018. doi: 10.1101/353474.

[23] Miyake J, Kaneshita Y, Asatani S, Tagawa S, Niioka H, Hirano T. Graphical classification of DNA sequences of HLA alleles by deep learning. Hum Cell. Apr. 2018;31(2):102–5. doi: 10.1007/s13577-017-0194-6.

[24] Anggreainy MS, Widyanto MR, Widjaja B, Soedarsono N, Widodo PT. Family relation and STR-DNA matching using fuzzy inference. Int J Electr Comput Eng (IJECE). Apr. 2019;9(2):1335. doi: 10.11591/ijece.v9i2.pp1335-1345.

[25] Siino V, Sears C. Artificially intelligent scoring and classification engine for forensic identification. Forensic Sci Int Genet. Jan. 2020;44:102162. doi: 10.1016/j.fsigen.2019.102162.

[26] Namaa DS, AL-Zubaidi MM, AL-Rubai HK, Sabbah MA, Al-Janabi TY, Hameed, SN, et al. Comparison between allele frequencies of several Strs Loci in Najaf City of Iraq and middle Province in Iraqi population. Indian J Forensic Med & Toxicol. Oct. 2019;13(4):578. doi: 10.5958/0973-9130.2019.00353.0.

[27] Manjunath BC, Chandrashekar BR, Mahesh M, Vatchala Rani RM. DNA Profiling and forensic dentistry – A review of the recent concepts and trends,. J Forensic Leg Med. Jul. 2011;18(5):191–7. doi: 10.1016/j.jflm.2011.02.005.

[28] Nahar Sultana GN. Mitochondrial DNA and Methods for forensic identification. J Forensic Sci Crim Investig. May 2018;9:1. doi: 10.19080/jfsci.2018.09.555755.
[29] Roewer L. Y-chromosome short tandem repeats in forensics—Sexing, profiling, and matching male DNA. WIREs Forensic Sci. Jul. 2019;1(4). doi: 10.1002/wfs2.1336.
[30] Budowle B, van Daal A. Forensically relevant SNP classes. Biotechniques. Apr. 2008;44(5):603–10. doi: 10.2144/000112806
[31] Panneerchelvam S, Norazmi MN. Forensic DNA profiling and database. Malays J Med Sci. Jul. 2003;10(2):20–6.
[32] Wyner N, Barash M, McNevin D. Forensic autosomal short tandem repeats and their potential association with phenotype. Front Genet. Aug. 2020;11:1–7. doi: 10.3389/fgene.2020.00884.
[33] Marano LA, Fridman C. DNA phenotyping: current application in forensic science. Res Rep Forensic Med Sci. Feb. 2019;9:1–8. doi: 10.2147/RRFMS.S164090.
[34] Grubwieser P, Zimmermann B, Niederstätter H, Pavlic M, Steinlechner M, Parson W. Evaluation of an extended set of 15 candidate STR loci for paternity and kinship analysis in an Austrian population sample. Int J Leg Med. Mar. 2007;121(2):85–9. doi: 10.1007/s00414-006-0079-9.
[35] Tang B, Pan Z, Yin K, Khateeb A. Recent Advances of Deep Learning in Bioinformatics and Computational Biology. Front Genet. Mar. 2019;10:1–10. doi: 10.3389/fgene.2019.00214.
[36] Bera M. Artificial Intelligence in Bioinformatics, 2021. www.ijisrt.com.
[37] Min S, Lee B, Yoon S. Deep learning in bioinformatics. Brief Bioinforma. Sep. 01, 2017;18(5):851–69. doi: 10.1093/bib/bbw068.
[38] Liu W, Wang Z, Liu X, Zeng N, Liu Y, Alsaadi FE. A survey of deep neural network architectures and their applications. Neurocomputing. Apr. 2017;234:11–26. doi: 10.1016/j.neucom.2016.12.038.
[39] Bouwmans T, Javed S, Sultana M, Jung SK. Deep neural network concepts for background subtraction: A systematic review and comparative evaluation. Neural Netw. Sep. 2019;117:8–66. doi: 10.1016/j.neunet.2019.04.024.