Iraqi Journal of Information and Communications Technology(IJICT)
Vol. 8, Issue 1, April 2025
www.ijict.edu.iq
ISSN: 2222-758X
e-ISSN: 2789-7362

EFFICIENT DEEP LEARNING MODEL FOR DNA FORENSIC INVESTIGATIONS

Donya A. Khalid [1], **Nasser N. Khamiss** [2]
[1] Department of Cyber Security Engineering, College of Information Engineering, Al-Nahrain University, Jadriya, Baghdad, Iraq
[2] Department of Information and Communication Engineering, College of Information Engineering, Al-Nahrain University, Jadriya, Baghdad, Iraq
donya.abbas@nahrainuniv.edu.iq[1], nassrnafea@gmail.com[2]
Corresponding Author: **Nasser N. Khamiss**
Received:01/05/2023; Revised:01/10/2023; Accepted:28/10/2023
DOI:10.31987/ijict.8.1.246

***Abstract*- Recent advances in genetics have increased the sensitivity and reliability of the forensic sciences. Therefore, there is a need for efficient forensic investigation techniques. Deep learning is becoming increasingly important in forensic science as it offers the potential to increase the accuracy and effectiveness of various forensic tasks like paternity testing, missing person identification, and potentially connecting suspects to crime sites. Although it's performance at solving many problems, the deep learning model may suffer from overfitting problems that occur when it fails to generalize well and instead fits more precisely to the training dataset. This work presents two deep learning models, Deep Neural Network (D-DNN) and Gated Recurrent Neural Network (D-GRU), for human identification based on Deoxyribonucleic Acid-Short Tandem Repeat (DNA-STR) as the input sequence. These models are built and tested in such a way as the best performance since two regularization techniques are introduced: dropout and data augmentation to avoid overfitting problems. Two datasets are used: one with a size of 53530 and another with a size of 151580. Whereas 80% for training purposes of the first dataset is equal to 42824, while it is equal to 121264 in the second dataset. A comparison of performance was held between these two models by using dropout or not. The results show that D-GRU using dropout has the best performance with overfitting invisibility, training and testing accuracy equal to 1.0, the loss equal to 1.5276 $\times 10^{-8}$, and the validation loss equal to 6.1561 $\times 10^{-6}$.**

keywords: **DNA, STR, Deep learning, Regularization technique.**

I. INTRODUCTION

Deoxyribonucleic Acid (DNA) profiling analyses sequences of individual or mixed DNA profiles to identify the persons to whom these profiles belong. In forensic science, one of the most crucial applications of DNA profiling is to find a match between a suspect's blood sample and their DNA profile recovered at the crime scene. There are also other applications including: paternity testing, identifying disaster victims, locating missing persons, and mapping hereditary illnesses [1]. DNA profiling involves the analysis of specific sequences of DNA, known as genetic markers. Short Tandem Repeat (STR) is the most genetic marker commonly used [2]. Approximately 3% of the human genome consists of STRs, which are highly repetitive sequences of 2-6 base pairs. When assessing individuals for identification purposes, the variation in the number of repetition units provides a powerful discrimination factor [3].

This paper presents a deep learning model for forensic investigation cases such as paternity testing, missing person identification or determining the criminal in a crime scene by finding the similarity between DNA profiles of suspects with reference samples. It proposes a deep learning system using the DNA-STR profile as input data that stores human identity consisting of 15 loci with two alleles in each locus. Output from this system is the value that indicates whether

ISSN: 2222-758X
e-ISSN: 2789-7362

the person belongs to the reference samples or not. Besides introducing the deep learning model to perform optimally, the structuring and management of the creating dataset play a key role in system development.

Fig. 1 presents the general overview of the proposed system, where the identification model assumes two deep learning models called the DNA-deep Neural Network (D-DNN) and the DNA-Gated Recurrent Neural Network (D-GRU). These two models are built with and without the aid of two common regularization techniques: dropout and data augmentation for efficient learning performance. A comparison will be made between the two systems; both are structured to ensure efficient accuracy with minimum loss.

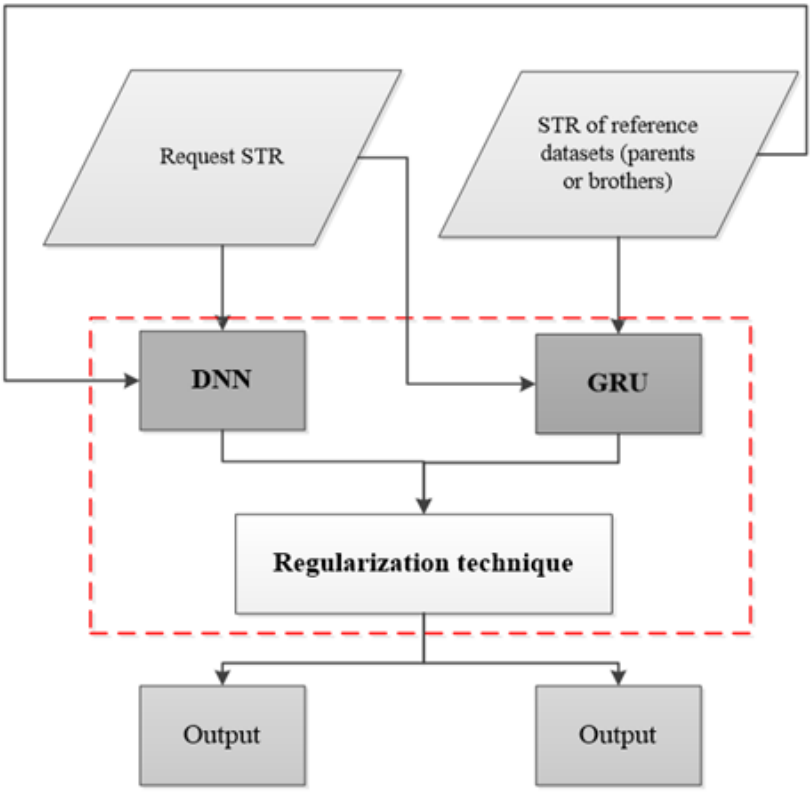

Figure 1: The proposed forensic system.

The rest of this paper is organized as follows. Section II gives a brief overview of DNA in forensic investigations. Section III presents deep learning in general and describing the regularization techniques. Section IV presents the proposed methodology. Section V shows the dataset distribution. Section VI gives the main results and discussions. Finally, the conclusions are presented in Section VII.

ISSN: 2222-758X
e-ISSN: 2789-7362

II. RELATED WORKS

This paper introduces deep learning technique for identifying an unknown person by finding his or her correct family. Based on the survey, some researchers concentrate their works on identifying DNA in human profiling, as follows.

Michael A. Marciano et al. 2017 [4] employed a probabilistic support vector machine technique to estimate the number of contributors in a DNA mixture with 96% accuracy. Miyake, et al. 2018 [5] employed a deep learning autoencoder to extract the properties of Human Leukocyte Antigen (HLA)-A long-chain DNA, perhaps aiding in the creation of a vaccine. The authors looked at the HLA-A DNA sequences. In terms of obtaining features from the human leukocyte antigen (HLA-A), deep learning autoencoders have demonstrated potential. Maria Anggreainy et al. 2019 [6] developed a technique to use fuzzy similarity to gauge how similar human DNA profiles are. DNA profile information is used as an input in this fuzzy system to record a person's identification in addition to their DNA profile. The information entered is the outcome of an electropherogram made up of 16 loci, each of which has two alleles, and was identified using the Polymerase Chain Reaction (PCR). The value of each individual's similarity to the reference and to the three other similarity levels-small, medium, and high is the output of this fuzzy system.

Corina C.G. et al 2019 [7] used machine learning techniques to extract as much value as possible from profile data. Utilizing the advantages of machine learning, they put forth a probabilistic method for determining the total number of contributors in a DNA mixture. They demonstrated a performance accuracy of 83% using the random forest as their classifier. Yao-Yuan Liu et al. 2020 [8] introduced Fragsifier, a machine learning method for detecting and extracting STR sequences from massively parallel sequencing data. This method finds the longest repeat stretches on each read, then uses k-mers in a machine learning sequence model to predict the locus to identify STRs on each read. Viviane Siino et al. 2020 [9] presented an Artificial Intelligence (AI) system that employs a prediction cascade using gradient descent logistic regression. This enables the iterative resolution of cases involving several missing persons. In addition, the AI has the ability to assess the level of confidence associated with likelihood ratios for various pedigrees, irrespective of the quantity of genetic data accessible and the number of individuals whose information is missing. Hamdah Alotaibi et al. in 2022 [1] created a software application, named TAWSEEM, that uses PROVEDIt dataset to estimate the number of unknown contributors in DNA mixture profiles using a Multi-Layer Perceptron (MLP) as a neural network deep learning model. The accuracy of the suggested method is equivalent to 97%.

However, this study introduces a deep learning framework designed for forensic investigation scenarios, including but not limited to paternity testing, missing person identification, and criminal identification in crime scenes. The proposed model aims to identify similarities between DNA profiles of suspects and reference samples.

III. DNA IN FORENSIC INVESTIGATIONS

DNA, or deoxyribonucleic acid, is also referred to as "blueprint of life" due to the fact that it carries all the information necessary for the proper functioning and reproduction of an organism [10]. Since the discovery of DNA's use in forensic investigation, it has played a significant role in the criminal justice system. In the majority of criminal cases, the DNA profile of an evidence sample collected from the crime site is compared to the DNA profile of a reference sample. However,

ISSN: 2222-758X
e-ISSN: 2789-7362

when no reference sample is available for comparison, familial DNA analysis can provide crucial investigation leads by identifying an individual during a criminal investigation. In addition, this analysis is proving effective in identifying an individual's ethnicity and ancestry [11], [12]. Typically, forensic biological materials obtained from a crime site are typed for autosomal STR loci [13]. The human genome has less than 25,000 genes, which are coded for proteins, among its 3 billion bases, which are distributed across the 23 chromosomes [14], and the repetitive sequences may be located in the non-coding regions of the genome [15]. In forensic DNA typing, STR markers are used to identify the human remains of missing people, confirm familial relationships, and possibly link suspects to crime scenes [16], [17]. Although different genetic markers are used for different purposes in forensic DNA analysis, STR typing is still the mainstay [3]. This is because each person's repetition of the pattern is unique. This variation provides a high level of discrimination and creates an individual profile from each person's STR pattern [18].

IV. EFFICIENT DEEP LEARNING

The term "deep" in "deep learning" refers to any type of deeper understanding acquired by the technique; rather, it refers to the concept of successive layers of representations. The multilayer architecture of deep neural networks echoes the structure of visual neuroscience, and nonlinear modules can transform the data representation into a more abstract form [19], [20]. The input layer, with several hidden layers, and an output layer comprise the fundamental architecture of deep neural networks. Since many difficult problems can be resolved by training deep neural networks, they have attracted an incredible amount of attention from scientists. Good performance results are possible for deep neural networks if enough data is provided during training. However, overfitting and underfitting issues arise in the predefined neural network model if training data are insufficient [21]. In order to address these issues, a number of regularization approaches have been developed and put into widespread use in the fields of application development and data analysis [22], [23]. Regularization is one of the key components of deep learning, which is a method for making minor adjustments to the learning process to improve the model's ability to generalize. This enhances the model's performance on the unobserved data as well [24], [25]. Regularization consists of different techniques and methods used to address the issue of overfitting by reducing the generalization error without affecting the training error much. Some of the techniques are weight decay, adding noise, dropout, data augmentation, and early stopping [26], [27], [28].

V. PROPOSED MODEL

This work presents a unifying, systematic taxonomy to identify the human profile. Two deep learning methods are distinguished with the aid of two regularization techniques that affect data, network structure, error values, terms of regularization, and optimization procedures. The suggested approach is primarily related to data augmentation and dropout, two widespread regularization methods.

Here, an analysis is conducted on how these approaches were employed to train two deep learning models (i.e., DNN and GRU), which are suggested here for forensic investigations like paternity testing or missing person identification. To confirm human profiling in the forensic investigation, data preparation is the first step in this system. 15 loci-STR are used for matching purposes between the alleles of DNA profiles of suspects and reference samples. The datasets used are

ISSN: 2222-758X
e-ISSN: 2789-7362

samples from real persons from Iraq, Najaf province [29], where each locus has a pair of two values called alleles. The Table I shows the applied STR-15 loci.

TABLE I
Applied STR-15 LOCI [29]

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

The proposed data creation model is presented in Fig. 2. It assumes familial data creation for the purpose of human identification and kinship determination, where two datasets with different data sizes are created: the first one is of size 53,530 samples based on the original DNA familial dataset with shuffling, while the second one is with augmented data results of size 151,580 samples.

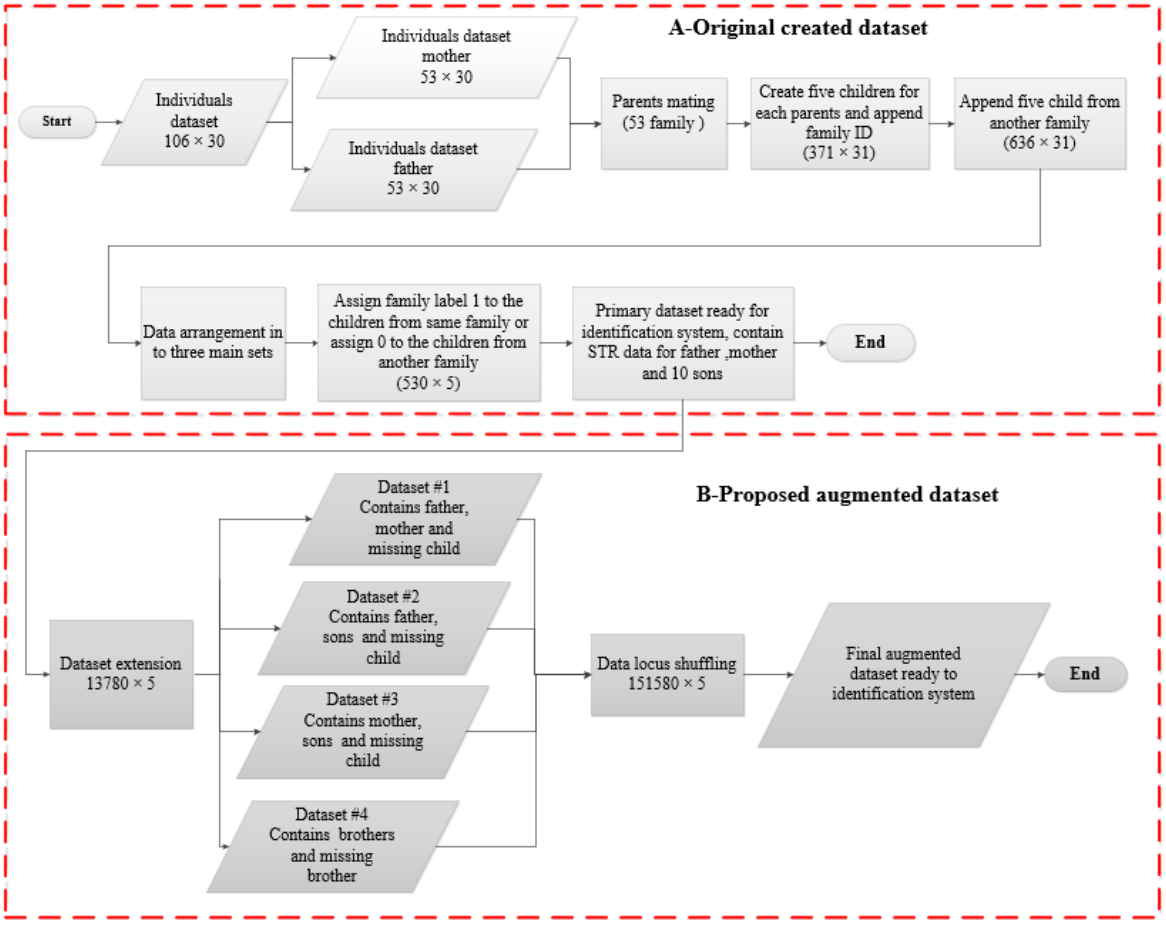

Figure 2: Data creation.

ISSN: 2222-758X
e-ISSN: 2789-7362

The proposed method of familial datasets creation can be summarized in the following steps:

1) It takes 106 real individuals as parents to create 53 familial datasets, each familial dataset consists of a mother, father, and five correct children.
2) Append five incorrect children to the assigned family; this process results in 530 data samples. This step aims to teach the proposed system the relationship between the family (mother, father) and their corresponding children.
3) A family ID from 0 to 52 is assigned for each family.
4) The labels are also appended to all family datasets, in which label (1) indicates whether the son truly belongs to this family or does not belong to the family (0). Fig. 3 shows the proposed family structure.
5) The created dataset of 53,530 samples, part A in Fig. 2, is iterated 11 times with a randomly shuffling technique for each set (father, mother, and children); this is the first level of augmentation. Part B in Fig. 2 results in 151,580 samples, which is the second level of the augmentation process, and it is described in the flowing cases as follows:
   - **Case 1:** The system supposes that both parents (mother and father) are alive, and they have a missing child to be identified among many undefined profiles. Both mother and father are used as a reference.
   - **Case 2 and case 3:** One party is dead, either mother or father, and they have a missing child to be identified among many unknown profiles. Only one of them is used as a reference.
   - **Case 4:** Both parents (mother and father) are dead, and they have a missing child to be identified among many profiles. His or her brothers are used as references.

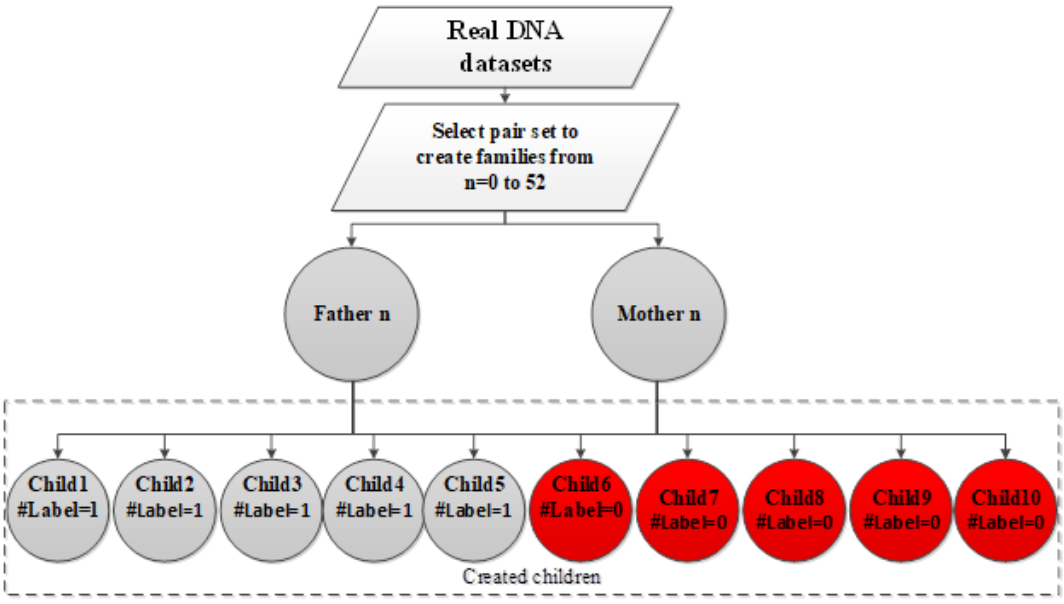

Figure 3: Structure of one family.

ISSN: 2222-758X
e-ISSN: 2789-7362

The proposed deep learning networks are structured for the purpose of an efficient human identification system. A D-DNN-identification system (refer to Fig. 4) is constructed from an embedded layer, flatten layer, a dense layer, and a dropout layer. This model takes input from the missing profile and the corresponding reference persons. The output of this model is a decision as to whether or not the missing person relates to the same reference or not. The D-DNN model has 8 layers: an embedding layer, five dense layers, and three dropout layers. It has a total number of trainable parameters equal to 806,401, and there are no non-trainable parameters. Table II summarizes the model parameters.

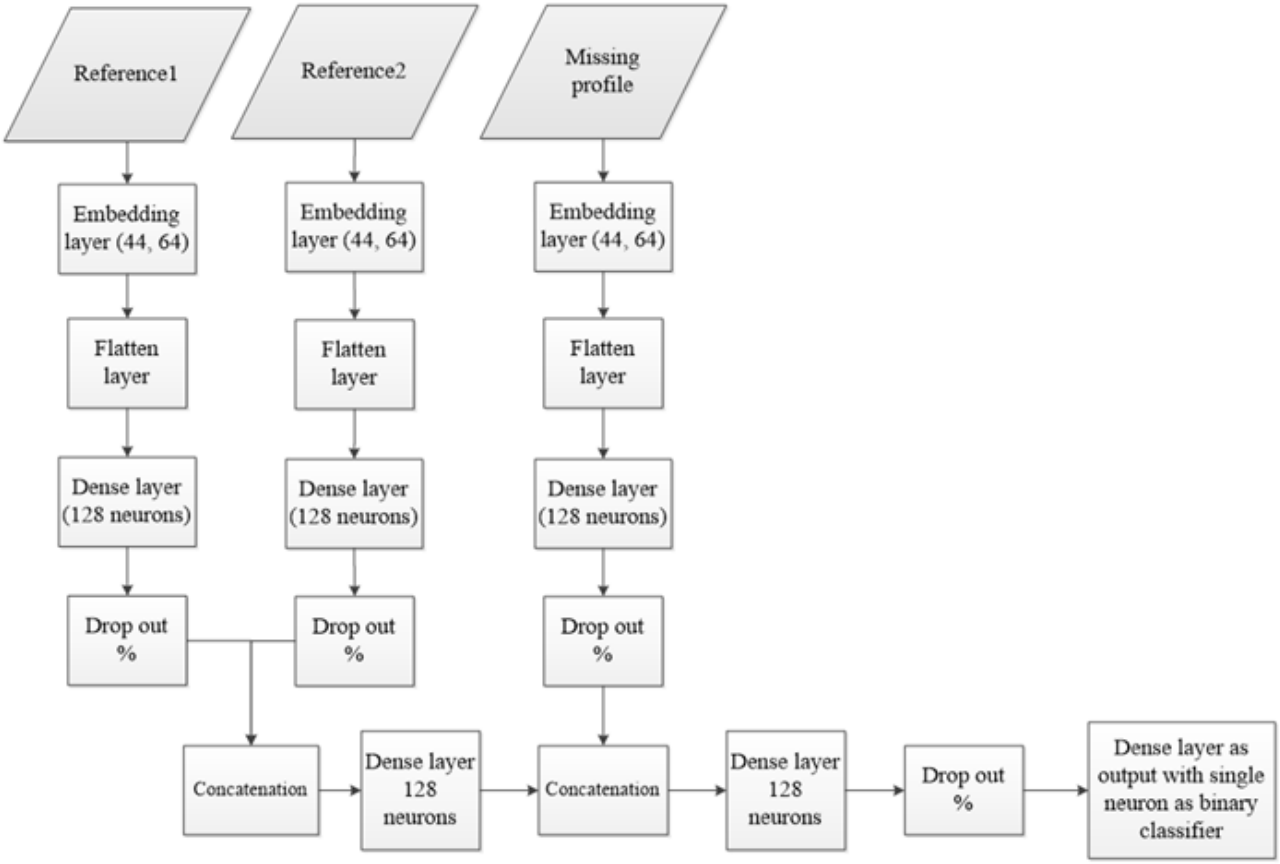

Figure 4: Proposed D-DNN model

While the structure of the D-GRU model (refer to Fig. 5) is constructed from the embedded layer, the GRU layer, and the dropout layer. This model also takes input datasets for the missing person's identity and the corresponding reference profiles. The output of this model is a decision as to whether the missing profile belongs to the same references or not. Table III describes DNA-GRU model parameters. It contains seven layers: an embedding layer, five GRU layers, two dense layers, and a single dropout layer. It has a total number of trainable parameters equal to 157,377.

ISSN: 2222-758X
e-ISSN: 2789-7362

TABLE II
D-DNN Model Parameter.

DNA Model: DNN		
Layer (shape)	**Output shape**	**Parameters**
embedding (Embedding)	multiple	2816
dense (Dense)	multiple	245,888
dropout (Dropout)	multiple	0
$dense_1$ (Dense)	multiple	245,888
$dropout_1$ (Dropout)	multiple	0
$dense_2$ (Dense)	multiple	32,896
$dropout_2$ (Dropout)	multiple	0
$dense_3$ (Dense)	multiple	245,888
$dropout_3$ (Dropout)	multiple	0
$dense_3$ (Dense)	multiple	32,896
$dropout_4$ (Dropout)	multiple	0
$dense_4$ (Dense)	multiple	129
Total parameters= 806,401 Trainable parameters= 806,401 Non-trainable parameters: 0		

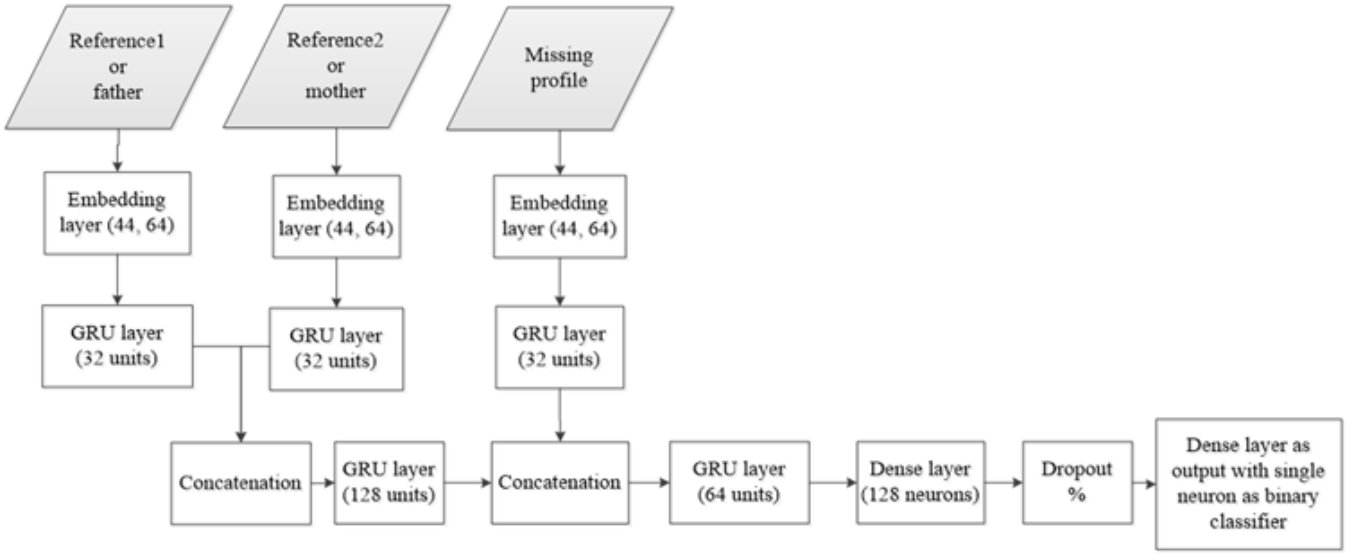

Figure 5: Proposed D-GRU model.

ISSN: 2222-758X
e-ISSN: 2789-7362

TABLE III
D-GRU Model Parameter

DNA Model: GRU		
Layer (shape)	**Output shape**	**Parameters**
embedding (Embedding)	multiple	2816
gru (GRU)	multiple	9408
gru_1 (GRU)	multiple	9408
gru_2 (GRU)	multiple	74,496
gru_3 (GRU)	multiple	9408
gru_4 (GRU)	multiple	43,392
dense (Dense)	multiple	8320
dropout (Dropout)	multiple	0
$dense_1$ (Dense)	multiple	129
Total parameters=157,377 Trainable parameters= 157,377 Non-trainable parameters: 0		

VI. RESULTS AND DISCUSSION

During the training and testing phases, the entire two datasets are divided into 80% training and 20% testing sets. The training dataset is further divided into balanced labels, with half of the data labelled as 1, which refers to correct children, and the other half labelled as 0 for false children. Fig. 6-a shows the distribution of labels in a histogram for the first dataset that has a size of 50,503 samples, while Fig. 6-b shows the histogram for 151,580 samples.

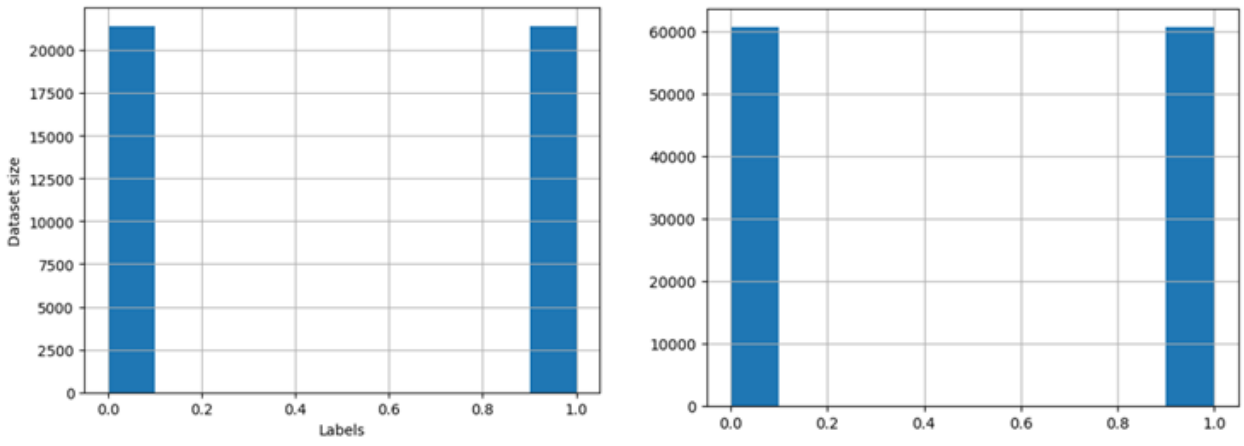

Figure 6: Histogram of the training datasets.

The training and testing processes are carried out separately through 20 epochs for each of the proposed DNA deep learning

models and their associated regularization techniques to make comparisons and evaluations. The results are grouped into two main groups: one illustrates the performance of two deep learning models without using dropout over a data size of 50,503, and the second group uses dropout equal to 30% for the same dataset. The same evaluations are repeated again on the augmented data size, which equals to 151,580 samples. The outcomes are acquired using the Google Colab platform, which relies on the Python programming language. Essential libraries, such as TensorFlow, Pandas, Numpy, Matplotlib, Seaborn, Scikit-learn, and Keras, are employed in the process.

A. Results of first DNA dataset 53530

Fig. 7 (a, b, c, and d) shows the accuracy for both D-DNN and D-GRU with and without using dropout for the first dataset of size 53,530. Fig. 8 (a, b, c, and d) shows the loss for both D-DNN and D-GRU with and without using dropout

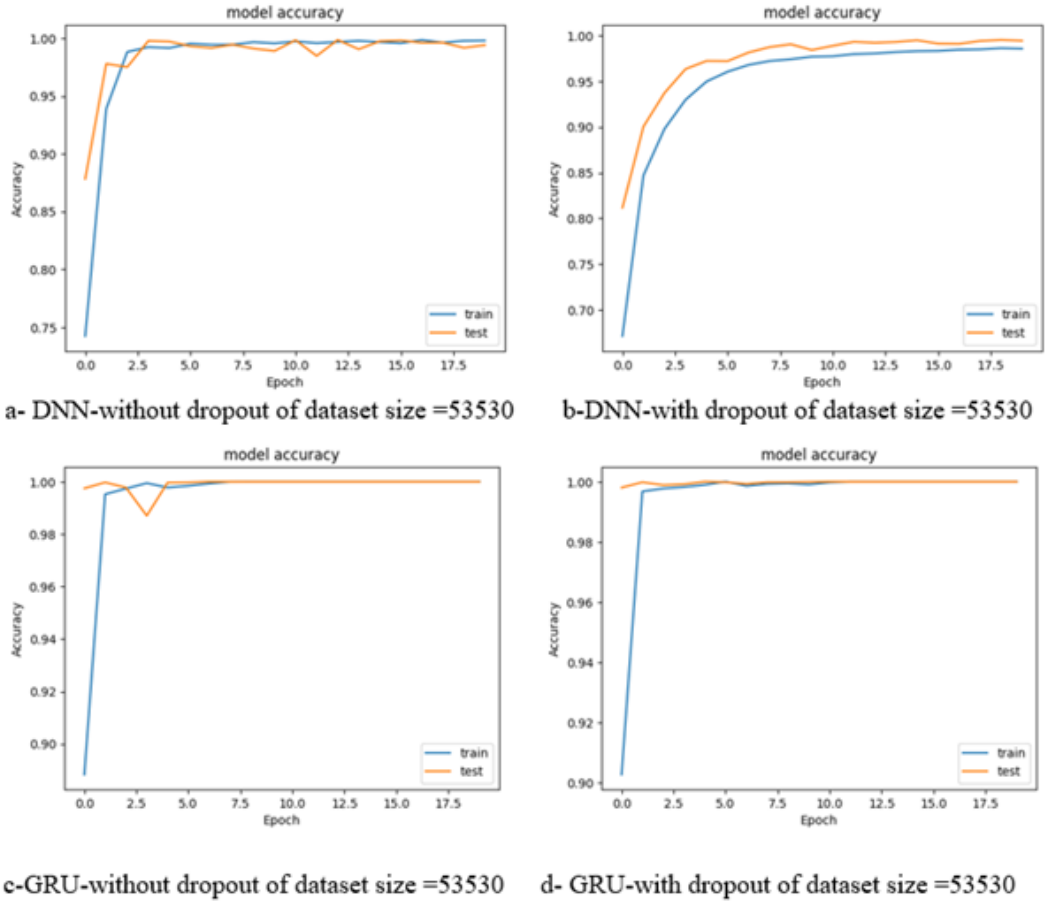

Figure 7: Accuracy of two models with and without dropout for 53,530 datasets.

for the first dataset of size 53,530. Fig. 7 and Fig. 8 show the performance of two different neural network models, D-DNN and D-GRU, with and without dropout regularization, over 20 epochs of training. They show the training and

ISSN: 2222-758X
e-ISSN: 2789-7362

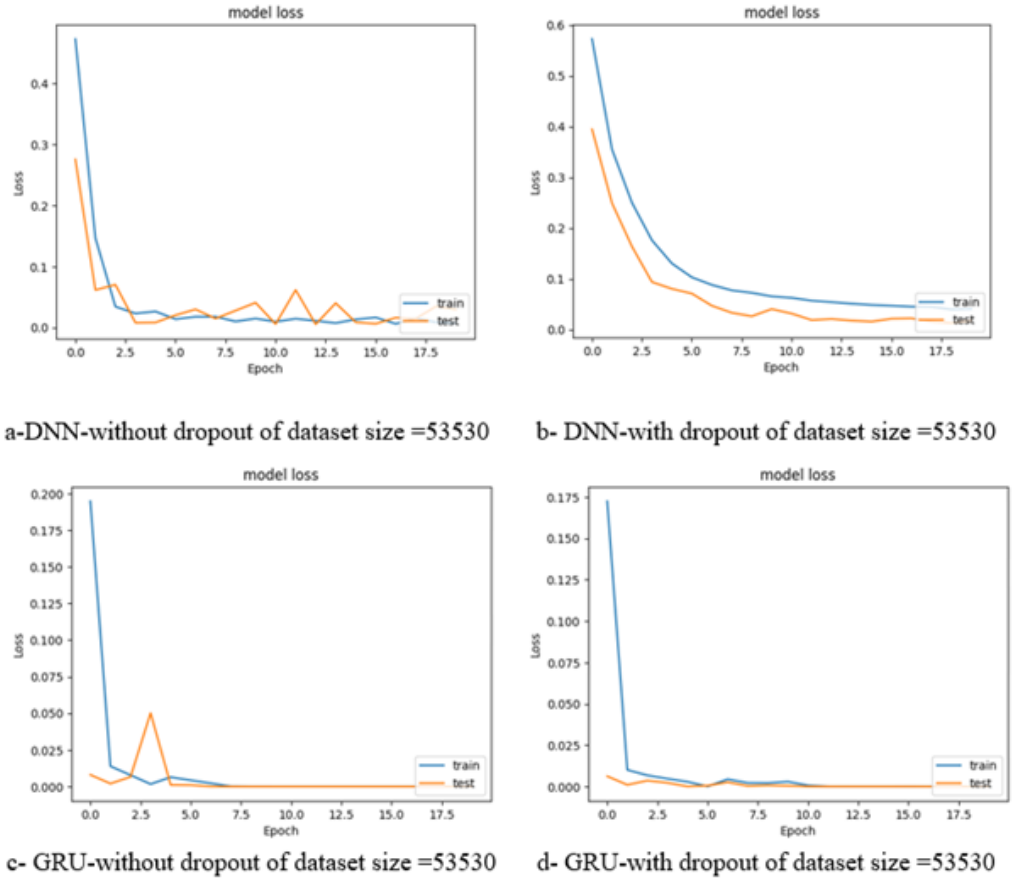

Figure 8: The loss of two models with and without dropout for 53,530 datasets.

validation loss for each epoch of training. The lower the loss, the better the model's performance; it shows that both models with dropout regularization perform better than their counterparts without dropout regularization; which is indicated by the disappearance of overfitting in contrast to the overfitting appearance when no dropout is applied.

It also seems that the D-GRU model performs better than the D-DNN model, with lower loss values for both the training and validation sets. However, it's worth noting that the D-DNN architecture seems to have reached convergence faster, as indicated by the much lower training and validation loss values after the first epoch. Overall, the D-GRU architecture with dropout regularization seems to be the best-performing model, with the lowest validation loss values throughout the training process.

Fig. 9 (a, b, c, and d) shows the confusion matrix for both D-DNN and D-GRU. With and without using dropout for the 53,503 datasets, a confusion matrix is presented in this paper to measure classification performance using four evaluation parameters: True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN).

ISSN: 2222-758X
e-ISSN: 2789-7362

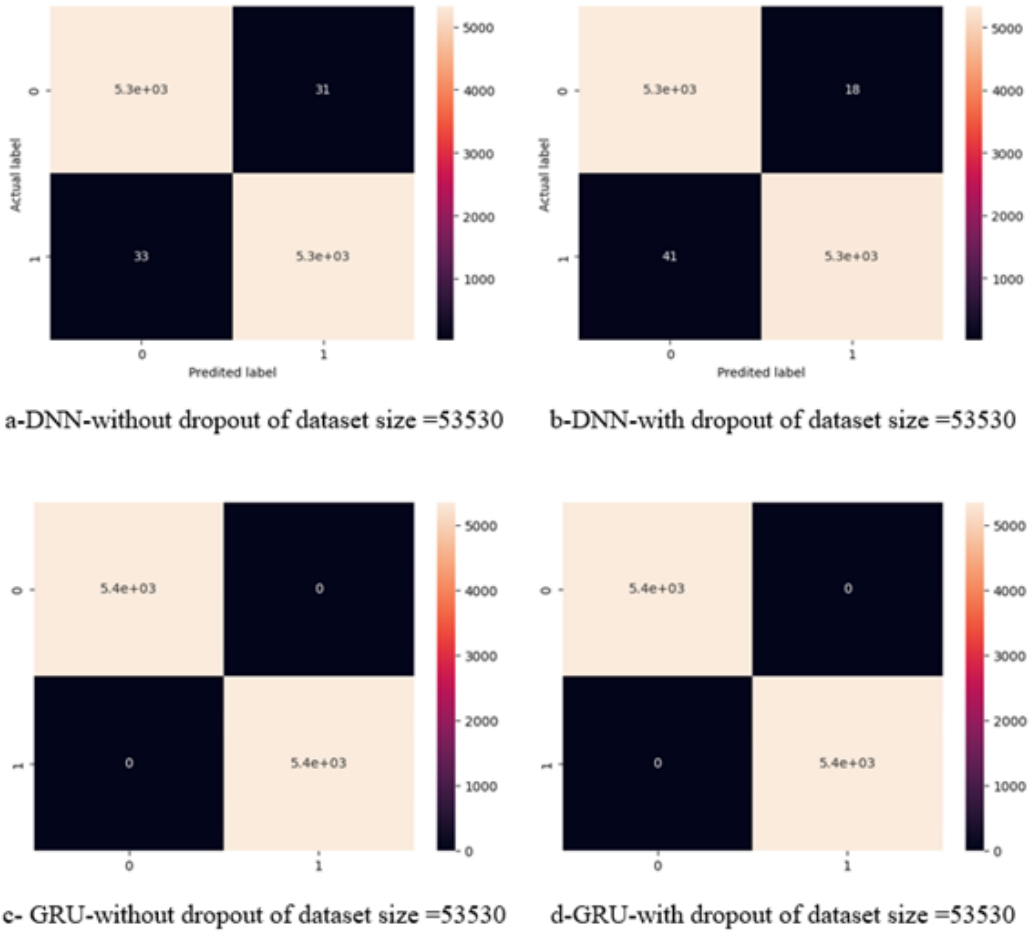

Figure 9: Confusion matrix of two models with and without dropout for 53,530 datasets.

Based on the confusion matrix in Fig. 9, it can be deduced again that D-GRU performs better than D-DNN since no false predictions have occurred in either situation (applying dropout or not). In the case of the D-DNN model, it shows that it performs better with dropout because it has incorrect predictions (FP+FN) equal to 59, which is smaller than 64 in the case without applying dropout due to overfitting taking place.

Table IV and Table V show the performance comparison between D-DNN and D-GRU over 20 epochs for both loss and accuracy respectively. The tabulated results were obtained over a dataset of size 53,530 samples, assuming two cases: with and without dropout. Each row in Table V shows the accuracy and validation accuracy of each model at a given epoch. The table shows that the models generally improve in accuracy and validation accuracy as the number of epochs increases, although the rate of improvement starts to level off after a certain point. It also appears that the models with dropout regularization tend to have higher validation accuracy than the models without dropout, indicating that dropout is

ISSN: 2222-758X
e-ISSN: 2789-7362

helping to prevent overfitting. Overall, it seems that the D-GRU model with dropout achieves the best performance, with a validation accuracy of 0.9953 at epoch 6 and 1.0000 at epochs 13-20.

TABLE IV
Performance Comparison Between Loss and Validation Loss of Two Deep Learning Models With and Without Dropout Regularization for Dataset Size =53,530

Model	**D-DNN #53,530**				**D-GRU #53,530**			
	Without Dropout		**With Dropout**		**Without Dropout**		**With Dropout**	
Epoch	**Accuracy**	**Validation Accuracy**	**Accuracy**	**Validation Accuracy**	**Accuracy**	**Validation Accuracy**	**Accuracy**	**Validation Accuracy**
1	0.4731	0.2757	0.5730	0.3950	0.2127	0.0159	0.1724	0.0062
2	0.1468	0.0615	0.3554	0.2495	0.0096	0.0018	0.0101	0.0010
3	0.0342	0.0703	0.2503	0.1640	0.0038	$2.38e^{-4}$	0.0068	0.0036
4	0.0231	0.0078	0.1763	0.0941	0.0068	0.0417	0.0047	0.0022
5	0.0263	0.0082	0.1306	0.0806	0.0024	$1.07e^{-4}$	0.0029	$4.64e^{-5}$
6	0.0137	0.0201	0.1033	0.0713	0.0025	$5.41e^{-5}$	$9.37e^{-5}$	$5.39e^{-4}$
7	0.0176	0.0298	0.0884	0.0473	$2.07e^{-5}$	$1.07e^{-5}$	0.0044	0.0026
8	0.0177	0.0146	0.0776	0.0335	$4.77e^{-6}$	$5.88e^{-6}$	0.0022	$3.56e^{-4}$
9	0.0100	0.0277	0.0728	0.0266	$2.08e^{-6}$	$3.04e^{-6}$	0.0021	$6.83e^{-4}$
10	0.0147	0.0407	0.0657	0.0407	$9.55e^{-7}$	$1.78e^{-6}$	0.0029	$4.57e^{-4}$
11	0.0096	0.0056	0.0629	0.0318	$4.55e^{-7}$	$1.04e^{-6}$	$7.02e^{-4}$	$5.61e^{-5}$
12	0.0144	0.0616	0.0573	0.0192	$2.16e^{-7}$	$5.77e^{-7}$	$7.70e^{-6}$	$3.54e^{-5}$
13	0.0110	0.0052	0.0573	0.0213	$1.06e^{-7}$	$3.52e^{-7}$	$1.38e^{-6}$	$3.92e^{-5}$
14	0.0073	0.0400	0.0514	0.0180	$5.27e^{-8}$	$2.57e^{-7}$	$9.53e^{-7}$	$3.39e^{-5}$
15	0.0133	0.0086	0.0489	0.0161	$2.64e^{-8}$	$1.77e^{-7}$	$3.73e^{-7}$	$3.62e^{-5}$
16	0.0165	0.0061	0.0473	0.0218	$1.34e^{-8}$	$1.10e^{-7}$	$2.50e^{-7}$	$5.11e^{-5}$
17	0.0060	0.0161	0.0454	0.0225	$7.12e^{-9}$	$8.79e^{-8}$	$1.23e^{-7}$	$2.18e^{-5}$
18	0.0152	0.0145	0.0450	0.0157	$3.81e^{-9}$	$5.65e^{-8}$	$1.21e^{-7}$	$7.07e^{-6}$
19	0.0086	0.0344	0.0401	0.0131	$2.19e^{-9}$	$5.41e^{-8}$	$3.39e^{-8}$	$3.58e^{-6}$
20	0.0070	0.0295	0.0410	0.0178	$1.28e^{-9}$	$3.61e^{-8}$	$1.53e^{-8}$	$6.16e^{-6}$

TABLE V
Performance Comparison Between Loss and Validation Loss of Two Deep Learning Models With and Without Dropout Regularization for Dataset Size =151,580

Model	D-DNN #151,580				D-GRU #151,580			
	Without Dropout		With Dropout		Without Dropout		With Dropout	
Epoch	Loss	Validation Loss	Loss	Validation Loss	Loss	Validation Loss	Loss	Validation Loss
1	0.7424	0.8783	0.6712	0.8116	0.8722	0.9953	0.9027	0.9980
2	0.9390	0.9779	0.8471	0.9005	0.9968	0.9995	0.9968	0.9998
3	0.9882	0.9751	0.8980	0.9373	0.9987	0.9999	0.9977	0.9989
4	0.9923	0.9977	0.9299	0.9635	0.9977	0.9865	0.9982	0.9992
5	0.9915	0.9971	0.9498	0.9724	0.9991	1.0000	0.9989	1.0000
6	0.9953	0.9931	0.9606	0.9722	0.9993	1.0000	1.0000	0.9998
7	0.9943	0.9915	0.9680	0.9816	1.0000	1.0000	0.9986	0.9993
8	0.9944	0.9945	0.9723	0.9875	1.0000	1.0000	0.9993	0.9998
9	0.9966	0.9910	0.9742	0.9906	1.0000	1.0000	0.9994	0.9998
10	0.9956	0.9890	0.9770	0.9843	1.0000	1.0000	0.9991	0.9999
11	0.9972	0.9983	0.9774	0.9889	1.0000	1.0000	0.9997	1.0000
12	0.9958	0.9847	0.9798	0.9933	1.0000	1.0000	1.0000	1.0000
13	0.9968	0.9985	0.9805	0.9922	1.0000	1.0000	1.0000	1.0000
14	0.9977	0.9905	0.9821	0.9931	1.0000	1.0000	1.0000	1.0000
15	0.9967	0.9974	0.9830	0.9950	1.0000	1.0000	1.0000	1.0000
16	0.9957	0.9980	0.9833	0.9913	1.0000	1.0000	1.0000	1.0000
17	0.9985	0.9961	0.9846	0.9910	1.0000	1.0000	1.0000	1.0000
18	0.9961	0.9962	0.9850	0.9942	1.0000	1.0000	1.0000	1.0000
19	0.9978	0.9917	0.9862	0.9953	1.0000	1.0000	1.0000	1.0000
20	0.9980	0.9940	0.9857	0.9945	1.0000	1.0000	1.0000	1.0000

B. Results for second DNA dataset 151,580

The same procedure is repeated in the second dataset of size 151,580 samples, and the best results were conducted in epoch 20 with the following values shown in Table VI and Table VII. This proves that the D-GRU model with dropout is better at removing overfitting and generalizing to new data, as it has lower validation loss values throughout training.

TABLE VI
Performance Comparison Between Accuracy and Validation Accuracy of Two Deep Learning Models With and Without Dropout Regularization for Dataset Size =151,580

Model	D-DNN #151,580				D-GRU #151,580			
	Without Dropout		With Dropout		Without Dropout		With Dropout	
Epoch	Loss	Validation Loss	Loss	Validation Loss	Loss	Validation Loss	Loss	Validation Loss
20	0.0294	0.5428	0.1554	0.1351	0.0021	0.0046	0.0019	0.0031

TABLE VII
Performance Comparison Between Accuracy and Validation Accuracy of Two Deep Learning Models With and Without Dropout Regularization for Dataset Size =53,530

Model	D-DNN #53,530				D-GRU #53,530			
	Without Dropout		With Dropout		Without Dropout		With Dropout	
Epoch	Accuracy	Validation Accuracy	Accuracy	Validation Accuracy	Accuracy	Validation Accuracy	Accuracy	Validation Accuracy
20	0.9901	0.9094	0.9405	0.9492	0.9993	0.9987	0.9993	0.9989

To make a comparison between the nearest related works and the proposed model in terms of applied methods, and reported accuracy, Table VIII illustrates that the proposed models have the highest accuracy among them.

TABLE VIII
Comparison with Related AI Works

Research	Applied Method	Accuracy
Anggreainy et al. [2]	Fuzzy	80%
Siino V and Sears C [9]	Gradient descent logistic regression	95%
Proposed work: DNN for 53,530 samples	Without Dropout	99%
	With Dropout	99%
Proposed work: DNN for 151,580 samples	Without Dropout	90%
	With Dropout	94%
Proposed work: GRU for 53,530 samples	Without Dropout	100%
	With Dropout	100%
Proposed work: GRU for 151,580 samples	Without Dropout	99%
	With Dropout	99%

VII. CONCLUSION

The purpose of this paper is to propose a method for identifying human profiles based on STR-DNA and to evaluate the performance of two deep learning models (i.e., D-DNN and D-GRU) with the aid of two regularization techniques for this task. The models are binary classifiers that aim to identify human profiles by comparing two DNA-STR samples, one from a known relative and the other from a missing person. The results prove that the D-GRU model with dropout is better than the D-DNN model at preventing overfitting and generalizing to new data; it seems to be the better choice for human identification problems as it has lower validation loss values throughout training. Overall, using dropouts improved the performance of both models because no overfitting occurred, as opposed to the case of no dropout overfitting that had appeared in both models. The dropout minimizes the structure of the network by 30% since it makes the model select neurons randomly. This reduction makes the deep learning model generalize better.

The results of applying dropout for the 50,530-dataset size show a significant improvement in the values of accuracy and loss of the D-GRU model. It gives the highest training and testing accuracy equal to 1.0, a minimum loss equal to

1.5276×10^{-8}, a validation loss equal to 6.1561×10^{-6}, and total false predictions equal to 59. D-DNN also performs better with no overfitting appearance in case of dropout, with the highest training accuracy equal to 0.9857, a testing accuracy equal to 0.9857, a loss value equal to 0.0410, a validation loss equal to 0.0178, and total false predictions equal to 64. The conclusion is drawn that the D-GRU model is the best choice for DNA forensic human identification because D-GRU deals better with sequence data like DNA than DNN. In the future the proposed system can be used for medical disease diagnosis with suitable datasets.

FUNDING

None.

ACKNOWLEDGEMENT

The author would like to thank the reviewers for their valuable contribution in the publication of this paper.

CONFLICTS OF INTEREST

The author declares no conflict of interest.

REFERENCES

[1] H. Alotaibi, F. Alsolami, E. Abozinadah, and R. Mehmood, "TAWSEEM: A Deep-Learning-Based Tool for Estimating the Number of Unknown Contributors in DNA Profiling," Electronics (Basel), vol. 11, no. 4, p. 548, Feb. 2022, doi: 10.3390/electronics11040548.
[2] N. Wyner, M. Barash, and D. McNevin, "Forensic Autosomal Short Tandem Repeats and Their Potential Association With Phenotype," Front Genet, vol. 11, Aug. 2020, doi: 10.3389/fgene.2020.00884.
[3] A. Keerti and S. Ninave, "DNA Fingerprinting: Use of Autosomal Short Tandem Repeats in Forensic DNA Typing," Cureus, vol. 14, no. 10, Oct. 2022, doi: 10.7759/cureus.30210.
[4] M. A. Marciano and J. D. Adelman, "PACE: Probabilistic Assessment for Contributor Estimation— A machine learning-based assessment of the number of contributors in DNA mixtures," Forensic Sci Int Genet, vol. 27, pp. 82–91, Mar. 2017, doi: 10.1016/j.fsigen.2016.11.006.
[5] J. Miyake, Y. Kaneshita, S. Asatani, S. Tagawa, H. Niioka, and T. Hirano, "Graphical classification of DNA sequences of HLA alleles by deep learning," Hum Cell, vol. 31, no. 2, pp. 102–105, Apr. 2018, doi: 10.1007/s13577-017-0194-6.
[6] M. S. Anggreainy, M. R. Widyanto, B. Widjaja, N. Soedarsono, and P. T. Widodo, "Family relation and STR-DNA matching using fuzzy inference," International Journal of Electrical and Computer Engineering (IJECE), vol. 9, no. 2, p. 1335, Apr. 2019, doi: 10.11591/ijece.v9i2.pp1335-1345.
[7] C. C. G. Benschop, J. van der Linden, J. Hoogenboom, R. Ypma, and H. Haned, "Automated estimation of the number of contributors in autosomal short tandem repeat profiles using a machine learning approach," Forensic Sci Int Genet, vol. 43, p. 102150, Nov. 2019, doi: 10.1016/j.fsigen.2019.102150.
[8] Y.-Y. Liu, D. Welch, R. England, J. Stacey, and S. Harbison, "Forensic STR allele extraction using a machine learning paradigm," Forensic Sci Int Genet, vol. 44, p. 102194, Jan. 2020, doi: 10.1016/j.fsigen.2019.102194.
[9] V. Siino and C. Sears, "Artificially intelligent scoring and classification engine for forensic identification," Forensic Sci Int Genet, vol. 44, p. 102162, Jan. 2020, doi: 10.1016/j.fsigen.2019.102162.
[10] J. L. Bukyya et al., "DNA Profiling in Forensic Science: A Review," Glob Med Genet, vol. 08, no. 04, pp. 135–143, Dec. 2021, doi: 10.1055/s-0041-1728689.
[11] R. M. Mateen, M. F. Sabar, S. Hussain, R. Parveen, and M. Hussain, "Familial DNA analysis and criminal investigation: Usage, downsides and privacy concerns.," Forensic Sci Int, vol. 318, p. 110576, Jan. 2021, doi: 10.1016/j.forsciint.2020.110576.
[12] J. Ge, H. Sun, H. Li, C. Liu, J. Yan, and B. Budowle, "Future directions of forensic DNA databases," Croat Med J, vol. 55, no. 2, pp. 163–166, Apr. 2014, doi: 10.3325/cmj.2014.55.163.
[13] J. Ge and B. Budowle, "Forensic investigation approaches of searching relatives in DNA databases," J Forensic Sci, vol. 66, no. 2, pp. 430–443, Mar. 2021, doi: 10.1111/1556-4029.14615.
[14] I. López-Flores and M. A. Garrido-Ramos, "The Repetitive DNA Content of Eukaryotic Genomes," 2012, pp. 1–28. doi: 10.1159/000337118.
[15] R. Youngest, V. Saamia, D. A. Oktaviani, S. B. Aritonang, I. M. Wiranatha, and I. Rofiq, "Str Locus Mutations In Paternity Case," Jurnal Biosains Pascasarjana, vol. 24, no. 1, pp. 34–49, Jun. 2022, doi: 10.20473/jbp.v24i1.2022.34-49.
[16] J. M. Butler, "The future of forensic DNA analysis," Philosophical Transactions of the Royal Society B: Biological Sciences, vol. 370, no. 1674, p. 20140252, Aug. 2015, doi: 10.1098/rstb.2014.0252.
[17] M. K. Ibrahem and A. S. Sadiq, "INTEGRATED BIOMETRIC DNA IDENTIFICATION SYSTEM," Iraqi Journal of Information and Communications Technology (IJICT), vol. 1, no. 1, 2018, [Online]. Available:https://ijict.edu.iq
[18] Emily Niedzwiecki, Sara Debus-Sherrill, and Michael B. Field, "Understanding Familial DNA Searching: Coming to a Consensus on Terminology," 2016.
[19] W. Liu, Z. Wang, X. Liu, N. Zeng, Y. Liu, and F. E. Alsaadi, "A survey of deep neural network architectures and their applications," Neurocomputing, vol. 234, pp. 11–26, Apr. 2017, doi: 10.1016/j.neucom.2016.12.038.
[20] Narmin Majid Dahham and Lahieb Mohammed Jawad, "State-of-the-Art for Email Phishing Detection Techniques based on Deep Learning," in MIDDLE EASTERN SIMULATION AND MODELLING CONFERENCE 2023 "MESM'2023," Baghdad : EUROSIS-ETI, Nov. 2023, pp. 18–21.

[21] C. F. G. Dos Santos and J. P. Papa, "Avoiding Overfitting: A Survey on Regularization Methods for Convolutional Neural Networks," ACM Comput Surv, vol. 54, no. 10s, pp. 1–25, Jan. 2022, doi: 10.1145/3510413.
[22] I. Nusrat and S.-B. Jang, "A Comparison of Regularization Techniques in Deep Neural Networks," Symmetry (Basel), vol. 10, no. 11, p. 648, Nov. 2018, doi: 10.3390/sym10110648.
[23] R. Moradi, R. Berangi, and B. Minaei, "A survey of regularization strategies for deep models," Artif Intell Rev, vol. 53, no. 6, pp. 3947–3986, Aug. 2020, doi: 10.1007/s10462-019-09784-7.
[24] J. Kukačka, V. Golkov, and D. Cremers, "Regularization for Deep Learning: A Taxonomy," Oct. 2017, [Online]. Available: http://arxiv.org/abs/1710.10686
[25] H. Noh, T. You, J. Mun, and B. Han, "Regularizing Deep Neural Networks by Noise: Its Interpretation and Optimization," in 31st Conference on Neural Information Processing Systems (NIPS 2017), Long Beach, CA, USA.: Curran Associates Inc., 2017, pp. 5115–5124.
[26] E. W. J. Z. K. Leslie Rice, "Overfitting in adversarially robust deep learning," in 37th International Conference on Machine Learning (ICML'20), JMLR.org, Jul. 2020, pp. 8093–8104.
[27] T. DeVries and G. W. Taylor, "Improved Regularization of Convolutional Neural Networks with Cutout," Aug. 2017, [Online]. Available: http://arxiv.org/abs/1708.04552
[28] A. M. Hamad Alhussainy and A. D. Jasim, "ECG SIGNAL CLASSIFICATION BASED ON DEEP LEARNING BY USING CONVOLUTIONAL NEURAL NETWORK (CNN)," Iraqi Journal of Information and Communications Technology(IJICT), vol. 3, no. 3, 2020, [Online]. Available: https://ijict.edu.iq
[29] D. S. Namaa et al., "Comparison between Allele frequencies of several strs loci in Najaf city of Iraq and middle province in Iraqi population," Indian Journal of Forensic Medicine and Toxicology, vol. 13, no. 4, pp. 578–583, Oct. 2019, doi: 10.5958/0973-9130.2019.00353.0.

