Very Deep Neural Networks for Extracting MITE Families Features and Classifying them based on DNA Scalograms

Mael SALAH JRAD[1], Afef ELLOUMI OUESLATI[2], Zied LACHIRI[3]
University of Tunis El Manar, SITI Laboratory, National School of Engineers of Tunis (ENIT)
BP 37, le Belvédère, 1002, Tunis Tunisia[1, 2, 3]
University of Carthage, National School of Engineers of Carthage (ENICarthage)
Electrical Engineering Department, Tunisia[2]

***Abstract*—DNA sequencing has recently generated a very large volume of data in digital format. These data can be compressed, processed and classified only by using automatic tools which have been employed in biological experiments. In this work, we are interested in the classification of particular regions in C. Elegans Genome, a recently described group of transposable elements (TE) called Miniature Inverted-repeat Transposable Elements (MITEs). We particularly focus on the four MITE families (Cele1, Cele2, Cele14, and Cele42). These elements have distinct chromosomal distribution patterns and specific number conserved on the six autosomes of C. Elegans. Thus, it is necessary to define specific chromosomal domains and the potential relationship between MITEs and Tc / mariner elements, which makes it difficult to determine the similarities between MITES and TC classes. To solve this problem and more precisely to identify these TEs, these data are classified and compressed, in this study, using an efficient classifier model. The application of this model consists of four steps. First, the DNA sequence are mapped in a scalogram's form. Second, the characteristic motifs are extracted in order to obtain a genomic signature. Third, MITE database is randomly divided into two data sets: 70% for training and 30%for tests. Finally, these scalograms are classified using Transfer Learning Approach based on pre-trained models like VGGNet. The introduced model is efficient as it achieved the highest accuracy rates thanks to the recognition of the correct characteristic patterns and the overall accuracy rate reached 97.11% for these TEs samples classification. Our approach allowed also classifying and identifying the MITES Classes compared to the TC class despite their strong similarity. By extracting the features and the characteristic patterns, the volume of massive data was considerably reduced.**

Keywords—DNA scalograms; genomic signature; classification; deep learning; transfer learning; VGGNET; accuracy

I. Introduction

DNA is a molecule composed of a long chain of four nucleotides: Adenine (A), Thymine (T), Cytosine (C) and Guanine (G) [1, 2]. It comprises a multitude of periodic structures; the majority of which have an unknown biological function. This molecule adopts a three-dimensional double-helix having a curve shape [3]. In our work, the character's string was mapped into a scalogram form, based on wavelet transform applied on a signal extracted from experimental measurements of the DNA curve. From these scalograms, we extracted patterns to classify some DNA regions. We chose, as model, the Caenorhabditis Elegans organism, which is an invertebrate combining simplicity and complexity. This duality makes it the most widely used versatile model for nearly all aspects of biological and genomic research. We also investigated a recently-described ET group, called Miniature Inverted-repeat Transposable Elements (MITEs). The latter were first discovered when studying the genes of several grass species including maize [4,5], rice [6] and barley [7]. They are genomic components abundant in many species, such as green pepper [8] and Arabidopsis [9, 10], as well as in several animal genomes including Caenorhabditis elegans [11], insects [12], humans [13] and zebrafish [14]. These species represent 1 % to 2% of the total sequence of the genomes. In these MITEs, we focused on four families, which are Cele1, Cele2, Cele14 and Cele42, because they have distinct chromosomal distribution patterns. In fact, Cele14 MITEs show clustering near the autosomes' ends. In contrast, the Cele2 MITEs display an even distribution through the central autosome domains, with no evidence for clustering at the ends. These patterns complicate the classification tasks. So far, there is no model for the systematic classification of 4 MITEs family.

However, more extensive sequence relationships between the MITEs and the Tc / mariner elements were established for the first time in C. Elegans. Most MITE families of this genome share their endings (~ 20 bp to 150 bp) and their TSD sequence with, at least, one of the described Tc1 / mariner transposons in this species. The comparison of the Tc elements coding of transposase and the numerous MITE families suggests possible scenarios for the origin of MITE in the C. Elegans genome.

As the distinction between the MITE families and the "transposable elements" (TC1, TC2, TC5) [15, 16, 17, 18] is a very difficult task, we thought about creating an efficient automatic model to classify them. In this paper, we introduce a new approach to classify DNA scalograms employing VGGNET while considering these scalograms as characteristic motifs of DNA. Our proposed method started first by converting the DNA string into DNA scalograms.

Afterward, a deep learning approach, that formed a Deep Neural Network (VGGNET) [19] for the prediction of database-derived tags from the original scalogram, was used. It allowed extracting high-level abstraction of characteristics from minimal preprocessing data. An evaluation of different CNN architectures namely, ResNet [37,38,39], inceptionv3 [40,41], Mobilnet [35,36] and Xception [42] was performed. This assessment shows that transfer learning achieved top-scoring performance.

The paper is organized as follows. Section II describes the utilized materials (the MITE and Transposons families, etc.) and the applied methods (DNA coding, continuous wavelet transform and the VGGNET classification methodology). It also details the criteria considered to evaluate the model performance (Accuracy and Confusion Matrix). Section III presents the different proposed approaches applied to classify and identify the four MITEs families applying these classification techniques. Section IV presents the experiments carried out to classify the MITE classes of C. Elegans and discusses the obtained results. Finally, Section V presents some concluding remarks.

II. Materials and Method

A. Materials

In this study, we focus on Caenorhabditis Elegans as an invertebrate combining simplicity and complexity, which makes it efficiently used to examine the important biological processes relevant to all eukaryotes. C. Elegans sequences were extracted from the National Center for Biotechnology Information (NCBI) public database [20]. Two sets of genomic data (the MITE dataset, composed of Cele1, Cele2, Cele14, and Cele42 [12], and non-ITE sequences which are the TEs TC1, TC2 and TC5 [15, 16, 17, 18]) are considered.

The MITEs are small non-autonomous elements derived from transposons. Their identification is usually based on the presence of target site duplications and terminal inverted repeats [10,11,12]. These elements are structurally comparable to defective class II. They are characterized by their small size (usually varying between 100bp and 458 bp in length) and their lack of coding capacity for transposase. They carry Terminal Inverted Repeats (TIR) and two adjacent short direct repeats called Target Site Duplications (TSD). MITEs are often located near or within genes, where they can affect gene expression [13,14]. They are preferentially located in single or weak copy regions. Thus, they can be used as genetic markers, especially for large genomes with low gene content [21]. MITEs can be grouped into super-families based on their association with TEs because they have almost the same TIRs. A relation between a given MITE family and its potential source of transposase is often based on limited sequence similarity in TIRs. The choice of a given family of TC as a non-MITE is justified by the fact that MITEs themselves contain TC sequences; which increases considerably MITE recognition rates in most bioinformatics tools. The studied MITE families are: CELE 1, CELE 2, CELE 14 and CELE 42. They have complex and variable structures and sizes.

Our database is composed of 7862 MITEs elements whose frequency occurrence in the C. Elegans genome of varies from 20 to 458, according to the class of family they belong to (Table I). The variability of length, composition and structure of these regions complicate their identification. Table I shows that MITEs have also a non-uniform distribution in the chromosomes. In fact, chromosome I (Chr I) contains the largest number of MITEs which is equal to 1799 with a size varying between 29 base pair(pb) and 380pb. Table I also demonstrate a high variability characterizing the sequences of MITE family; hence it is challenging to introduce an automated algorithm to predict them. Table II reveals that the sequences of Transposon family (TC1, TC2, TC5) are completely different. Although TC1, TC2 and TC5 are structurally characterized by more reduced numbers, they have big sizes (usually varying between 12 and 2088 bp in length).

N_{Occ} is the number of occurrences of a class in 6 chromosomes of C. Elegans, and $S_{\mathrm{min-max}}$ represents the range of the minimum and maximum sizes and occurrences of a class in 6 chromosomes of C. Elegans.

In this research work, DNA scalograms are used to characterize these regions and transfer learning is applied to classify them.

TABLE I. Number of Occurrences of the Four MITEs Families in 6 Chromosomes of C. Elegans and Size of the Scalograms of each Class of these Families

		CELE1	CELE2	CELE14	CELE42
Chr I	N_{Occ}	509	643	761	336
	$S_{min-max}$	[32-371]	[36-380]	[29-201]	[33-251]
Chr II	N_{Occ}	148	578	438	101
	$S_{min-max}$	[45-372]	[34-367]	[30-202]	[34-247]
Chr III	N_{Occ}	362	714	429	179
	$S_{min-max}$	[22-382]	[37-363]	[30-207]	[30-245]
Chr IV	N_{Occ}	179	430	360	129
	$S_{min-max}$	[49-373]	[38-379]	[43-445]	[37-251]
Chr V	N_{Occ}	366	394	677	178
	$S_{min-max}$	[20-458]	[36-363]	[34-225]	[47-273]
ChrX	N_{Occ}	16	56	268	11
	$S_{min-max}$	[74-301]	[40-317]	[40-191]	[54-237]
Total occurrence		1180	2815	2933	934
Total		**7862**			

TABLE II. Number of Occurrences of Transposan Families in 6 Chromosomes of C. Elegans and Size of the Scalograms of each Class of these Families

		TC1	TC2	TC5
Chr I	N_{Occ}	36	24	29
	$S_{min-max}$	[50-1610]	[53-230]	[54-1606]
Chr II	N_{Occ}	42	19	26
	$S_{min-max}$	[62-1611]	[47-2074]	[24-1611]
Chr III	N_{Occ}	21	15	22
	$S_{min-max}$	[111-1610]	[61-157]	[12-1608]
Chr IV	N_{Occ}	23	31	34
	$S_{min-max}$	[15-1610]	[12-154]	[35-844]
Chr V	N_{Occ}	91	33	62
	$S_{min-max}$	[64-1611]	[39-2088]	[28-1611]
ChrX	N_{Occ}	83	52	28
	$S_{min-max}$	[33-1610]	[12-155]	[64-1631]
Total occurrence		296	174	201
Total		**671**		

In bio-informatics field, two sequences are considered homologous if they come from a common ancestor. Multiple sequence alignment techniques allow specifying the homologous regions of each sequence. Fig. 1, shows that the DNA scalograms highlight DNA homology, a degree of identity or similarity, between scalograms of different regions of CELE2 and similar homology for different elements of TC2. It also reveals a slight difference between the CELE2 and TC2 images [12,21].

B. Methods

To classify the MITE families, it is necessary to parameterize the DNA sequences regardless to their heterogeneity. Thus, we choose the DNA mapping into image based on scalograms. For this reason, we use PNUC coding technique [22, 23] and Continuous Wavelet Transform (CWT) [24, 25, 26] to highlight features. Then, we extract these features from DNA images using VGGNet, a powerful CNN architecture, pre-trained on ImageNet. Finally, a classification is performed based on deep learning model (VGG19 and VGG16) [19, 27].

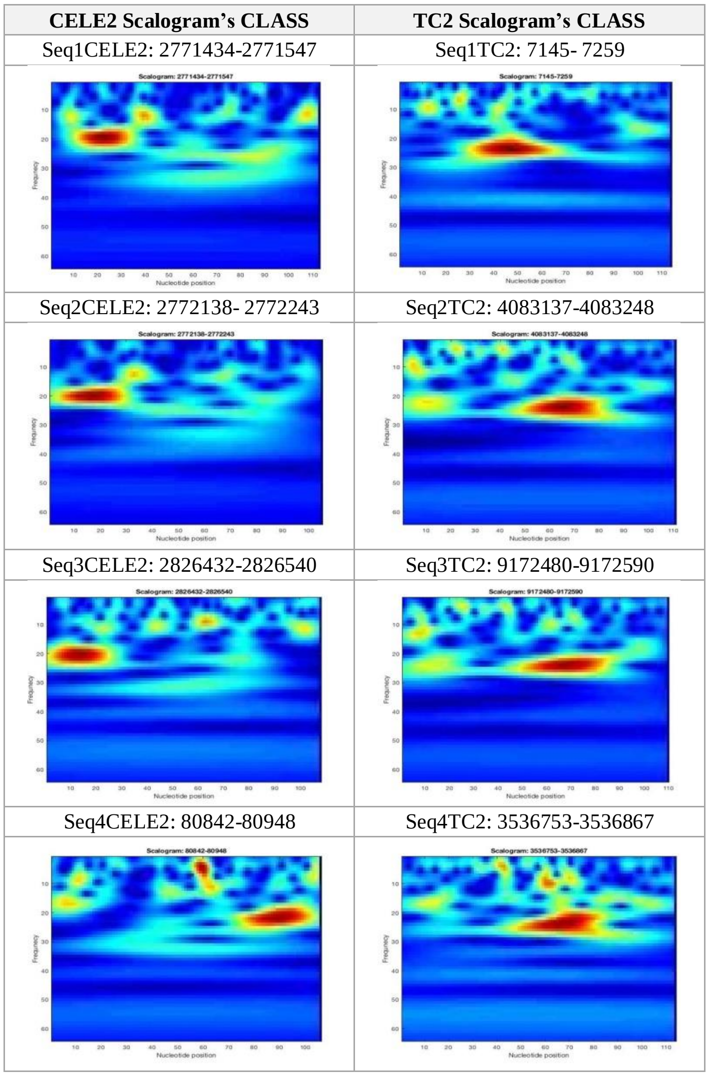

Fig. 1. Samples taken from the CELE2 Scalogram Class and TC2 Scalogram Class.

Transfer Learning consists first in training a base network on a dataset and then transferring the learned features to a second target network to train them to a target dataset.

1) PNUC coding technique and Wavelet Transform: For our classification technique, we consider DNA images. These images represent scalograms which are energy distributions obtained by taking the square module of the continuous wavelet transform applied to sequences encoded in PNUC [22, 23]. Considering the square module, time-frequency localization is enhanced, and a new database of DNA scalograms is generated.

PNUC coding is based on curvature measurements. This curvature is directly related to the nucleosome structures presence. Applying this technique, the pairing of the two DNA helix (A-T and CG) along the helix is taken into account. PNUC coding consists in assigning, to each codon or trinucleotide, the numerical value given by the experimental values associated with each codon [23].

For example, the S_{DNA} is replaced by the numerical sequence.

S_{DNA}= 'AAG TTT CTT GTG AAA ACG TGC AGC'

The Pnuc coding of S_{DNA} is :

CS_{DNA} = '7.3 0 7.3 9.2 0 7.6 8.5 1'

DNA has a multitude of periodic structures and the wavelet analysis was proposed to reveal the local and frequential properties of the DNA periodic motifs. The analysis based on the Morlet Complex wavelet allows detecting the different periodicities in various types of C. Elegans chromosomal DNA [24, 25, 26].

Wavelet analysis relies essentially on the signal's decomposition into a sum of time-frequency atoms. The latter, called "wavelets", are obtained by dilating or contracting a Mother Wavelet ψ (t) [28, 29] and translatin g it along the time axis. The versions obtained after these transformations are noted ψ [(t-b) / a].

The dilation and compression of a mother wavelet depend on a scaling factor (a), while the translation is ensured using a translation parameter (b). The wavelet family of scales and positions is then generated by the following expression:

$$\psi_{a,b}(t) = \frac{1}{\sqrt{a}}\psi^*\left(\frac{t-b}{a}\right), a \succ 0, b \in \mathbb{R} \quad (1)$$

In general, the wavelet transform of a signal f(t) is given by Equation (2):

$$T_\psi(f)(a,b) = \frac{1}{\sqrt{a}}\int_{-\infty}^{\infty} f(t)\psi^*\left(\frac{t-b}{a}\right)dt \quad (2)$$

where the symbol * indicates the complex conjugate. The obtained Tψ (a, b) numbers are called coefficients of wavelets. The Morlet Complex wavelet is the most efficient technique applied to analyze and characterize DNA structures [24] and presented as an exponential-modulated Gaussian envelope. It is defined by the following equation:

$$\psi(t) = \pi^{-\frac{1}{4}}\left(e^{i\omega_0 t} - e^{-\frac{1}{2}\omega_0{}^2}\right)e^{-\frac{1}{2}t^2} \quad (3)$$

where the parameter ω0 designates the number of oscillations of the mother wavelet. It must be greater than 5 to satisfy the eligibility requirement. Using M scales, we obtain a matrix of N × M coefficients representing the time-frequency plane where N is the length of the analyzed signal. The modulus of wavelet coefficients | Tψ (a, b) | is called "scalogram".

In our study, the performed analysis consists in applying a continuous wavelet transform based on the complex Morlet wavelet [24, 25, 26]. This analysis highlights the periodicities that reside in the DNA (represented by its inverse the frequency on the y axis) with precision on location in the nucleotides position (equivalent to time in the x axes). The result of analysis generates scalograms which are the images used in the classification.

2) VGGNET model for classification of DNA scalograms: We are interested in studying the VGGNET which is a convolutional neural network trained on more than one million images from the ImageNet database [30].

We use transfer learning to classify the DNA scalograms. The main idea of transfer learning based on very deep neural networks is to apply a pre-trained deep learning model, previously trained on a large-scale dataset such as ImageNet. Containing 1.2 million images with another 50,000 images for validation and 100,000 images for testing, on 1000 different categories, and re-purpose, to handle an entirely different problem [31].

The used model treats the input image. Then, it outputs the vector containing 1000 values. This vector represents the corresponding class classification probability. If a model is utilized to predict that an image belongs to class 0, class 1, class 2, class 3, class 780, class 999 with probability 1, 0.05, 0.05, 0.03, 0.72, 0.05, respectively and the remaining classes with probability 0, the classification vector of this model will be:

$$\hat{y} = \begin{bmatrix} \widehat{y_0} = 0.1 \\ 0.05 \\ 0.05 \\ 0.03 \\ . \\ . \\ . \\ \widehat{y_{780}} = 0.72 \\ . \\ . \\ \widehat{y_{999}} = 0.05 \end{bmatrix}$$

Softmax function, defined below, is used to ensure that these probabilities add to 1:

$$P\left(y = j \middle| \theta^{(i)}\right) = \frac{e^{\theta^{(i)}}}{\sum_{j=0}^{k} e^{\theta_k^{(i)}}} \qquad (4)$$

where:

$$\theta = w_0\, x_0 + \, w_1\, x_1 + \cdots + w_k\, x_k = \sum_{i=0}^{k} w_i x_i = w^T x$$

After learning certain features from a large dataset (ImageNET), they are used by VGGNet model as a base to learn the presented classification problem. As demonstrated in Fig. 2, we employ a popular and reliable CNN architecture called VGGNet with 16 convolutional and 3 fully-connected layers [27]. The width of convolutional layers (the number of channels) is rather small, starting from 64, in the first layer, and increasing by a factor of 2, after each max-pooling layer, up to 512. The input of the CNN is a fixed-size 224 x 224 RGB image. Each image passes through a stack of convolutional (conv.) layers. Subsequently, the convolution stride is added such that the spatial resolution will be preserved after convolution, i.e. the padding is considered also in Conv. layers. Spatial pooling is carried out by five max-pooling layers, which follow some but not all of the Conv. Layers. Max-pooling is performed over a specific pixel window; with stride. A stack of Conv. Layers, having different depths in various architectures, are followed by three fully-connected (FC) layers: each of the two first layers has 4096 channels, while the third one performs the classification of 2 after each max-pooling layer, up to 512 [32,33].

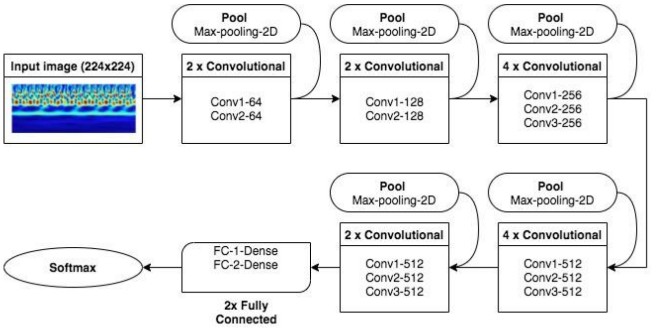

Fig. 2. Overview of the VGG16 Layer Structure (Left) and Corresponding Parameters (Right).

III. Proposed Approach

The adopted methodology includes three steps. Fig. 3 represents the flowchart describing our proposed approach whose application consists in:

- Extracting the MITEs sequences (CELE1, CELE2) and the TEs (TC1, TC2 and TC5) of all the chromosomes of C. Elegans from the NCBI database. The extraction phase can be divide into the following two sub-steps:
    - Generating the corresponding PNUC sequences to convert the DNA string into a 1D signal.
    - Applying Continuous Wavelet analysis to transform the signal to scalogram images.
- Extracting features using convolutional neural networks
- Using the VGGNET model to classify the studied sequences.

A. Creating MITE Signal Database

In the first step of our methodology, we extract the entire DNA sequences corresponding to the C. Elegans genome from the NCBI database [20]. Then, we apply PNUC coding on all chromosomes (6 chromosomes). Thereby, a chromosomal signal database (1D signal) is created after applying the module on the square of the continuous wavelet transform [24, 25, 26], which enhances time-frequency localization and generates a new DNA database of DNA containing images (or scalograms) which represent energy distributions).

B. Extraction of Features using Convolutional Neural Networks

Several models were used to extract the characteristics of AND scalograms in the field of deep learning. In this work, we use a model adapted for their extraction (Fig. 4). These Different values of independent variables are also considered as the input of the classifier to predict the corresponding class to which the independent variable belongs. The architecture of introduced model is presented in Table III.

As shown in Fig. 4, the shape of the input image is (224, 224, 3) and the last layer produced from VGGNet has the shape (7, 7, 512). This means that VGGNet returns a feature vector of 7×7×512 = 25088 features. In order to perform transfer learning with VGGNet, we first save the extracted features (bottleneck features) from the pre-trained model. Then, top model is trained to classify our data using the saved bottleneck features. Finally, we combine our training data and the VGGNet model with the top model to predict DNA Pattern of scalograms [34].

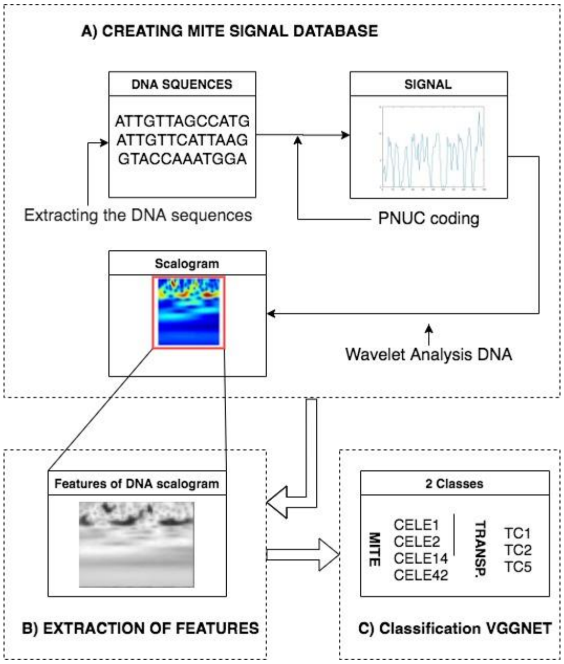

Fig. 3. Flow Chart of the Proposed Approach.

TABLE III. CONVOLUTIONAL NEURAL NETWORK ARCHITECTURE

Feature	value
Convolution layer	3x3
Max pooling layer	2x2
Convolution stride	1 pixel
Padding	2 pixel
Rectification	ReLU
Fac layer	softmax
Fac layer nodes	4096
Total layers	19 Layers

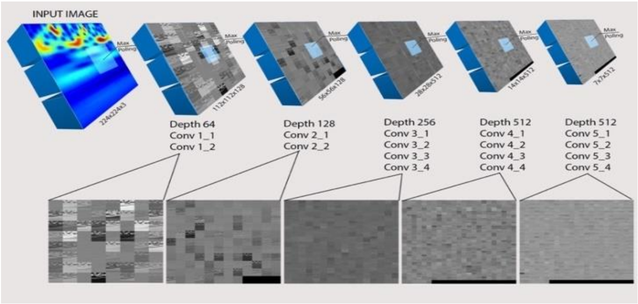

Fig. 4. Feature Extraction by VGG19 Model and Corresponding Parameters.

C. Classification Algorithm

For our classification algorithm, we use the third convolutional layer containing only two channels (one for each class). The final layer is the soft-max layer. All hidden layers are equipped with the non-linearity rectification [19,27]. For each image X of study type T in the training set, the weighted binary cross-entropy loss is optimized. The VGGNet specifications are described in Fig. 3.

The major limitation of VGGNet lays in the fact that this architecture necessitates huge memory requirements. Because of the number of fully-linked nodes and its depth, the size of VGGNet is equal to 574 MB, which complicates its use as features extractor.

We also employ the VGG16 and compare its results with those provided by the VGG19. The VGG-16 is a 16-layer CNN developed by Simon et al. for image recognition in the 2014 ImageNet large scale visual recognition challenge (ILSVRC) [19]. The filters 3 × 3 are employed for all convolutional layers. This network accepts the input image with a dimension of 224 × 224. The image passes through a sequence of 16 convolutional layers. A multilayer perceptron (MLP) classifier, including three fully connected (FC) layers and the convolutional layers, is utilized in the classification step. The Rectified linear unit (ReLU) layers and max-pooling layers are also used in the whole network to prevent overfitting.

To evaluate our classification model, we apply the classification rate calculation and the confusion matrix as classification criteria. The performance of the proposed approach is tested in terms of accuracy, recall, precision, sensitivity, specificity, F-measure (F1), Confusion matrix illustrated in Fig. 5 and loss functions value to select features of DNA scalograms. These measures are described below:

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \tag{5}$$

$$\mathrm{Recall} = \frac{\mathrm{TP}}{\mathrm{TP+FN}} \tag{6}$$

$$Precision = \frac{TP}{TP+FP} \tag{7}$$

$$Sensitivity = \frac{TP}{TP+TN} \tag{8}$$

$$Specificity = \frac{TN}{TN+FP} \tag{9}$$

$$F1 = 2 \times \frac{Specificity \times Sensitivity}{Specificity+Sensitivity} \tag{10}$$

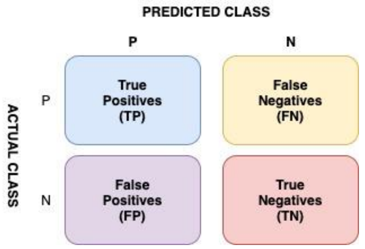

Fig. 5. Confusion Matrix.

where "TP" (True Positives) refers to the CELE samples correctly labeled by the classifier, "TNs" (True Negatives) are the Transposon samples correctly labeled by the classifier, "FPs" (False Positives) denotes the CELE scalograms incorrectly labeled as Transposons TC and "FNs" (False Negatives) are the transposon samples mislabeled as CELE.

The two most crucial and most intensively-employed loss functions are: the cross entropy function and the MSE function. Both of them are applied in regression and classification problems, respectively. The can be formulated as follows:

$$\mathcal{L}_{MSE}(\boldsymbol{W}) = \frac{1}{n-1}\sum_{i=0}^{n-1}(y_i-\hat{y}_i)^2 \quad (11)$$

$$\mathcal{L}_{\text{cross-entropy}}(\mathbf{W})=-\sum_{i=0}^{n-1}\sum_{c=0}^{M} y_{i,c}\log\widehat{y_{i,c}} \quad (12)$$

where n is the total number of samples in the dataset, M denotes the number of classes with in the dataset, $y_{i,c}$ designates a binary indicator indicating if class c represents the correct classification for sample i and $\widehat{y_{i,c}}$ refers to the predicted probability of sample i which belongs to class c. The first previously-mentioned loss produces high loss when the predicted value is close to the true value; whereas the cross-entropy loss punishes uncertain prediction probabilities.

IV. Results

The main objectives of this study are to characterize MITE families and distinguish them from other regions. For this purpose, genomic sequences used in this work are composed of two parts: a part containing MITE family sequences (CELE1, CELE2, CELE14 and CELE42) and another one including TC1, TC2, TC5.

The examination of the structure and distribution of MITEs reveals that the number of appearances of these elements are variable and that the TC family number of scalograms is reduced compared to them, which complicates the MITE classification process. To solve this problem of unbalanced data, we enlarge the database of TC and MITEs signals by grouping all the elements of the TC family in the same class and applying a binary classification. Here, the idea is based on the identification of the MITE family of elements (CELE1 CELE2 CELE14 and CELE42) with respect to other non-MITE families (Tc1, Tc2, Tc5). Thereafter, this dataset is split into two parts 70% for training and 30% for test). Then, VGGNET is applied with softmax activation mode. The Recognition process consists of two stages: features extraction and features recognition. The performance of the proposed system strongly depends on the choice of the extraction method.

The experimental results demonstrate that most of the CELE elements are correctly recognized with the Tc elements. Obviously, the VGG16 trained model achieves an accuracy rate of 97.11% for CELE14 identification, 93.38% for CELE1 identification, 91.79% for CELE42 identification and 89.66% for CELE2 identification. Fig. 6 illustrates the accuracy of the VGG-16 model over the Test images.

Similarly, Fig. 8 illustrates the accuracy rate obtained by applying the VGG-19 model on the validation dataset. The trained model reaches an accuracy rate equal to 96.44%, 94.52 %, 91.05 and 90.17 for the identification of CELE14, CELE42, CELE1 and CELE2, respectively.

Fig. 6, 7, 8 and 9 demonstrate that the learning and validation curves are remarkably enhanced for VGG16 and VGG19 Models. It is also clear that the network converges from the second epoch and, with the rise in the epochs, and the cross entropy loss tends to zero.

These figures represent the accuracy curves of train set and those of validation set. Each point of the precision curve corresponds to the accurate prediction rate for train or validation images. The accuracy curve follows similar smooth processing as that adopted by the loss curve. It is obvious that the train set accuracy and the validation set accuracy approach 100% after 2 epochs.

Fig. 6 and 7 show that the VGG16 model accuracy value is higher, compare to that of VGG19 model. However, this is not true depending on the element to be identified. Thus, the accuracy average is computed to classify the 4 MITE families. The accuracy rate attains 92.98, for VGG16, and 93.045, for VGG19, revealing that VGG16 is more effective in the classification of MITES scalograms, compared to VGG19 model.

Additionally, testing results are given in Fig. 10 and 11 representing the Confusion Matrix for the validation data.

The performance measurement is with four different combinations of predicted and target classes which are the true positive, false positive, false negative, and the true negative. In this format, the number and percentage of the correct classifications performed by the trained network are indicated in the diagonal.

The confusion matrix shows that the used models clearly differentiate the families of MITE, compared to TC1, TC2, TC5, despite the similarities between the CELE and TC1, TC2, TC5, as cited in the first part (Section 3) of this paper [21].

As seen in Fig. 10, all the classes of MITEs are correctly classified. Our model, using VGG16, recognizes CELE2 with a very promising rate of 99.52%, and 99.19%, for CELE14 identification, 96.48%, for CELE1 identification, and 86.19% for identification of CELE42.

Comparison of our models with other CNN architectures

Similarly, a comparative analysis of the results obtained by the VGGNET framework, employing four well-known methods, was carried out to shed light on the efficiency of VGGNET in identifying the four MITE families, as given in this table (Table IV). We show that Mobilnet [35,36], Resnet [37,38,39], InceptionV3 [40,41] and Xception [42] give average accuracy rates (Acc.) of 88.85%, 86.92%, 86.23% and 88.06%, respectively, to classify the four MITE families.

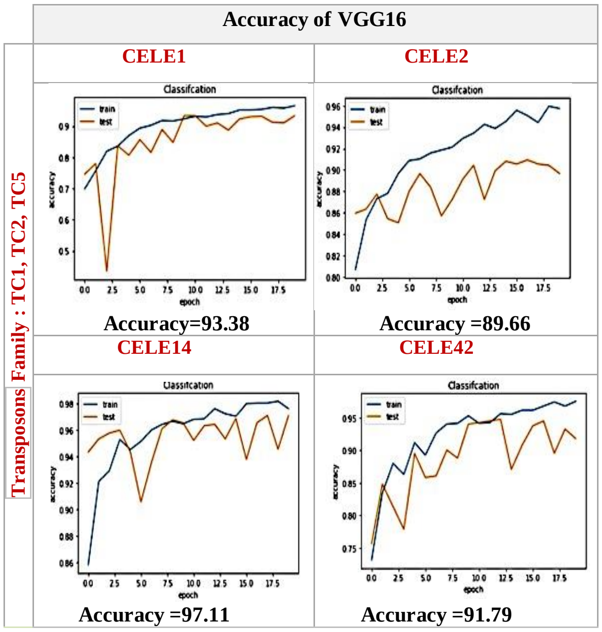

Fig. 6. Accuracy Rate Obtained for the Identification of Mites Families to Transposon Families DNA Scalograms VGG16.

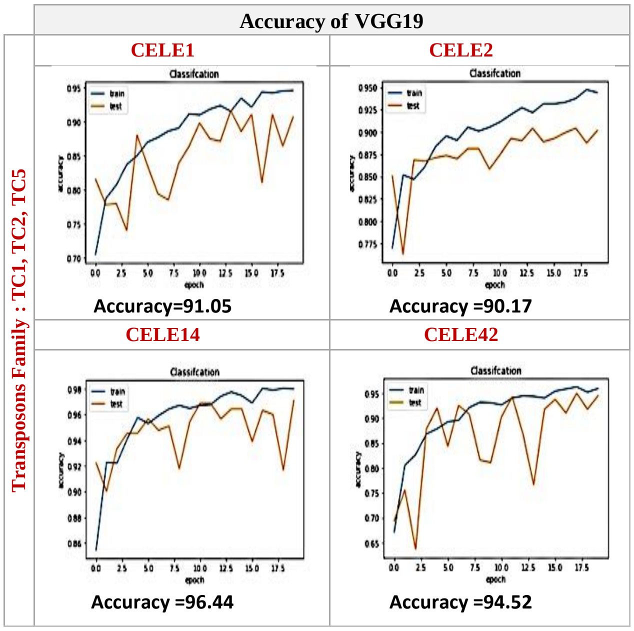

Fig. 7. Accuracy Rate Obtained for the Identification of Mite Family to Transposon Family DNA Scalograms VGG19.

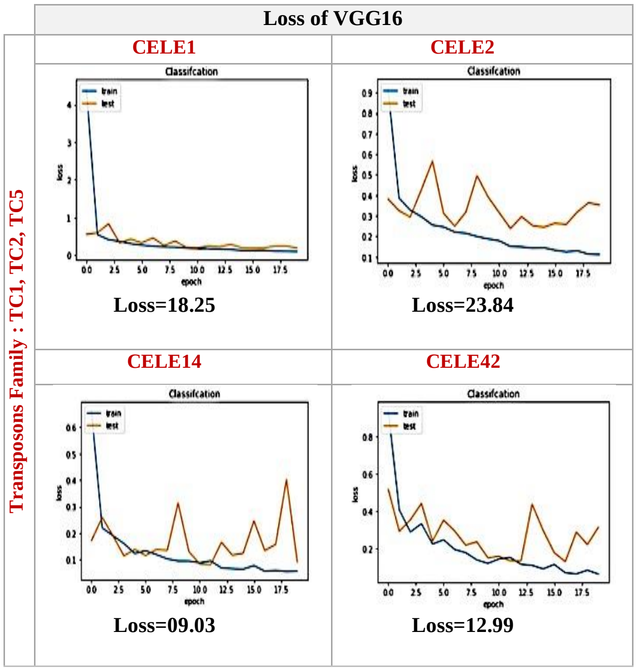

Fig. 8. Loss for the Identification of Mite Family to Transposon Family DNA Scalograms VGG16.

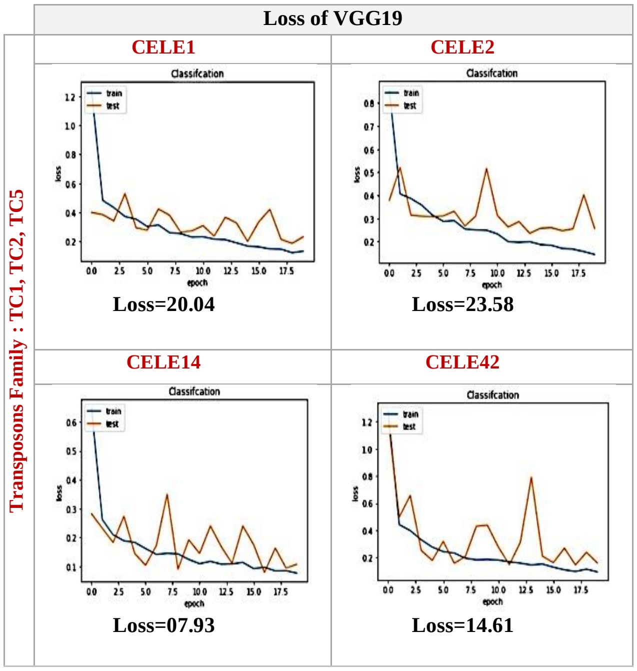

Fig. 9. Loss for the Identification of Mite Family to Transposon Family DNA Scalograms VGG19.

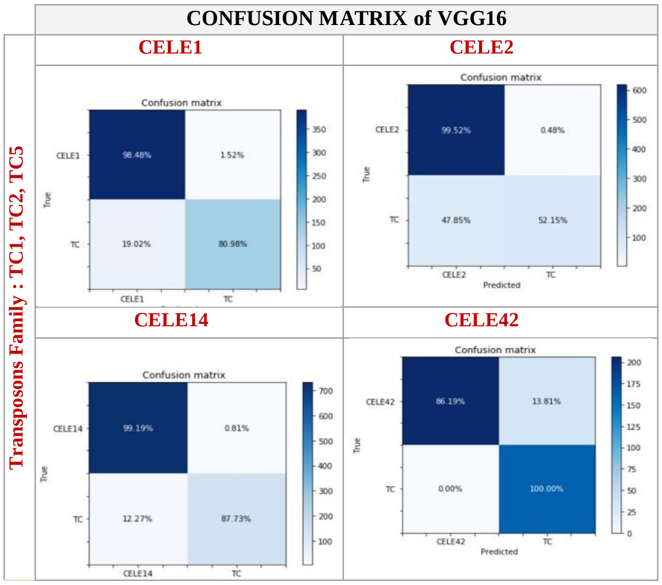

Fig. 10. Confusion Matrix for the Identification of MITE Families to Transposon Family of DNA Scalograms VGG16.

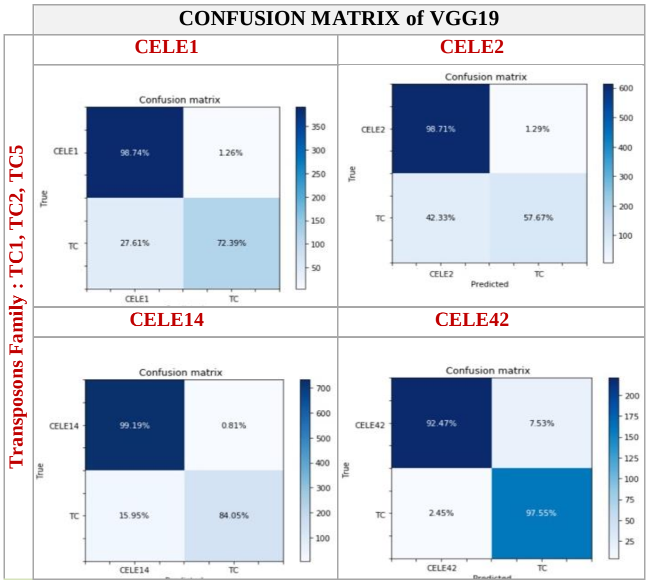

Fig. 11. Confusion Matrix for the Identification of MITE Families to Transposon Family of DNA Scalograms VGG19.

However, findings obtained using the VGGNET provide the highest accuracy of 97.11% for the identification of the CELE1 element. These results reveal that our architectures more powerful and promising than other deep models

Table IV demonstrates also the performance achieved by other CNN architectures [43,44] using Loss, Recall (Rec.), Precision (Pre.), sensitivity (Sens.), specificity (Spec.) and F1 metrics.

The proposed approach shows high performance by achieving accuracy, loss, recall, precision and f1-score of 97.11, 09.03, 99.18, 97.34, and 85.64, respectively.

TABLE IV. SPLATTING BASED COMPARATIVE WITH OTHER CNN ARCHITECTURE TO IDENTIFY CELE 1, CELE2, CELE14 AND CELE42

CELE1	AcC.	LosS	REc	Pre.	Sen.	Spec.	F1
VGG16	**93,38**	**18.25**	98.48	**92.63**	74.71	80.98	**77.71**
VGG19	91.05	20.04	98.73	89.67	76.81	72.39	74.53
Resnet152	83.89	35.92	97.97	82.55	82.72	49.69	62.08
Mobilnet	84.97	33.67	85.60	92.62	71.36	**83.43**	76.92
InCEPtionV3	85.86	34.26	**99.24**	83.80	**83.79**	53.37	65.2
Xception	88.01	29.28	95.20	88.70	76.62	70.55	73.45

CELE2

CELE2	AcC.	LosS	REc	Pre.	Sen.	Spec.	F1
VGG16	89.66	23.84	99.51	88.79	87.90	52.14	65.45
VGG19	**90.17**	**23.58**	98.71	89.88	86.88	57.66	69.25
Resnet152	84.69	34.32	100	83.80	**93.52**	26.38	41.15
Mobilnet	86.47	20.578	**100**	85.41	91.59	34.96	50.60
InCEPtionV3	86.66	34.66	93.23	**90.04**	85.65	**59.50**	**70.21**
Xception	87.11	33.97	98.71	86.82	89.75	42.94	58.08

CELE14

CELE14	AcC.	LosS	REc	Pre.	Sen.	Spec.	F1
VGG16	**97.11**	09.03	99.18	97.34	83.65	**87.73**	**85.64**
VGG19	96.44	**07.93**	99.18	96.56	84.23	84.04	84.13
Resnet152	93.56	14.27	99.72	92.92	87.30	65.64	74.93
Mobilnet	90.45	12.536	**100**	89.56	**90.55**	47.23	62.07
InCEPtionV3	91.56	21.28	99.45	91.06	88.96	55.82	68.59
Xception	94.56	15.30	99.18	**99.45**	85.91	73.61	79.28

CELE42

CELE42	AcC.	LosS	REc	Pre.	Sen.	Spec.	F1
VGG16	91.79	**12.99**	86.19	**100**	55.82	**100**	71.64
VGG19	**94.52**	14.61	92.46	98.22	58.15	97.54	**72.86**
Resnet152	85.57	26.66	98.32	81.13	68.31	66.87	67.58
Mobilnet	93.53	33.62	98.32	91.43	62.50	86.50	72.56
InCEPtionV3	80.84	26.93	**99.58**	75.79	**73.23**	53.37	57.83
Xception	82.58	28.48	71.54	98.84	51.50	98.77	67.70

Table V presents the existing works based on supervised machine learning algorithms used in the classification step and compares the results obtained employing DNA sequences database. As shown in this table, several studies utilized CNN (Convolutional neural networks) [47], C-KNN [45] and support vector machines utilized [46].

Nevertheless, successful categorization rate ranges between 70% and 90%. Obviously, a successful categorization relies mainly on the entry variability. The majority of the studies listed in Table V and those based on the machine learning dealt with correctly detecting and classifying the highest number of defects applying features extractions and classifiers. Measures reported in the literature (classification, accuracy, number of features and computation time) are also compared. They demonstrate that transferred VGGNET models attain the highest accuracy rate. The major benefit of pre-trained VGGNET models, compared to those applied in the existing studies, lays in the fact that they do not necessitate a feature extraction mechanism or an intermediate feature selection phase.

TABLE V. COMPARISON OF THE TRANSFER LEARNING-BASED VGGNET MODELS WITH THE EXISTING WORKS

Study	Method	Accuracy
[43] Nguyen, N.G. **[44] Amerah Kassim**	DNA Sequence Classification by CNN	82 %
[45] Mochammad Anshori	LDA-SVM	92.7%
[46]Alhersh	C-KNN	73.72 % -91.82%
Proposed Approch		97.11%

V. CONCLUSION

In this paper, we focused on DNA images. Our main purpose is to identify the MITES Families from Transposan families and classify them. The DNA images represent the scalograms. In fact, the ATCG chain was first converted, using a PNUC coding technique, into a signal based on the experimental DNA curve measures. Then, the Continuous Wavelet Transform by Morlet Complex wavelet allowed converting this signal into particular images. Thirdly, a selection of features was performed applying Transfer learning approach. Finally, each produced feature set was tested by several classifiers to validate the proposed model.

This approach showed high performance by achieving accuracy, loss, recall, precision and f1-score of 97.11%, 09.03, 99.18, 97.34, and 85.64, respectively. The obtained results are the highest among all known published works on the same dataset, even if compared to other convolutional network models. In fact, the classification rate obtained in previous works did not exceed 90%.

REFERENCES

[1] R. B. Macgregor, and G. M. Poon, "The DNA double helix fifty years on.Computational biology and chemistry", 27(4), pp.461-467, 2003.

[2] F. Rechenmann, and C. Gautier,"Interpreting the genome. La recherche", (332), pp.39-45, 2000.

[3] Craig NL, Craigie R, Gellert M, Lambowitz AM. Mobile DNA II. Washington, DC: Am. Soc. Microbiol. Press, 2002.

[4] Bureau, T. E. & Wessler, S. R. Plant Cell 4, pp.1283–1294, 1992.

[5] Bureau, T. E. &Wessler, S. R. Proc. Natl. Acad. Sci. USA 91, pp.1411–1415, 1994.

[6] Bureau, T. E., Ronald, P. C. &Wessler, S. R. Proc. Natl. Acad. Sci. USA , 93, pp.8524–8529, 1996.

[7] Shirasu, K., Schulman, A. H., Lahaye, T. & Schulze-Lefert, P. Genome Res. 10, pp.908–915, 2000.

[8] Pozueta-Romero, J., Houlne, G. & Schantz, R.Gene 171, pp. 147–153, 1996.

[9] Casacuberta, E., Casacuberta, J. M., Puigdomenech, P. & Monfort, A., Plant J. 16, pp.79–85, 1998.

[10] Surzycki, S. A. & Belknap, W. R. J. Mol. Evol. 48, pp. 684–691, 1991.

[11] Oosumi, T., Belknap, W. R. & Garlick, B. Nature (London) 378, 672, 1995.

[12] Surzycki, S. A., and W. R. Belknap, Repetitive-DNA elements are similarly distributed on Caenorhabditis elegans autosomes.Proc. Natl. Acad. Sci. USA 97: pp.245–249, 2000.

[13] Tu, Z. Proc. Natl. Acad. Sci. USA 94, pp.7475–7480, 1997.

[14] Izsvak, Z., Ivics, Z., Shimoda, N., Mohn, D., Okamoto, H. & Hackett, P. B. J. Mol. Evol. 48, pp.13–21, 1999.

[15] Mello CC, Kramer JM, Stinchcomb D, Ambros V. Efficient genetransfer in C. elegans: extrachromosomal maintenance and integration of transformingsequences. Embo J ; 10 : 959-70, 1991.

[16] Duret, L., G. Marais & C. Biemont. Transposons but not retrotransposons are located preferentially in regions of high recombination rate in Caenorhabditis elegans. Genetics 156: pp. 1661–1669, 2000.

[17] Kidwell MG. Transposable elements and the evolution of genome size in eukaryotes. Genetica 115: pp. 49–63, 2002.

[18] Feschotte C, Pritham EJ. DNA transposons and the evolution of eukaryotic genomes. Annu Rev Genet 41: pp.331–368, 2007.

[19] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. In ICLR, 2015.

[20] The NCBI GenBank database. [Online]. Available: http://www.ncbi.nlm.nih.gov/Genbank/. Accessed 15 Sept 2005.

[21] Casa, Alexandra M., et al. "The MITE family Heartbreaker (Hbr): molecular markers in maize." Proceedings of the National Academy of Sciences 97.18: pp.10083-10089, 2000.

[22] AE Oueslati, Messaoudi I, Z Lachiri, N Ellouze, ed. by SalihSalih Dr, Spectral analysis of global behaviour of C. elegans chromosomes, in Fourier Transform Applications, pp. 205–228, 2012.

[23] Oueslati, A. E., Messaoudi, I., Lachiri, Z., & Ellouze, N. A new way to visualize DNA's base succession: the Caenorhabditis elegans chromosome landscapes. Medical & biological engineering & computing, 53(11), pp.1165-1176, 2015.

[24] I Messaoudi, A Elloumi, Z Lachiri, in C. elegans, International Conference on Control, Engineering & Information Technology (CEIT2013), vol. 3,. Complex Morlet wavelet analysis of the DNA frequency chaos game signal and revealing specific motifs of introns (Sousse, Tunisia,4–7June), pp. 27–32, 2013.

[25] Messaoudi, I., Oueslati, A. E., & Lachiri, Z. Revealing Helitron signatures in Caenorhabditis elegans by the Complex Morlet Analysis based on the Frequency Chaos Game Signals. In IWBBIO (pp. 1434-1444), 2014.

[26] Messaoudi, I., Oueslati, A. E., & Lachiri, Z. Wavelet analysis of frequency chaos game signal: a time frequency signature of the C. elegans DNA. EURASIP Journal on Bioinformatics and Systems Biology, 2014(1), 16.

[27] B. Liu, X. Zhang, Z. Gao, and L. Chen, "Weld defect images classification with vgg16-based neural network," in Proceedings of the International Forum on Digital TV and Wireless Multimedia Communications, pp. pp.215–223, 2017.

[28] NAJMI AH, SADOWSKY J, The continuous wavelet transform and variable resolution timefrequency analysis. Johns Hopkins APL Tech Dig 18:1pp. 34–140,1997.

[29] Ngui, Wai Keng, et al. "Wavelet analysis: mother wavelet selection methods." Applied mechanics and materials. Vol. 393. Trans Tech Publications, 2013.

[30] Russakovsky, Olga, et al. "Imagenet large scale visual recognition challenge." International journal of computer vision 115.3 : pp.211-252, 2015.

[31] Chen, L.-C., Papandreou, G., Kokkinos, I., Murphy, K. &Yuille, A. L. Deep lab:Semantic image segmentation with 290 deep convolutional nets, a trous convolution, and fullyconnectedcrfs. IEEE Transactions on pattern analysis machine 291 intelligence 40, pp.834–848 , 2017.

[32] Sahlol, Ahmed T., Philip Kollmannsberger, and Ahmed A. Ewees. "Efficient classification of white blood cell leukemia with improved Swarm optimization of deep features." Scientific Reports 10.1 (2020): 1-11.

[33] Busia, Akosua, et al. "A deep learning approach to pattern recognition for short DNA sequences." *BioRxiv* : pp.353474, 2019.

[34] Bengio, Y., Courville, A. & Vincent, P. Representation learning: A review and new perspectives. IEEE Transactions on 249 pattern analysis machine intelligence 35, pp.1798–1828, 2013.

[35] A. G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, and H. Adam. Mobilenets: Efficient convolutional neural networks for mobile vision applications, 2017.

[36] Z. Qin, Z. Zhang, X. Chen, and Y. Peng, “FD-MobileNet: Improved MobileNet with a Fast Downsampling Strategy,” 2018.

[37] C. Szegedy, S. Ioffe, V. Vanhoucke, and A. A. Alemi, “Inceptionv4, inception-ResNet and the impact of residual connections on learning,” in Proceedings of the AAAI Conference on Artificial Intelligence, San Francisco, CA, USA, 2017.

[38] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In Proceedings of CVPR, pp. 770–778, 2016.

[39] S. Xie, R. Girshick, P. Dollár, Z. Tu, and K. He, “Aggregated Residual Transformations for Deep Neural Networks,” in Proc. of CVPR, 2017.

[40] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich. Going deeper with convolutions. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp.1–9, 2015.

[41] Szegedy, Christian, et al. "Rethinking the inception architecture for computer vision." Proceedings of the IEEE conference on computer vision and pattern recognition, 2016.

[42] Chollet, “Xception: Deep Learning with Depthwise Separable Convolutions,” in Proc. of CVPR, 2017.

[43] Nguyen, N.; Tran, V.; Ngo, D.; Phan, D.; Lumbanraja, F.; Faisal, M.; Abapihi, B.; Kubo, M.; Satou, K. DNA sequence classification by convolutional neural network. J. Biomed. Sci. Eng, 9, pp.280–286, 2016.

[44] Nurul Amerah Kassim1, and Dr Afnizanfaizal Abdullah2. Classification of DNA Sequences Using Convolutional Neural Network Approach.Innovations in Computing Technology and Ap.

[45] Alhersh, Taha, et al. "Species identification using part of DNA sequence: evidence from machine learning algorithms." Proceedings of the 9th EAI International Conference on Bio-inspired Information and Communications Technologies (formerly BIONETICS), 2016.

[46] Anshori, Mochammad, Wayan Firdaus Mahmudy, and Ahmad Afif Supianto. "Classification Tuberculosis DNA using LDA-SVM." Journal of Information Technology and Computer Science 4.3:pp. 233-240, 2019.

[47] Vu, Duong, Marizeth Groenewald, and Gerard Verkley. "Convolutional neural networks improve fungal classification." *Scientific reports* 10.1 : pp.1-12, 2020.