EAI Endorsed Transactions
on Pervasive Health and Technology

Research Article EAI.EU

Breast cancer early detection in TP53 SNP protein sequences based on a new Convolutional Neural Network model

Saifeddine Ben Nasr[1], Imen Messaoudi[2,*], Afef Elloumi Oueslati[3,*], Zied Lachiri[4,*]

[1]SITI Laboratory, National School of Engineers of Tunis (ENIT), University of Tunis El Manar, Tunis, Tunisia.

[2]University of Carthage, Industrial Computing Department, Higher Institute of Information Technologies and Communications (ISTIC), Tunis, Tunisia.

[3]University of Carthage, Electrical Engineering Department, National School of Engineers of Carthage (ENICarthage), Tunis, Tunisia

[4]SITI Laboratory, National School of Engineers of Tunis (ENIT), University of Tunis El Manar, Tunis, Tunisia.

Abstract

INTRODUCTION: Breast Cancer is the most commonly occurring cancer and the second leading cause for women's disease death. The breast cancer cases are associated with mutations that are inherited from older generations or acquired overtime. If the diagnosis is done at the first stage, effects associated with certain treatments can be limited, costs can be saved and the diagnostic time can be minimized. This can also help specialists target the best treatment to increase the rate of cures. Nevertheless, its discovery in patients is very challenging due to silent symptoms aside from the fact that the routine screening is not recommended for women under 40 years old.

OBJECTIVES: Several efforts are aimed at the breast cancer early detection using machine and deep learning systems. The proposed algorithms use different data types to distinguish between cancerous and non-cancerous cases; as images from mammography, ultrasound and magnetic resonance imaging. Then, different learning tools were applied on this data for the classification task. Despite the classification rates which exceed 90%, the major drawback of all these methods is that they are applicable only after the appearance of the cancerous tumors, which reduces the cure rates.

METHODS: We propose a new technique for early breast cancer screening. Here, we focus on cancerous and non-cancerous Single Nucleotide Polymorphism protein sequences of the TP53 gene in chromosome 17. This gene is shown to be linked to different single amino acid mutations on which we will shed light. The method we propose transforms SNP textual sequences into digital vectors by coding. Then, RGB scalogram images are generated using the continuous wavelet transform. A pretreatment of color coefficients is applied to scalograms aiming at creating four different databases. Finally, a Convolutional Neural Network deep learning network is used for the binary classification of cancerous and non-cancerous images.

RESULTS: During the validation process, we reached good performance with specificity of 97.84%, sensitivity of 96.45%, an overall accuracy of 95.29% and an equal run time of 12 minutes 3 seconds. These values ensure the efficiency of our method.To enhance more these results, we used the Oriented FAST Rotated and BRIEF features detection technique. Consequently, the classification rates have been slightly improved to reach 95.9% as accuracy, but it reduced the run time to 9 minutes 36 seconds.

CONCLUSION: Our method will allow significant saving time and lives by detecting the disease in patients whose genetic mutations are beginning to appear.

Received on 08 April 2023; accepted on 23 November 2023; published on 28 November 2023

Keywords: CNN classification, breast cancer, scalogram, ORB, SNP, tumor suppressor genes, TP53 gene

Copyright © 2023 S. Ben Nasr *et al.*, licensed to EAI. This is an open access article distributed under the terms of the CC BY-NC-SA 4.0, which permits copying, redistributing, remixing, transformation, and building upon the material in any medium so long as the original work is properly cited.

doi:10.4108/eetpht.9.3218

1. Introduction

In 2020, the large increase in cases diagnosed with breast cancer (BC) makes it the most common cancer; with 2.3 million diagnosed cases representing a rate of 11.7%. In Europe and the United States, the incidence of BC is relatively stagnant; but in Africa and Asia there is a sharp increase in the number of detected cases[1, 2].

This increase can be caused by unhealthy changes that affected our lifestyle. Mutations that can alter human genes are inherited from older generations or acquired by the radiation, exposure to certain chemicals early in life in the womb, during puberty, and during pregnancy. Other factors linked to the change of life such as: obesity, physical inactivity, late pregnancy and reduced breastfeeding can cause breast cancer. Pesticide DDT which persists for long periods in the environment, dioxins, which are formed when fuels (wood, coal, or oil are burned), polluted air, gasoline and organic solvents used in industry increase the risk of developing breast cancer later in life[3]. Thus, early diagnosis of breast cancer can decrease the number of deaths and minimize the duration and cost of treatment[4].

The clinical stage of BC diagnosis is one of the most important prognostic factors of survival[5].

In fact, there is an association of advanced clinical stage of BC with delays of more than three months between discovery and treatment start; which reduces survival. That is why the issue of BC early detection is important. Till now, it is difficult to detect the breast cancer at its primary stages. One can promote the early detection only with some acts such as: doctor's consultation as soon as changes in the breast occur, a clinical examination of the breast, a screening mammogram, etc[6]. On the other hand, the routine screening is not recommended for women under 40 years old because their breast tissue is generally denser than the breast tissue in older women. As a result diagnosis becomes more difficult. Nowadays, machine learning (ML) and deep learning (DL) are proposed as innovative technologies for detecting breast cancer from breast mammography, ultrasound, MRI, and IR imaging[7][8]. These algorithms process datasets that are individually collected or downloaded from the public databases. DDSM[9], INbreast[10], CBIS-DDSM[11] and MIAS[12] are examples of databases that include mammogram images. Only the MIAS database does not contain images of breast cancer, but it marks the table locations of any possible abnormality.

The following Table 1 summarizes datasets presented in these databases[13].

Table 1. Mammogram images in relation with breast cancer.

Data Set	Total number of mammogram Images	Cancerous images
DDSM	10480	2620
INbreast	410	115
CBIS-DDSM	10239	6775
MIAS	322	-

For a long time, mammography was the most used process to examine the human breast for diagnosis and screening. Different automated classifiers were tested on mammogram images which aim at distinguishing between benign and malignant patterns. In Table 2, we provide a summary of some recent studies using data from Table 1. In these works, the accuracy varies from 76% to 97.52%. However, the main drawback of all these works resides in the inability of early detection of BC[14, 15].

Table 2. Studies using machine and deep learning to detect breast cancer.

Dataset	Classifier	ACC
Private dataset	-Random forest [16]	95.30
	-RBF-SVM [17]	96.00
	-3 CNN architectures : GoogLeNet,VGGNet and ResNet [18]	97.52
DDSM	-You Only Look Once (YOLO-V1) [19]	97.00
	-YOLO-V1/ Fully connected neural network (FC-NN)[20]	85.52
	-CNN[20]	94.50
	-ResNet50 [20]	95.83
	-InceptionResNet -V2[20]	97.50
-INbreast	-CNN[20]	88.74
	-ResNet50[20]	92.55
	InceptionResNet -V2[20]	95.32
	-ResNet [20]	91.00
	-InceptionV3[20]	95.50
CBIS-DDSM	DCNN [20]	76.00

As a new orientation, many biologists target their research towards protein sequences whose expression favors the appearance of cancer[21, 22]. Indeed, genetic carriers have become one of the most valuable pieces of information in the field of early detection of human disease especially cancers. According to the Canadian Cancer Society, our DNA can undergo changes that lead to mutations in genes[23]. Each genetic mutation can have several effects such as inappropriate growth of cells at rest, a large amount of protein production and abnormal or insufficient protein production[24].

The protein is a molecule whose sequence is composed of 20 amino acids that can undergo modifications. These modifications can go completely unnoticed or can generate cancerous pathologies. The

★Breast cancer early detection in TP53 SNP protein sequences based on a new Convolutional Neural Network model

*Corresponding author. Email: saiffedine.bennasr@enit.utm.tn imen.messaoudi@enit.rnu.tn Afef.Elloumi@enit.utm.tn Zied.lachiri@enit.utm.tn

Human DNA mutation gives rise to Single Nucleotide Polymorphism (SNP)[4, 25–28]. More than 335 million SNPs have been found in humans, part of which may fall within coding sequences of genes. These gene modifications are found to be associated with inherited diseases. In fact, several biologists have identified many SNPs during the study of healthy and sick subjects carrying different alleles of a given gene. These studies have shown a difference in sensitivity to well-defined cancers among them, breast cancer represents a highly serious disease with the highest female incidence and mortality. Hence the importance of studying these mutations from an artificial intelligence point of view[22, 29, 30].

In this paper, we provide a new approach for the breast cancer early identification even before the stage 0. For this, we base our work on using pathological SNP mutations that are linked to the BC genesis or development as well as a simple architecture of the convolutional neural network (CNN). Genetic mutations represent the main regulators for the triggering of the treatment response as for the development of this type of cancer. On the other hand, women with breast cancer may not have symptoms during the primary stage where the diagnostic difficulty comes from. All of this makes us to seek for a new diagnostic method based on genomic sequences such as DNA, RNA, or protein sequences[31].

Our choice goes to protein sequences to determine whether there is cancer or not. Therefore, this type of sequence is considered an important decision element for the early diagnosis of breast cancer. Our study is based on deep learning image classification to identify the appropriate protein sequences linked to breast cancer. The basic idea was to use a deep classifier that returns a higher performance rate than other traditional machine learning. As for data, we propose a new method to represent the protein sequences into images as an alternative to mammography. This type of sequences represents the most abundant form of genetic variation in the human genome. Further, they are linked to susceptibility to disease and individual response to drugs[32–34]. To transform the string characters of SNP sequences, we have to apply a coding technique aiming at transforming them into digital vectors. Each vector will be then transformed into a scalogram-image using the continuous wavelet transform with the complex Morlet function, After that, we consider, the Convolutional Neural Network to classify data into healthy and cancerous. We further, adopt the Oriented FAST and Rotated BRIEF (ORB) before classification which aims to obtain more efficient BC prediction. Indeed, this serves as a data reduction tool since we locate the relevant descriptors and eliminate therefore the unnecessary parts of the scalogram by a simple crop.

This paper is divided into four parts. The second part presents the proposed method. Part 3 demonstrates the experimental results obtained by our new approach and part 4 concludes the paper.

2. Materials and methods

2.1. Proposed method

In this paper, we introduce an efficient solution to classify SNPs linked to breast cancer, and thus for early detection purposes. The contribution of this work can be summarized in five major parts. The correspondent flowchart is given in (figure 1). The first step is collecting protein sequences of SNP type. A database comprising cancerous sequences and non-cancerous ones will be generated thereafter. The next step consists of coding this data in order to transform them into 1-D digital vectors. For this goal, we will use the pseudo-Electron-Ion Interaction Potential coding technique. In the third step, the 1-D digital vectors will be transformed into images (scalograms) using the continuous wavelet transform. Following the images creation, we will subdivide the scalogram database into two groups of data as shown in figure 1. In this step, we will consider a database that contains 462 cancerous sequences and 104 non-cancerous sequences. Hence the need to find the best distribution that gives a sufficient number of images for learning our CNN model, on the one hand, and evaluating the model on the other hand. Here, we will consider 70% of the data for training and 30% for testing. Finally, we will use these two subgroups of data in a CNN model to complete the classification task. This step first consists of using the data that was designed for training in order to estimate the best CNN model. After that, the testing data will be used to check the performance of the given CNN model. In the following we will give a brief overview of the required methods to develop our system. Then, we will give more details about the SNP images database establishment, the preprocessing of these images and their classification by CNN.

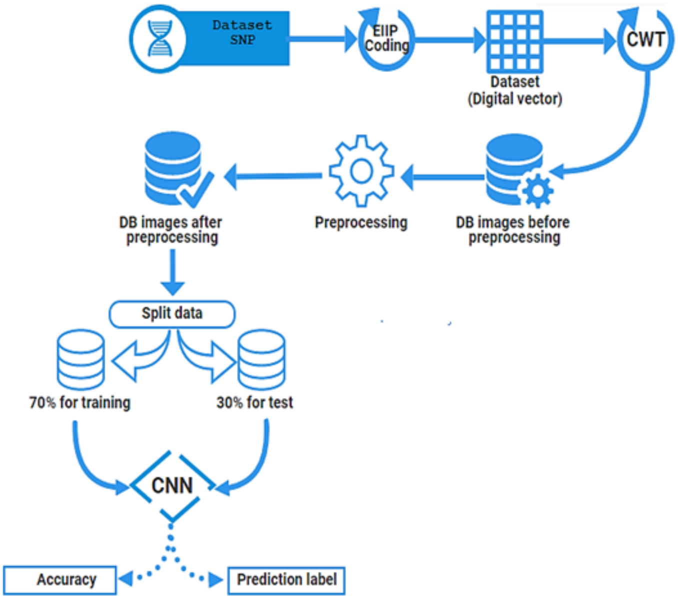


Figure 1. Our methodology's flowchart for breast cancer early detection.

2.2. Existing theories

We define here the methods we need to develop our work.

Pseudo-Electron-Ion Interaction Potential coding:. In order to recover digital vectors from the string characters of the protein sequences, several approaches are proposed. Some of these methods are based on the physicochemical properties of amino acids such as volume, charge, area, dipole moment, alpha, etc. Among the experimental coding techniques, the pseudo-Electron-Ion Interaction Potential (EIIP) attributes to each amino acid the value of the electrons' valence energy. The corresponding numerical values are provided in Table 3 [35, 36]. A wide range of scientific works used EIIP to identify hot spots in proteins and to design peptides that are useful in drug discovery.

Table 3. EIIP values of amino acids.

AMINO ACID	EIIP	AMINO ACID	EIIP
A	0.0373	I	0.0000
R	0.0959	K	0.0371
N	0.0036	M	0.0823
D	0.1263	F	0.0946
C	0.0829	P	0.0198
Q	0.0761	S	0.0829
E	0.0058	T	0.0941
G	0.0050	W	0.0548
H	0.0242	Y	0.0516
L	0.0000	V	0.0057

Continuous wavelet analysis:. The wavelet transform provides a frequency spectrum in relation to the locality. This gives the possibility to follow the information content change over the protein's entire length. The procedure consists of using a set of basic functions obtained by translation and extension operations of the so-called mother wavelet $\psi(t)$[36, 37]. These daughter wavelets are given by:

$$\psi_{a,b}(t) = \frac{1}{\sqrt{\alpha}}\psi^*\left(\frac{(t-b)}{\alpha}\right), \alpha > 0, b \in \Re \tag{1}$$

where b indicates the time localization and defines a translation of the wavelet, α is a scale parameter and * is the complex conjugate. Assuming the base wavelet is positioned around a frequency f_0 (maximum value of the mother wavelet spectrum), the frequency set is proportional to the scale one. The continuous wavelet transform (CWT) of a signal function X(t) is defined by $T_\psi(X)(a,b)$ as follows:

$$T_\psi(X)(a,b) = \frac{1}{\sqrt{4}}\int_a^b X(t)\psi^*\left(\frac{(t-b)}{a}\right)\mathrm{d}t \tag{2}$$

The modulus of the coefficients $|T_\psi(a,b)|$ is named scalogram. In this work, we choose the Complex Morlet function as the mother wavelet $\psi(t)$. It is a Gaussian-windowed complex sinusoid. Its mathematical formulation is given by:

$$\psi(t) = \pi^{\frac{-1}{4}}(e^{i\omega o^t} - e^{\frac{-1}{2}\omega o^2})e^{\frac{-1}{2}t^2} \tag{3}$$

Here ω_0 corresponds to the number of the wavelet oscillations ($\omega_0 = 2\pi f_0$). The value of ω_0 must be greater than 5 because of the mother wavelet invertibility.

The convolutional neural network:. The convolutional neural network is a type of artificial neural network which is commonly applied to image classification[38]. The CNN architecture is made up of a stack of independent processing layers. The figure below (Figure 2) shows the main layers (convolutional, pooling such as maxpool) and activation functions (that can be non-linear such as sigmoid and tanh) of a convolutional neural network.

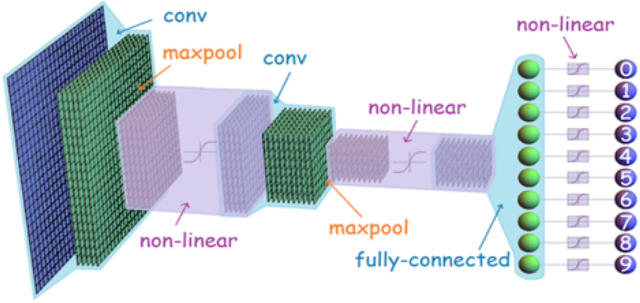


Figure 2. Structure of a CNN network.

2.3. The SNP images database establishment

Our goal is to make a difference between cancer sequences (CS: Cancer-related variations) and non-cancer ones (RS: Non-cancer specific variations) in an automatic way. To this end, we proceed in three steps:

- Step 1: Collecting CS and RS protein sequences.
- Step 2: Using a coding technique for the digital vector generation from the CS and RS sequences.
- Step 3: Creating images from the digital vectors we obtained in Step 2.

The protein dataset collection:. Here, we consider real data derived from recently published studies showing a relation with breast cancer genesis or development. Our choice goes to the CanProVar2 database[39]. This database is designed to store and display SNP sequences in the protein form. The CanProVar 2.0 database is created and fed from other public databases such as TCGA, COSMIC, HPI, BIOMART, and OMIM. It provides a full description of germline and somatic single amino acid changes in the human proteome. While inspecting this database we remarked on the diversity in terms of SNPs in humans. Among the human genes, the gene TP53 shows various SNP sequences linked to breast cancer. This gene is located in the p13.1 band of the chromosome 17 (p) arm [40, 41]. The TP53 gene is a tumor suppressor gene that causes the development of many types of human cancers. Several studies show that the TP53 protein is a transcription factor that controls the expression of many genes involved in apoptosis and cell cycle regulation. Other studies seek the relationship between chemotherapy and TP53 status. Despite the fact that the role of TP53 in human clinics is still controversial. Numerous animal studies show that inactivating this protein may increase sensitivity to some agents and resistance to others, giving it a key role in the response to anti-cancer agents[7].Besides breast cancer, this gene targets a large number of mutations linked to several types of cancer. After taking into account all the changes that can happen in the protein sequences of this gene, we generated an unbalanced database containing 462 protein sequences with breast cancer and 104 non-cancerous sequences.

The digital protein representation using the EIIP coding:. Based on Table 3, we assign a numerical value to each amino acid in a given sequence. For example, we take the protein sequence: Pseq = MDALKPP. After an EIIP coding, we obtain the following digital vector: Pnum=0.0823;0.1263;0.0373;0.0000;0.0371;0.0198;0.0198. Based on this technique, we generated a database of 1-D numerical representations related to the CS and RS protein sequences.

The protein image representation using the continuous wavelet analysis:. The final step in generating our database consists of converting the $1-D$ data of CS and RS sequences into color images. The protein sequences are of a complex nature. Their time-frequency (or time-scale) representations: the scalograms, allow us to describe the existing periodic structures with great precision. In fact, through this transformation, we get scalogram images under RGB color space with a size of [875×656×3] that well characterize SNP sequences linked to breast cancer (CS) and other unrelated (RS). These images present a multitude of repeating patterns at different scales of observation that are specific to SNP sequences. In Figure 3, we provide an illustrative example of two scalograms related to CS and RS sequences.

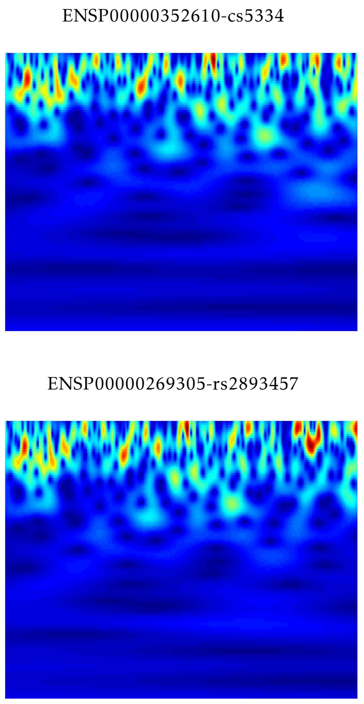


Figure 3. Scalogram representation of CS and RS sequences.

As it is shown in this figure, it is difficult to detect great differences between these two image classes with the naked eye. This is due to the fact that scalograms represent mutations of a single amino acid in the sequences. For that reason, we have thought of using the convolutional neuron network because of its advantage in identifying singular differences in images. On that basis, we will use thereafter a simple CNN architecture for the binary classification of CS and RS images. But before this step, we will proceed with some preprocessing operations on the scalogram images; which are generally required to enhance the classification rates.

2.4. Preprocessing of SNP images database

In our research, we have chosen to work on two sets of data. The first one contains the scalogram images having undergone two pre-processing steps such as: resizing and changing the color space to HSV, LAB, and grayscale. These steps will allow us to maximize the classification rates and reduce the CNN processing time. The second data set is created using the ORB features in the RGB scalogram images. After identifying the area that consists of the ORB descriptors, a cropping operation is done: we only keep the part with hotspots in the image. The figure below illustrates our methodology for image preprocessing (Figure 4).

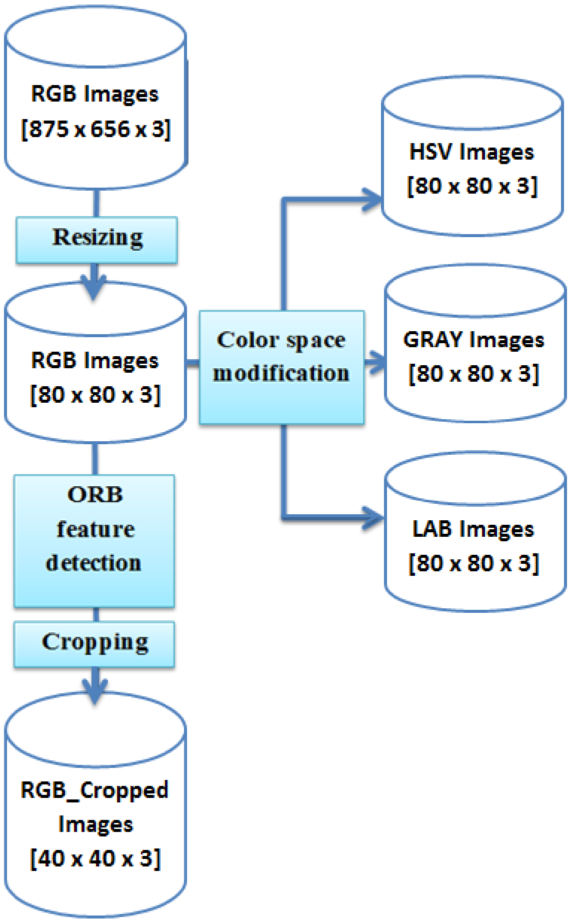


Figure 4. Our methodology for preprocessing the CS and RS scalogram images.

Resizing:. The large number of input pixels to the CNN network (which represents the number of neurons) can cause memory overflow. That is where a need to find a compromise between the data to be classified, the CNN model, and the material. Here, our preliminary database contains scalograms of size 875×656×3 (3 for RGB). So for each scalogram, we need 574,000 neurons (875×656×3) to be entered into the CNN network. This large number of neurons could be minimized by resizing our images. Nevertheless, this operation of resizing must preserve the distinctive characteristics of each scalogram image. Several techniques are available such as bi-cubic interpolation, bilinear interpolation, and the nearest neighbor. Taking into account the characteristics of each method, we found that the bi-cubic method is the most efficient in terms of time and quality of result. For the size of the scalograms to be obtained, we noticed after several tests that certain descriptors disappear as soon as we exceed the size [70×70]. That is why we have chosen to work on images of size [80×80].

Color space:. The CS and RS scalograms are presented under the RGB color space. Modifying the color space gives the possibility to present the descriptors of each image with other values. This can increase the classification rate in terms of the learning and test rates. Here, we consider the LAB, Grayscale, and HSV (Hue Saturation Value) color spaces[28, 42].

- The colorimetric space LAB (L*a*b) presents colors in three channels. The first channel L *, is reserved for the luminance (luminosity) and takes a value between 0 (for black) and 100 (for white). The two values a * and b * give both the chromaticity. Here, a * denotes the location of the color along the red-green axis, and b * denotes the location of the color along the blue-yellow axis [43].
- The second colorimetric space we have chosen is the grayscale. This space allows us to decrease the number of neurons entering the classification model since it replaces the three RGB color channels (red, green, and blue) in a single value of the luminosity of each pixel [44].
- The HSV space is represented with three coefficients. The shade H is coded according to the corresponding angle within the color circle. The saturation S represents the intensity of the color; it varies between 0 and 100%. The value V gives the "brightness" of the color; it varies also between 0 and 100% [43].

For example, we consider the two SNP scalograms illustrated in Figure 3, which belong to CS and RS sequences. After color space modifications, we obtain the results given by Figure 5.

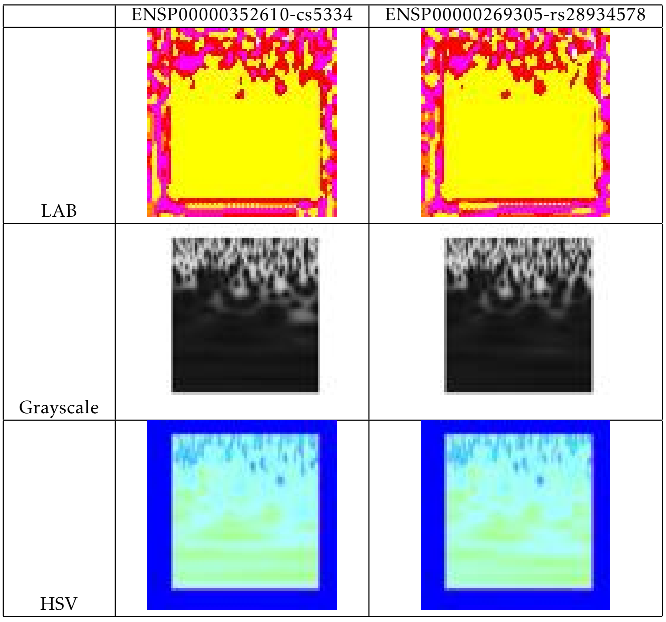


Figure 5. Scalogram color modification of CS and RS sequences.

After converting the CS and RS scalograms into L*a*b, gray level, and HSV, we obtained four databases to which we gave the name (respectively) of: Database_RGB, Database_LAB , Database_N and Database_HSV.

Cropping scalograms based on the Oriented FAST Rotated and BRIEF (ORB) features detection:. To reduce the amount of data entering our CNN neural network, we have found a way to get rid of the parts of the scalograms that do not contain interesting information. For this, we thought of using a method that identifies the main descriptors in the scalogram images and marks their position. After that, we can eliminate unnecessary data, which will improve the classification efficiency in time and model stability. Our choice goes to the technique Oriented FAST Rotated and BRIEF, also called ORB. The FAST (Features from Accelerated Segment Test) key point detection and binary BRIEF descriptor are used in the development of ORB. The finest features are extracted from an image using ORB which helps in feature selection. Before using a Harris corner detector to extract useful characteristics from those key points, ORB applies the FAST key point detector, which finds a lot of key points. Results from extracted characteristics are more accurate and less noise-sensitive. Equation 4 can be used to determine the centroid of the image using the patch moment in ORB[45].

$$m_{pq} = \sum {}_{x;y} x^p y^q (x;y) \tag{4}$$

The intensity centroid of image patches is used to determine the corner direction using equation 5.

$$C\left(\frac{m_{10}}{m_{00}}, \frac{m_{01}}{m_{00}}\right) \tag{5}$$

The angle between the patch's centroid and center is provided by equation 6.

$$atang2\left(\frac{m_{10}}{m_{00}}, \frac{m_{01}}{m_{00}}\right) = atang2(m_{01}, m_{10}) \tag{6}$$

Figure 6 shows the positions of the characteristic points after application of the ORB method in the RGB scalograms of a CS cancer sequence and a non-cancerous RS one. As can be seen, this method allowed us to identify the part of the image that contains the relevant information. Note that this information is rotation-invariant and noise-resistant.

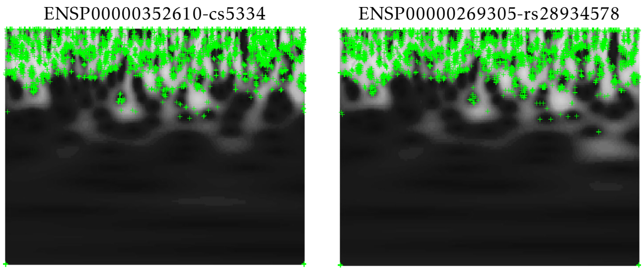


Figure 6. The ORB points in ENSP00000352610-cs5334 and ENSP00000269305-rs28934578 scalograms before cropping.

After determining the ORB points in scalograms, we found that the main descriptors are concentrated in the upper part of each image. Based on this observation, we crop this part and eliminate the remaining part that represents no important information, and thus for images belonging to cancerous and non-cancerous cases. This technique will be used later on only one database among Database_RGB, Database_LAB, Database_N, and Database_HSV. The chosen database shall be that which will have the highest classification rate without applying the ORB technique. Figure 7 shows the RGB images of ENSP00000352610-cs5334 and ENSP00000269305-rs28934578 after cropping.

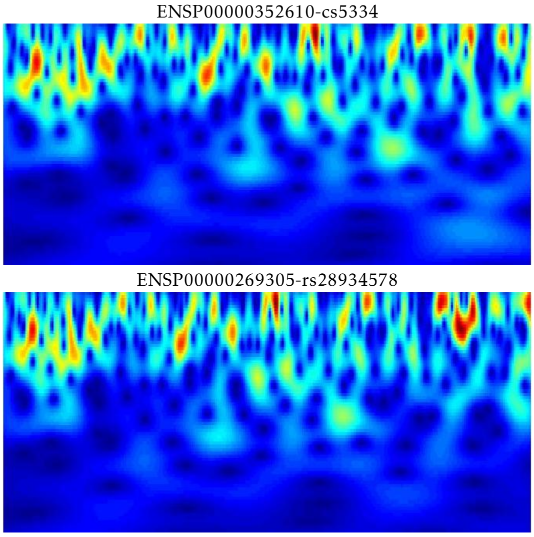


Figure 7. The ENSP00000352610-cs5334 and ENSP00000269305- rs28934578 cropped scalograms after ORB features localization.

2.5. CNN classification

CNN's role consists of determining the descriptors that allow us to identify CS cancerous from their scalograms instead of processing them in textual form. Features extracted are then fed to a fully connected layer for the classification to be done. We have a binary classification situation here; which aims at the distinction between protein sequences linked to breast cancer sequences and healthy ones. In this work, we created a simple CNN model, to which we progressively added more layers and modified the coefficients of each layer. The purpose of this operation is to find distinguishing descriptors of the CS images that enhance the classification rate. We used also small filters of size [3×3] or [5×5] for each layer of convolution in order to locate the CS images descriptors. Finally, to stabilize classification rates, we used normalization functions[46]. Figure 8 describes the overall convolutional neural network architecture of the system we implemented.

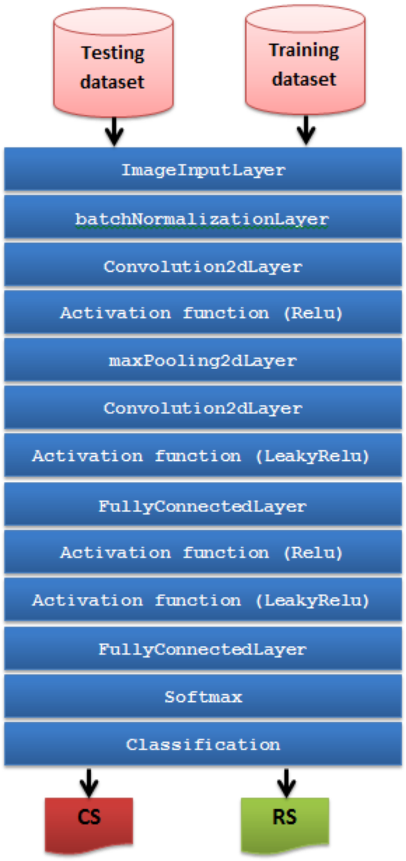
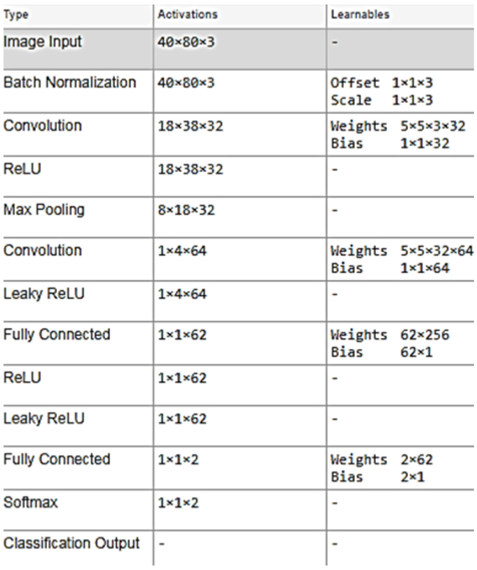


Figure 8. Our CNN model for CS and RS classification.

We present in Figure 9 the hyper-parameters of our CNN model.



Type	Activations	Learnables
Image Input	40×80×3	-
Batch Normalization	40×80×3	Offset 1×1×3 Scale 1×1×3
Convolution	18×38×32	Weights 5×5×3×32 Bias 1×1×32
ReLU	18×38×32	-
Max Pooling	8×18×32	-
Convolution	1×4×64	Weights 5×5×32×64 Bias 1×1×64
Leaky ReLU	1×4×64	-
Fully Connected	1×1×62	Weights 62×256 Bias 62×1
ReLU	1×1×62	-
Leaky ReLU	1×1×62	-
Fully Connected	1×1×2	Weights 2×62 Bias 2×1
Softmax	1×1×2	-
Classification Output	-	-

Figure 9. The CNN model hyper-parameters.

As for the input data, our model takes 70% of the data set for training and 30% for testing. Note that for each data set, the images are chosen at random without having the possibility that two data sets can exist at the same time.

3. Results and discussion

While designing our CNN model, we used several measures such as accuracy (ACC), sensitivity (TPR), specificity (TNR), and the precision rate (PPV) [47]. The following measurements are calculated according to the equations below:

$$Accuracy(ACC) = \frac{tp + tn}{tp + fp + fn + tn} \tag{7}$$

$$Sensitivity(TPR) = \frac{tp}{tp + fn} \tag{8}$$

$$Specificity(TNR) = \frac{tn}{fp + tn} \tag{9}$$

$$Precision(PPV) = \frac{tp}{tp + fp} \tag{10}$$

The *tp*, *fp*, *fn*, and *tn* values used in the formulas of the CNN model evaluators denote respectively: true positive, false positive, false negative, and true negative. For each database, we have made three iterations with a random selection of samples.

3.1. Results of not cropped images

In Table 4, we provide the values of the performance measures we obtained considering not cropped RGB images.

Table 4. Classification results by dataset.

Test	Dataset	ACC	PPV	TPR	TNR
1	Database_RGB	91.76%	94.96%	94.96%	77.42%
2		92.94%	94.24%	97.04%	77.14%
3		95.29%	97.84%	96.45%	89.66%
Average		93.33%	95.68%	96.15%	81.41%
1	Database_HSV	91.76%	94.24%	95.62%	75.76%
2		90.00%	97.84%	90.67%	85.00%
3		93.53%	96.40%	95.71%	83.33%
Average		91.76%	96.16%	94.00%	81.36%
1	Database_N	90.00%	97.12%	91.22%	81.82%
2		91.18%	97.12%	92.47%	83.33%
3		91.18%	99.28%	90.79%	94.44%
Average		90.79%	97.84%	91.49%	86.53%
1	Database_LAB	89.41%	94.96%	92.31%	74.07%
2		93.53%	94.96%	97.06%	79.41%
3		90.00%	98.56%	90.13%	88.89%
Average		90.98%	96.16%	93.17%	80.79%

As we can see, the best accuracy with the value of 95.29% is obtained with Dataset_RGB in test 3. As for PPV and TNR, it is Dataset_N which gives the best results with rates of 99.28% and 94.44% respectively in test 3. In terms of TPR, we reach 97.06% with Dataset_LAB in test 2. On average the best values of ACC and TPR are reached with Dataset_RGB (93.33% and 96.15% respectively). With respect to PPV and TNR, the best rates are obtained with Dataset_N with the value of 97.84% and 86.53% respectively.

Overall, there is no wide discrepancy between the values obtained with data in the RGB, grayscale, HSV, and LAB color spaces. The efficiency of our model in terms of breast cancer classification, is especially shown when data are expressed in the RGB color space (Dataset_RGB). In fact, with this dataset, we reached 95.29% of accuracy, 97.84% of precision, 89.66% of specificity, and 96.45% of sensitivity. As for the average classification time, it was equal to 12 minutes and 3 seconds. We can also show in Figure 10 the results of the learning progression of our model with random data when we considered Dataset_RGB.

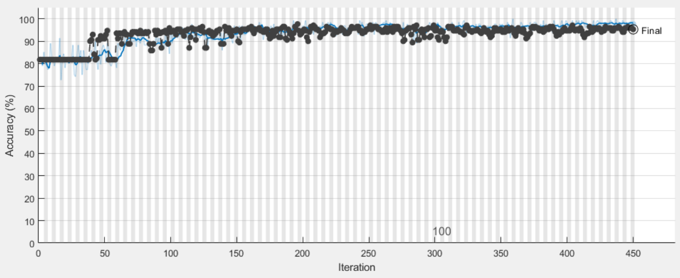


Figure 10. Example of the training progression of our CNN model with Dataset_RGB as data.

We further provide, in figure 11 the confusion matrix corresponding to the best test of Dataset_RGB.

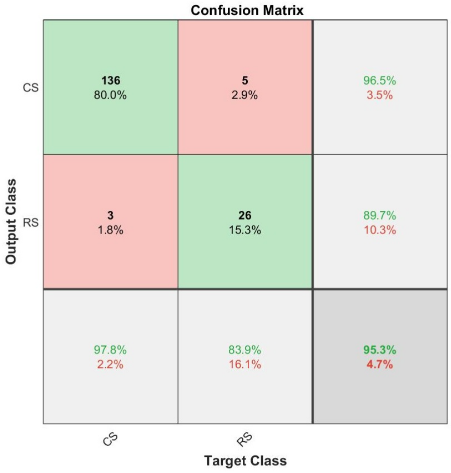


Figure 11. Confusion matrix corresponding to the best test of Dataset_RGB.

3.2. Results of cropped images

Now, we apply the same CNN model to the database whose images are cropped. The image cropping is done after locating the area where the descriptors of the CS and RS scalograms are concentrated. Three tests were made. For these tests, we have used the same learning and test data of the three tests already carried out on the Dataset_RGB database. This task consists of reducing the data processed by our CNN model and stabilizing this model by eliminating unnecessary data. Table 5 gives a comparison between the results obtained with the Dataset_RGB database and the Dataset_RGB_Cropped database.

Table 5. Comparison between the classification of Cropped and entire scalogram images.

Test	Dataset	ACC	PPV	TPR	TNR
1	Database_RGB	91.76%	94.96%	94.96%	77.42%
2		92.94%	94.24%	97.04%	77.14%
3		95.29%	97.84%	96.45%	89.66%
Average		93.33%	95.68%	96.15%	81.41%
1	Database_RGB _Cropped	91.76%	96.40%	93.71%	81.48%
2		92.94%	95.68%	95.68%	80.65%
3		95.88%	99.28%	95.83%	96.15%
Average		93.53%	97.12%	95.07%	86.09%

According to Table 5, cropping the main informative part of the scalograms is effective in terms of minimizing data entered into the CNN model, improving the classification rates, and giving better stability to the classification operation (Figure 12).

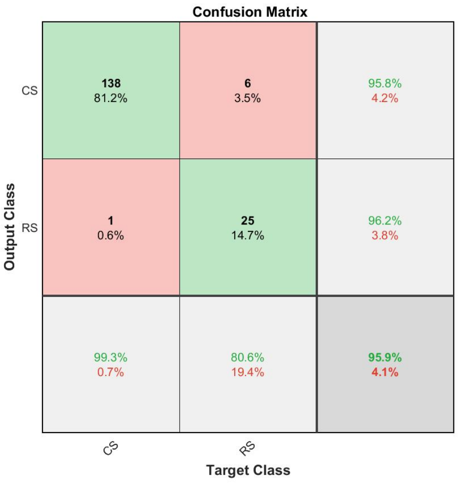


Figure 12. Confusion matrix corresponding to the best test of Dataset_RGB_Cropped.

With our method, we achieved a prediction of 95.88% with an equal run time of 9 minutes 36 seconds, which exceeds the rates obtained from several previously published studies (Table 2). As we are the first to have worked on SNP data from the CanProVar database targeting breast cancer, we compare our results with those of other works linked to other cancers. We, also consider one of our previous works, in which we classified SNPs linked to intestinal cancer. The following Table 6 exposes works using data from the CanProvar database.

Table 6. Studies using data extracted from the CanProvar database.

Method	Dataset	ACC	PPV	TPR	TNR
CanSavPre	set 30	87.43%	27.52%	58.44%	89.42%
CanSavPre	set 40	85.50%	33.94%	58.39%	88.30%
DEOGEN2	set 30	72.22%	12.25%	55.17%	73.37%
DEOGEN2	set 40	69.27%	14.47%	46.93%	71.56%
Our CNN Intestinal cancer	1192 sequences	96.18%	98.32%	97.42%	85.46%
Our CNN breast cancer	**566 sequences**	**95.88%**	**99.28%**	**95.83%**	**96.15%**

According to the Table 6, our method exceeds CanSavPre and DEOGEN2[48] which aim to identify cancerous mutations of a single amino acid in other cancers. They do not exceed the value of 88% as accuracy rate and present a low precision of less than 34%. Instead, with another simple CNN architecture and the same way to represent the biological data we obtained close results in the case of intestinal cancer. Indeed, we reached 96.19% in accuracy, 97.42% in sensitivity, 85.46% in specificity, and 98.32% in precision with a run time of 9 minutes 9 seconds[28]. In this work, we succeeded the breast cancer early detection even with the absence of symptoms with very encouraging rates.

4. Conclusion

The purpose of this paper is to present a new method for the early identification of cancerous lesions linked to breast cancer. Our methodology counts on screening the SNP protein sequences based on a deep learning model. For this, we divided SNPs into two types: CS (which designates the SNP protein sequences linked to breast cancer) and RS (which designates non-pathological SNP protein sequences). Here, we used SNPs of the TP53 gene which is located in the p13.1 band of chromosome 17. To these sequences, we applied the EIIP coding technique in order to transform the protein SNPs into digital vectors. The 1D signals we obtained were transformed after that into scalogram images using the complex Morlet wavelet transform. The resulting database: Dataset_RGB contains 462 CS images and 104 RS images. After resizing images using the bi-cubic method, this database was presented in different color spaces in an attempt to enhance the inherent descriptors. This led to the design of three other databases which are: Dataset_NG, Dataset_HSV, and Dataset_LAB. On the other hand, the CNN neural network architecture can reveal distinguishable characteristics of images. That is why we were able to recognize CS images from RS ones although it was impossible to detect with the naked eye the difference between the two classes (because of the presence of only a single mutation of an amino acid). With a simple CNN architecture, we have obtained good performance in terms of accuracy, precision, specificity, and sensitivity, and thus for the four considered databases. To reduce the amount of data in the CNN network input, we adopted the Oriented FAST Rotated and BRIEF technique. After applying ORB to the scalograms, we eliminated unnecessary data with the aim of improving the classification efficiency in time and ensuring the model stability. The results are close to those previously obtained, and the best precision was that of the cropped RGB images (Dataset_RGB_Cropped). In fact, with this dataset, we reached 95.88% of accuracy, 99.28% of precision, 96.15% of specificity, and 95.83% of sensitivity with an equal run time of 9 minutes 36 seconds. Based on these significant results, we can rely on our classifier for early detection of breast cancer even in the absence of symptoms. It can be further used for BC screening in fetuses since genetic factors are among the main causes of this type of pathology. Finally, we hope that this tool for early BC diagnosis will increase the cure rate and shorten treatment times in the future.

References

[1] https://www.who.int/news-room/fact-sheets/detail/breast-cancer.

[2] https://www.cdc.gov/cancer/breast/basic_info/index.htm.

[3] https://www.breastcancer.org/research-news/chemical-exposure-early-in-life-increases-risk.

[4] Nkondjock, A. and Ghadirian, P. (2005) Facteurs de risque du cancer du sein. *médecine/sciences* **21**(2): 175–180.

[5] www2.le.ac.uk/projects/vgec/highereducation/topics/dna-genes\protect\discretionary{\char\hyphenchar\font}{}{}chromosomes/resources.

[6] ADAKO, O.P. (2014) *KNOWLEDGE OF BREAST CANCER AND PREFERENCE OF EARLY DETECTION SCREENING MEASURES AMONG FEMALE UNDERGRADUATE STUDENTS OF EKITI STATE UNIVERSITY, ADO EKITI, NIGERIA*. Ph.D. thesis.

[7] Khuriwal, N. and Mishra, N. (2018) Breast cancer detection from histopathological images using deep learning. In *2018 3rd international conference and workshops on recent advances and innovations in engineering (ICRAIE)* (IEEE): 1–4.

[8] Ragab, D.A., Sharkas, M., Marshall, S. and Ren, J. (2019) Breast cancer detection using deep convolutional neural networks and support vector machines. *PeerJ* **7**: e6201.

[9] Shen, R., Yan, K., Xiao, F., Chang, J., Jiang, C. and Zhou, K. (2018) Automatic pectoral muscle region segmentation in mammograms using genetic algorithm and morphological selection. *Journal of digital imaging* **31**: 680–691.

[10] Moreira, I.C., Amaral, I., Domingues, I., Cardoso, A., Cardoso, M.J. and Cardoso, J.S. (2012) Inbreast: toward a full-field digital mammographic database. *Academic radiology* **19**(2): 236–248.

[11] Ma, Y., Yang, C., Zhang, J., Wang, Y., Gao, F. and Gao, F. (2020) Human breast numerical model generation based on deep learning for photoacoustic imaging. In *2020 42nd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)* (IEEE): 1919–1922.

[12] Yoon, W.B., Oh, J.E., Chae, E.Y., Kim, H.H., Lee, S.Y. and Kim, K.G. (2016) Automatic detection of pectoral muscle region for computer-aided diagnosis using mias mammograms. *BioMed research international* **2016**.

[13] Alkhaleefah, M., Ma, S.C., Chang, Y.L., Huang, B., Chittem, P.K. and Achhannagari, V.P. (2020) Double-shot transfer learning for breast cancer classification from x-ray images. *Applied Sciences* **10**(11): 3999.

[14] Krishna, K.S. and Prince, P.G.K. (2020) Deep learning based breast cancer detection-a survey .

[15] Ragab, D.A., Sharkas, M., Marshall, S. and Ren, J. (2019) Breast cancer detection using deep convolutional neural networks and support vector machines. *PeerJ* **7**: e6201.

[16] Kathale, P. and Thorat, S. (2020) Breast cancer detection and classification. In *2020 International Conference on Emerging Trends in Information Technology and Engineering (ic-ETITE)* (IEEE): 1–5.

[17] Al-Azzam, N. and Shatnawi, I. (2021) Comparing supervised and semi-supervised machine learning models on diagnosing breast cancer. *Annals of Medicine and Surgery* **62**: 53–64.

[18] Khan, S., Islam, N., Jan, Z., Din, I.U. and Rodrigues, J.J.C. (2019) A novel deep learning based framework for the detection and classification of breast cancer using transfer learning. *Pattern Recognition Letters* **125**: 1–6.

[19] Aly, G.H., Marey, M., El-Sayed, S.A. and Tolba, M.F. (2021) Yolo based breast masses detection and classification in full-field digital mammograms. *Computer Methods and Programs in Biomedicine* **200**: 105823.

[20] Hamed, G., Marey, M., Amin, S. and Tolba, M. (2021) Comparative study and analysis of recent computer aided diagnosis systems for masses detection in mammograms. *International Journal of Intelligent Computing and Information Sciences* **21**(1): 33–48.

[21] Zhang, M., Wang, B., Xu, J., Wang, X., Xie, L., Zhang, B., Li, Y. *et al.* (2017) Canprovar 2.0: an updated database of human cancer proteome variation. *Journal of proteome research* **16**(2): 421–432.

[22] Li, J., Duncan, D.T. and Zhang, B. (2010) Canprovar: a human cancer proteome variation database. *Human mutation* **31**(3): 219–228.

[23] da Cunha Santos, G., Dhani, N., Tu, D., Chin, K., Ludkovski, O., Kamel-Reid, S., Squire, J. *et al.* (2010) Molecular predictors of outcome in a phase 3 study of gemcitabine and erlotinib therapy in patients with advanced pancreatic cancer: National cancer institute of canada clinical trials group study pa. 3. *Cancer* **116**(24): 5599–5607.

[24] Weinberg, R.A. (1996) How cancer arises. *Scientific American* **275**(3): 62–70.

[25] Chen, X. and Gonçalves, M.A. (2018) Dna, rna, and protein tools for editing the genetic information in human cells. *Iscience* **6**: 247–263.

[26] Durand-Dubief, M. (2005) *Régulations génétique et moléculaire par ARN interférence chez Trypanosoma brucei*. Ph.D. thesis, Museum national d'histoire naturelle-MNHN PARIS.

[27] Snustad, D.P. and Simmons, M.J. (2015) *Principles of genetics* (John Wiley & Sons).

[28] Nasr, S.B., Messaoudi, I., Oueslati, A.E. and Lachiri, Z. (2021) Cnn model applied on snp protein sequences for intestinal cancer early detection. In *2021 18th International Multi-Conference on Systems, Signals & Devices (SSD)* (IEEE): 255–263.

[29] Pavlopoulou, A., Spandidos, D.A. and Michalopoulos, I. (2015) Human cancer databases. *Oncology reports* **33**(1): 3–18.

[30] Ghosal, R., Kloer, P. and Lewis, K. (2009) A review of novel biological tools used in screening for the early detection of lung cancer. *Postgraduate medical journal* **85**(1005): 358–363.

[31] Feng, Y., Spezia, M., Huang, S., Yuan, C., Zeng, Z., Zhang, L., Ji, X. *et al.* (2018) Breast cancer development and progression: Risk factors, cancer stem cells, signaling pathways, genomics, and molecular pathogenesis. *Genes & diseases* **5**(2): 77–106.

[32] Alwi, Z.B. (2005) The use of snps in pharmacogenomics studies. *The Malaysian journal of medical sciences: MJMS* **12**(2): 4.

[33] Barkur, S.S. (2007) Snps in disease gene mapping, medicinal drug development and evolution. *Journal of human genetics/Japan Society of Human Genetics* **52**(11): 871–880.

[34] Shastry, B.S. (2007) Snps in disease gene mapping, medicinal drug development and evolution. *Journal of human genetics* **52**: 871–880.

[35] Meher, J.K., Dash, G.N., Meher, P.K., Raval, M.K. *et al.* (2011) A reduced computational load protein coding predictor using equivalent amino acid sequence of dna string with period-3 based time and frequency domain analysis. *American Journal of Molecular Biology* **1**(02): 79.

[36] Messaoudi, I., Oueslati, A.E. and Lachiri, Z. (2014) 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): 1–13.

[37] Messaoudi, I., Elloumi, A. and Lachiri, Z. (2013) Detection of the 6.5-base periodicity in the c. elegans introns based on the frequency chaos game signal and the complex morlet wavelet analysis. *International Journal of Scientific Engineering and Technology* **2**(12): 1247–1251.

[38] Sivangi, K.B., Dasari, C.M., Amilpur, S. and Bhukya, R. (2022) Noas-ds: Neural optimal architecture search for detection of diverse dna signals. *Neural Networks* **147**: 63–71.

[39] `https://www.http://canprovar2.zhang-lab.org/`.

[40] Olivier, M., Langer d, A., Carrieri, P., Bergh, J., Klaar, S., Eyfjord, J., Theillet, C. *et al.* (2006) The clinical value of somatic tp53 gene mutations in 1,794 patients with breast cancer. *Clinical cancer research: an official journal of the American Association for Cancer Research* **12**(4): 1157–1167.

[41] Langerød, A., Zhao, H., Borgan, Ø., Nesland, J.M., Bukholm, I.R., Ikdahl, T., Kåresen, R. *et al.* (2007) Tp53mutation status and gene expression profiles are powerful prognostic markers of breast cancer. *Breast cancer research* **9**(3): 1–16.

[42] Nasr, S.B., Messaoudi, I., Oueslati, A.E. and Lachiri, Z. (2022) Pre-mirna sequence prediction using convolutional neural network. In *2022 6th International Conference on Advanced Technologies for Signal and Image Processing (ATSIP)* (IEEE): 1–6.

[43] Abbadi, N. and Razaq, E. (2020) Automatic gray images colorization based on lab color space. *Indonesian Journal of Electrical Engineering and Computer Science* **18**(3): 1501–1509.

[44] Kekre, H.B. and Thepade, S.D. (2008) Color traits transfer to grayscale images. In *2008 first international conference on emerging trends in engineering and technology* (IEEE): 82–85.

[45] Gupta, S., Kumar, M. and Garg, A. (2019) Improved object recognition results using sift and orb feature detector. *Multimedia Tools and Applications* **78**: 34157–34171.

[46] O'Shea, K. and Nash, R. (2015) An introduction to convolutional neural networks. *arXiv preprint arXiv:1511.08458* .

[47] Lee, K.W., Yoon, H.S., Song, J.M. and Park, K.R. (2018) Convolutional neural network-based classification of driver's emotion during aggressive and smooth driving using multi-modal camera sensors. *Sensors* **18**(4): 957.

[48] Liu, J.J., Yu, C.S., Wu, H.W., Chang, Y.J., Lin, C.P. and Lu, C.H. (2021) The structure-based cancer-related single amino acid variation prediction. *Scientific reports* **11**(1): 13599.