EAI Endorsed Transactions
on Pervasive Health and Technology

Research Article EAI.EU

Cancer disease multinomial classification using transfer learning and SVM on the genes' sequences

Ines Slimene[1,*], Imen Messaoudi[2], Afef Elloumi Oueslati[3] and Zied Lachiri [4]

[1, 4]Electrical Engineering Department, SITI Laboratory, National School of Engineers of Tunis, University of Tunis El Manar, Tunis, Tunisia.
[2]Industrial Computing Department, Higher Institute of Information Technologies and Communications, University of Carthage, Tunis, Tunisia.
[3]Electrical Engineering Department, National School of Engineers of Carthage. University of Carthage, Tunis, Tunisia.

Abstract

INTRODUCTION: The advent of high-throughput sequencing technologies has provided an opportunity to explore the role of miRNAs in cancer classification. Altered gene expression by miRNAs has been linked to various cancers and can serve as potential biomarkers for disease classification.
OBJECTIVES: The objective of this study is to develop a deep learning model for cancer disease classification based on the analysis of gene sequences which are altered by miRNAs.
METHODS: The study utilizes a comprehensive dataset of altered gene sequences for various cancer types. Gene features are extracted using the Frequency Chaos Game Representation technique. Deep learning models, such as DeepInsight and ResNet, are trained using the pre-processed data. The models are designed to learn complex patterns and relationships between miRNA expression profiles and gene sequences, enabling them to classify cancer types with high accuracy.
RESULTS: A combination of ResNet50 and SVM in classification, and $FCGR_6$ in feature extraction achieved an accuracy of 0.98 in terms of distinguishing between 18 different cancer types.
CONCLUSION: This study highlights the promising application of deep learning models for cancer disease classification based on the analysis of gene sequences altered by miRNAs. Further research and validation are required to explore the clinical utility of these models and incorporate them into clinical practice.

Keywords: Cancer, FCGR, DeepInsight, Transfer learning, SVM

Received on 09 April 2023, accepted on 28 May 2023, published on 10 July 2023

doi: 10.4108/eetpht.9.3220

1. Introduction

Detecting cancer disease still a challenging problem. Although many research are done using clinical data and biological tools, they still limited because cancer is usually detected in later stages after attacking and damaging one or many organs. In addition, those tools are resource consumers and need a lot of execution time. Moreover, used data needs a lot of treatment to extract information. Nevertheless, gene sequences are a good way to discover diseases in their first stages since diseases are the result of gene alterations. These gene alterations are caused by inhibition or repression of the translation and the protein production process by microRNAs (1) . Hence, analysing gene sequence expression provides information about medical condition. However, we need performant tools to manage the big size of those data. In bioinformatic fields, deep learning algorithms show their efficiency in disease prediction and classification despite the data size. Deep

[*]Corresponding author. Email: ines.slimene@enit.utm.tn

learning models are a collection of algorithms based on a neural network in a layered architecture inspired by the human brain neural connections. This complex structure helps in analysing and extracting information from complex medical images and videos. Today, we are in the presence of many studies that use deep learning algorithms for diseases detection. Shakeel et.al compared Instantaneously Trained Neural Networks (DITNN), Radial Neural Networks (RBNN), Convolution Neural Networks (CNN), and Hopfield Neural Networks (HNN) for Lung cancer detection. Used data are extracted from Cancer imaging Archive (CIA). DITNN gives the best accuracy value which is 0.98 (2).

Another research used Histogram Equalization (HE) to enhance lung cancer scan quality, then, a segmentation process which is Artificial Bee Colony (ABC) is done to detect the region of interest. Finally, the lung disease classification was done with Fuzzy Particle Swarm Optimization Convolution Neural Network (FPSOCNN), an optimized version of CNN, to ameliorate and reduce processing time and cost. Obtained average accuracy's value is 95% (3).

Allugunti et.al used CNN for cancer disease detection based on thermographic images. Obtained accuracy value was better than SVM and Random Forest models. However, other steps were done before classification which are segmentation and denoising (4).

Another work applied CNN model to Breast Cancer Histology Images and obtained an accuracy of 87% (5). Begum et al. Combined LSTM and random forest approach for breast cancer detection (6).After shaping and medical pre-processing of breast cancer infrared thermal images, Mambou et.al used DNN and SVM to detect whether it is a normal or a cancerous sample (7).

Ahmad et al. used un hybrid deep learning model on PCam Kaggle dataset to classify and detect lymph node breast cancer. Proposed model is a combination of AlexNet, a transfer learning model, and gated recurrent unit (GRU) model. DensNet model gave an accuracy of 95.4% for cancer disease classification. Used data are histopathological images (8). Mallick et al. presented in their research a novel model based on deep wavelet autoencoder (DWA) from image compression and Deep neural networks (DNN) for Brain MRI Image Classification (9).

The review conducted by Fadi Alharbi et al. (10) provides an extensive analysis of machine and deep learning models utilized in the identification of cancer diseases using expression gene datasets. The findings of this review indicate that RNA-seq data exhibits superior performance compared to microarray data, particularly in scenarios involving multiclassification and large-scale data analysis. Regarding feature engineering techniques, Alharbi et al. discuss three methods: filter, embedded, and wrapper approaches. While wrapper models such as random forest and support vector machines (SVM) achieve an accuracy rate of approximately 99%, it is worth noting that these models require significantly more computational resources and time compared to filter methods.

Numerous machine models have been employed in the classification of cancer diseases. A comparative analysis of support vector machines (SVM), random forest, and k-nearest neighbors (kNN) reveals that SVM gives the best results. However, SVM's performance may vary across different cancer types. Furthermore, the study compares several deep learning models, including artificial neural networks (ANN), convolutional neural networks (CNN), recurrent neural networks (RNN), graph neural networks (GNN), temporal neural networks (TNN), and transfer learning approaches. The results indicate that deep learning models consistently outperform traditional machine learning models in cancer disease classification tasks.

Even with promising performance and commendable accuracy achieved in recent state-of-the-art studies, a substantial amount of pre-processing is required to optimize image quality and detect Regions of Interest (ROI). Consequently, these studies impose significant demands on time and resources.

The main drawback of all these methods, is that they rely on data collected after the apparition of the cancer. Here, we try to provide a new alternative for very early screening of cancer even before signs indicating its development begin to appear.

The main objective of this research is to employ deep learning models to classify 18 cancer disease types. By leveraging the information obtained from gene alteration analysis. Thus, early detection of cancer diseases can be achieved without the need for expensive and complex biological pre-processing procedures.

For cancer disease classification, a novel dataset is generated from gene alteration datasets and miRNA binding site data. A combination of dimension reduction and transfer learning models is utilized to reduce training time. With the introduction of our approach, execution time is significantly reduced, eliminating the need for an extensive image pre-processing step.

2. Methods

In this section, we present our novel method which is a combination of the Frequency Chaos Game Representation (FCGR) for feature extraction, DeepInsight for feature selection and a transfer learning model for cancer disease multi-classification. Here, we consider 18 types of cancer. The input data is a set of 1D sequence of the altered version of the gene responsible of each disease. We start by calculating the appearance frequency of each group of k-nucleotides in the sequence, where k= {4, ..., 7}. Then, we apply the DeepInsight model to select the most relevant features and convert the dataset into a 2D signal which will be the input of a deep learning model. Therefore, we could use a deep learning model with our data. Here, we choose to work with a transfer learning model to reduce the training time; it is a pre-trained model.

Finally, for the classification, we use SVM or fully connected layer (FC). The flowchart of the proposed cancer multiclassification system is given in Figure 1.

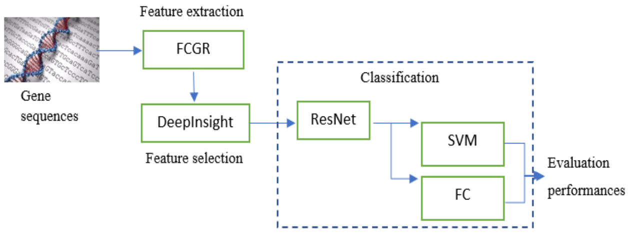


Figure 1. The proposed cancer disease multiclassification system.

2.1. Gene sequence database

In this work, used dataset downloaded from the DoCM website (11). It is a database of gene sequence mutations caused by different cancer disease types.

Here, we extracted 1903 mutations for 111 genes that belong to 18 cancer disease types, which are: skin cancer, breast cancer, leukemia, renal carcinoma, bladder carcinoma, ovarian cancer, brain cancer, colorectal cancer, endometrial cancer, esophageal carcinoma, gastric adenocarcinoma, head and neck cancer, hepatocellular carcinoma, lung cancer, myeloma, pancreatic cancer, prostate cancer, and cervix carcinoma. After eliminating redundancy, we obtained 472 samples. Figure 2 shows samples repartition by disease.

2.2 Frequency Chaos Game Representation

A genomic sequence is a sequence of nucleotides. It can be represented by a string using the alphabet of the four letters: Adenine A, Cytosine C, Guanine G, and Thymine T. The length of the complete sequence of an eukaryotic genome is about 10^6 characters. Hence, we use Frequency Chaos Game Representation (FCGR) to extract information from gene by looking for the k-mers nucleotide that composed the sequences.

Indeed, the chaos game representation is a 2D representation proposed by H.J.Jeffrey (12) to define the nucleotides patterns in a genomic sequence. Since gene sequences are long under their one-dimensional representation, the method converts the sequence into a 2D packed representation.

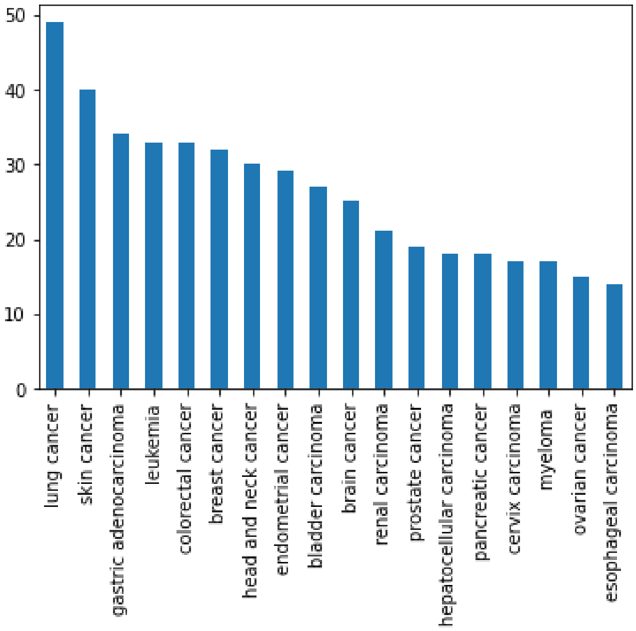


Figure 2. Data repartition by cancer disease type.

The Chaos Game Representation (CGR) provides a visual image that helps detecting and recognizing the characteristic patterns of a gene sequence.

To construct an FCGR, we begin by drawing the Chaos Game Representation (CGR), for this, we start by the position (0,0) then the four basic nucleotides are plotted at the four corners A (1,1), T(1,-1), G(1,1), and C(-1,-1). After that, for each nucleotide we move to the next location which is halfway between the nucleotide and the present location.

Figure 3 shows the CGR of the sequence “ATGC”.

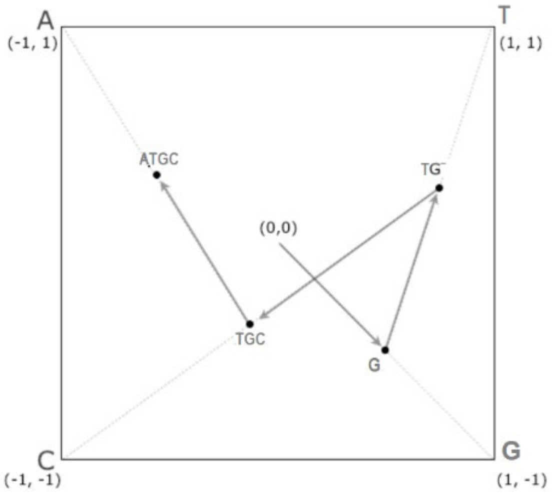


Figure 3. The CGR representation of the sequence ATGC.

Figure 4 shows the FCGR$_k$ representation of the MTOR gene with k is equal to 3, 4, 5, 6.

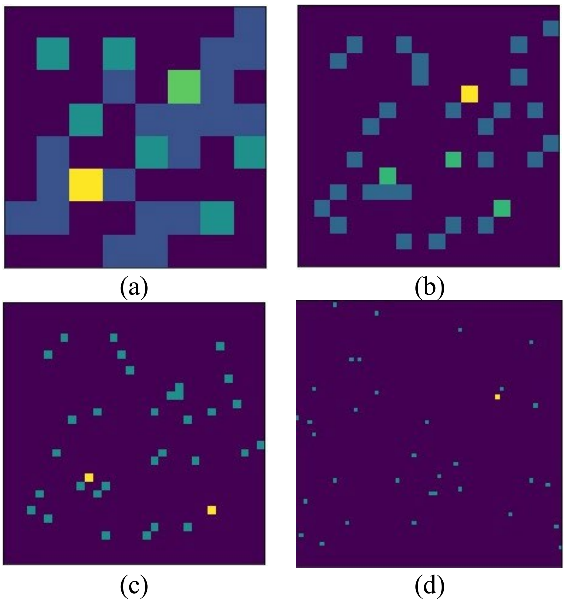


Figure 4. FCGR$_k$ representation of the MTOR gene with k = 3 (a), 4 (b), 5 (c), and 6 (d).

2.3 DeepInsight for dimension reduction

To classify complex datasets with several features, DeepInsight transforms 1D signal to 2D image by gathering similar features and giving them their average value (13). Here, we have to linearize the FCGR representation to be transformed by DeepInsight.

The DeepInsight model have two main hyperparameters:

(i) Data normalization**:**

For the data normalization, a min-max or a logarithmic method can be adopted.

- Min-Max normalization:

$$x_{scaled} = \frac{x - x_{min}}{x_{max} - x_{min}} \quad (1)$$

- Logarithmic normalization:

$$x_{scaled} = log\ (x + \|x_{min}\| + 1) \quad (2)$$

Where $x = (x_1, \dots, x_n)$ is a feature vector.

(ii) Dimension reduction:

For the dimension reduction, a Distributed Stochastic Neighbour Embedding or a Kernel PCA method can be adopted.

- *Distributed Stochastic Neighbor Embedding (t-SNE):* an unsupervised dimensionality reduction algorithm developed by Laurens van der Maaten and Geoffrey Hinton, it allows analyzing data described in a high dimensional space (14).
- *Kernel PCA (kPCA):* a generalization of PCA to non-linear dimensionality reduction (15). The two generally used kernels are polynomial and gaussien kernels.

In our work, we deal with a big number of features since the FCGR method generates 4^k features, for example for k = 7 we have 4^7 = 16384 features. Therefore, we use DeepInsight method to select the most relevant features.

For illustration, we provide in Figure 5, the DeepInsight output of the FCGR$_3$ of four cancers: breast cancer (BRCA), ovarian cancer (OV), esophageal carcinoma (GBML), and pancreatic cancer (KIPAN). The dark colours present the 3-mer nucleotides that are not frequent in the gene's sequences, While the light colours correspond to the most frequent 3-mers.

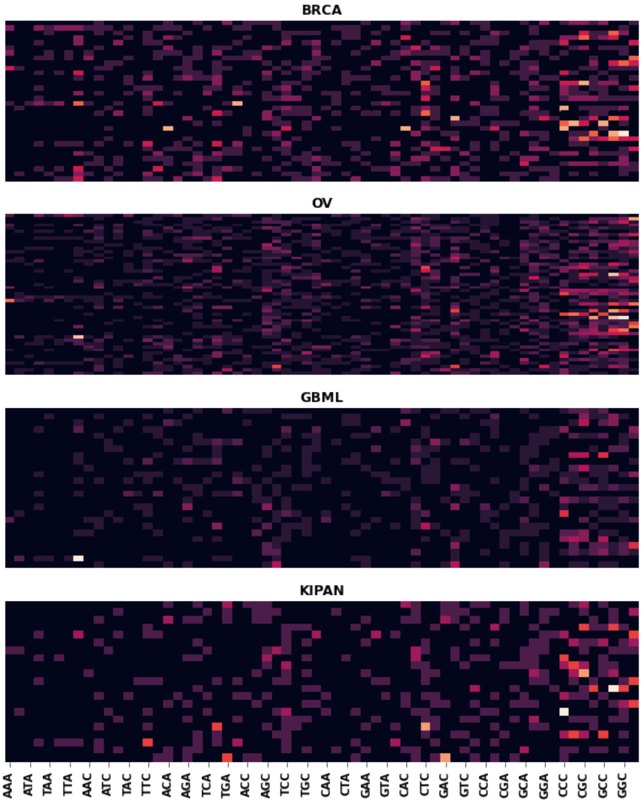


Figure 5. 3-mer nucleotides abundance in 4 diseases.

2.4 Support Vector Machine

SVM is a machine learning algorithm proposed by Vapnik in 1990 (16). Its objective is to find a hyperplane that

separates data in an n-dimensional space where n is the features number.
The main idea is to maximize the margin between the hyperplane and the data points.

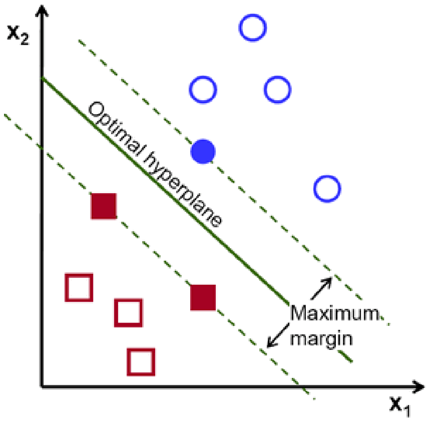


Figure 6. Support vector machine

This hyperplane is defined by equation (3):

$$\underset{w,h}{\boldsymbol{argmax}}\ \underset{k}{\boldsymbol{min}} \frac{l_{k\,(w^T x_k+b)}}{\|w\|} \qquad (3)$$

Where x = (x_1, ..., x_n) is the input data, ω = (ω_1, ..., ω_n) is the support vector, and *b* is the biais.
For real cases, datasets cannot be separated by a straight line. Therefore, for nonlinear data we use nonlinear SVM model which uses a kernel function to calculate similarity between two samples x_1 and x_2.
Examples of kernel functions are listed in Table 1.

Table 1. SVM kernel functions

Kernel	Mathematical function
Linear	$K(x_i,x_j) = x(i)x(j)$
RBF	$K(x_i,x_j) = \exp\,(-\gamma\|x(i) - x(j)\|^2)$
Polynomial	$K(x_i,x_j) = (\gamma x(i)x(j) + \mathrm{L})^D$
Sigmoid	$K(x_i,x_j) = \tanh\,(\sigma x_i^T x_j + r)$

2.5 Transfer learning

Transfer learning is a deep learning technique based on training model on a task then uses it when modelling a similar one to improve performance and reduces execution time. In fact, for traditional deep learning models, training needs a big amount of data, and it is computationally costly. The most used pre-trained models for images are AlexNet, GoogleNet, and VGGNet. They are convolutional neural network (CNN) architectures with multiple layers. AlexNet has only 5 layers. However, GoogleNet and VGGNet have more layers (17).Nevertheless, increasing the network depth could arise Vanishing gradient problem. This causes the gradient to become 0 or too large. Thus, when we increase the number of layers, the error rate of training and testing also increases.

2.5.1 Residual Network (ResNet)

The ResNet architecture introduces the concept of residual blocks to solve the gradient problem. In this network, a technique called skip connections is used (18) .The shortcut connection connects activations from one layer to other layers by skipping some intermediate layers. This forms a residual block as shown in Figure 7.

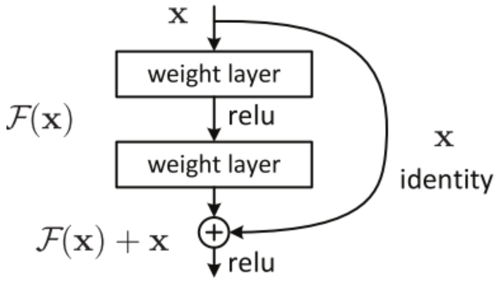

x
weight layer
relu
weight layer
$\mathcal{F}(\mathbf{x})$
x
identity
$\mathcal{F}(\mathbf{x}) + \mathbf{x}$
relu

Figure 7. Residual block architecture.

F(x) is the activation function. Here, w is the layer weight and b a bias term.

$$\boldsymbol{F(x) = wx + b} \qquad (4)$$

The ResNet architecture depends on layers number. Figure 8 shows the architecture of three ResNet variants which are ResNet50, ResNet101, and ResNet152 with respectively 50, 101, and 152 layers.

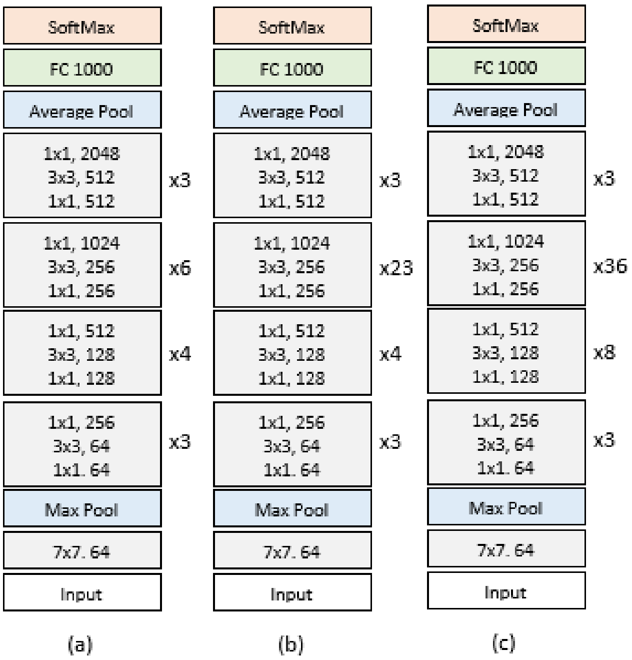


Figure 8. Architecture of: ResNet50 (a), ResNet101 (b), and ResNet152 (c).

In this figure, (k x k, n) is the kernels size of each block, and the numbers behind (x3, x4, . . .) is the repeating of each one.
The ResNet convolutional layers are used for features extraction. then, the multinomial classification is done by a fully connected layer. Finally, the SoftMax activation layer calculates the probabilities for each class.

2.6 Performance evaluation

To predict the proposed classifiers, we will use the following metrics: F1-score, accuracy, and Confusion matrix, whose equations are given by:

- F1-score:

$$\boldsymbol{F1} = 2 \times \frac{\boldsymbol{precision} \times \boldsymbol{recall}}{\boldsymbol{precision} + \boldsymbol{recall}} \quad (5)$$

Where

$$\boldsymbol{precision} = \frac{\boldsymbol{TP}}{\boldsymbol{TP}+\boldsymbol{FP}} \quad (6)$$

$$\boldsymbol{recall} = \frac{\boldsymbol{TP}}{\boldsymbol{TP}+\boldsymbol{FN}} \quad (7)$$

- Accuracy:

$$\boldsymbol{Accuracy} = \frac{\boldsymbol{TP}+\boldsymbol{TN}}{\boldsymbol{TP}+\boldsymbol{TN}+\boldsymbol{FP}+\boldsymbol{FN}} \quad (8)$$

- Confusion Matrix (CM):

An N by N matrix where N is the classes number. It presents a comparison between actual and predicted values. Terms are calculated using TN, TP, FN, and FP.

TP, TN, FP, FN are true positive, true negative, false positive, and false negative values.

3. Experiments and results

Initially, the input data extracted from the DoCM website, contain gene mutations caused by cancer disease. Hence, to predict cancer type we developed a python program to generate altered gene sequences. After eliminating redundant data, the length of the obtained dataset becomes 472.
To compare and evaluate our proposed models, data are split into training, validation, and test datasets with 302, 75, and 95 samples. Validation dataset is used to evaluate our models during the training phase.
Execution is done in a GPU environment.

3.1 FCGR for feature extraction

In this paper, our input data is a set of altered gene sequences. The objective of this work is to detect the cancer disease type by analysing the gene composition of nucleotides. FCGR is a technique for k-mer nucleotide extraction. In our work we use FCGR with k between 4 and 7. Those k-mers will be considered as features for our classification models.

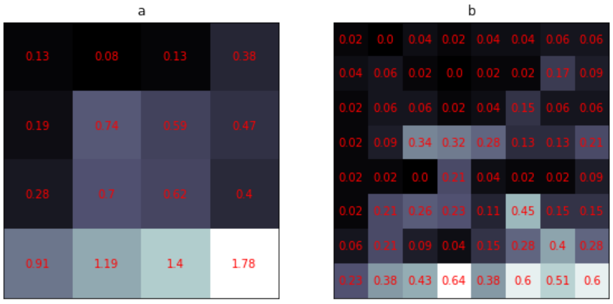


Figure 9. FCGR representation of a mutated sequence of the CTNNB1 gene by its 2-mer (a) and 3-mer (b) composition.

Figure 9 corresponds to the FCGR$_2$ and FCGR$_3$ whose k-mer matrices are presented by equations (9) and (10). As we can see, increasing the FCGR order increases the features number.

$$FCGR_2(s) = \begin{pmatrix} N_{AA} & N_{AG} & N_{GA} & N_{GG} \\ N_{AC} & N_{AT} & N_{GC} & N_{GT} \\ N_{CA} & N_{CG} & N_{TA} & N_{TG} \\ N_{CC} & N_{CT} & N_{TC} & N_{TT} \end{pmatrix} \qquad (9)$$

$$FCGR_3(s) = \begin{pmatrix} N_{AAA} & N_{AAT} & N_{ATA} & N_{ATT} & N_{TAA} & N_{TAT} & N_{TTA} & N_{TTT} \\ N_{AAC} & N_{AAG} & N_{ATC} & N_{ATG} & N_{TAC} & N_{TAG} & N_{TTC} & N_{TTG} \\ N_{ACA} & N_{ACT} & N_{AGA} & N_{AGT} & N_{TCA} & N_{TCT} & N_{TGA} & N_{TGT} \\ N_{ACC} & N_{ACG} & N_{AGC} & N_{AGG} & N_{TCC} & N_{TCG} & N_{TGC} & N_{TGG} \\ N_{CAA} & N_{CAT} & N_{CTA} & N_{CTT} & N_{GAA} & N_{GAT} & N_{GTA} & N_{GTT} \\ N_{CAC} & N_{CAG} & N_{CTC} & N_{CGT} & N_{GAC} & N_{GAG} & N_{GTC} & N_{GTG} \\ N_{CCA} & N_{CCT} & N_{CGA} & N_{CGT} & N_{GCA} & N_{GCT} & N_{GGA} & N_{GGT} \\ N_{CCC} & N_{CCG} & N_{CGC} & N_{CGG} & N_{GCC} & N_{GCG} & N_{GGC} & N_{GGG} \end{pmatrix} \qquad (10)$$

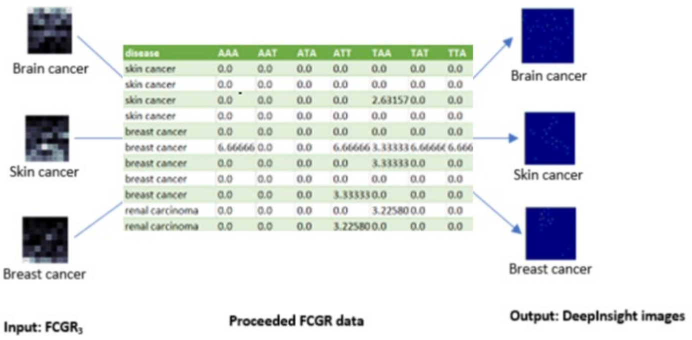
Where s is the gene sequence and N_x is the occurrence of the k-mer x.

3.2 Dimension reduction with DeepInsight

The number of generated features by FCGR technique is too large. For example, for $FCGR_6$ and $FCGR_7$ we have respectively 4096 and 16384 features. Therefore, to minimize the number of features, we use the DeepInsight technique.

It is a dimension reduction technique that converts a 1D signal into an image, allowing us to apply after that, the deep learning algorithms.

Before applying the DeepInsight technique, we modify the structure of our FCGR outputs with a python script to obtain all disease characteristics in one file. This process is presented by Figure 10.



disease	AAA	AAT	ATA	ATT	TAA	TAT	TTA
skin cancer	0.0	0.0	0.0	0.0	0.0	0.0	0.0
skin cancer	0.0	0.0	0.0	0.0	0.0	0.0	0.0
skin cancer	0.0	0.0	0.0	0.0	2.63157	0.0	0.0
skin cancer	0.0	0.0	0.0	0.0	0.0	0.0	0.0
breast cancer	0.0	0.0	0.0	0.0	0.0	0.0	0.0
breast cancer	6.66666	0.0	0.0	6.66666	3.33333	6.6666(	6.666
breast cancer	0.0	0.0	0.0	0.0	3.33333	0.0	0.0
breast cancer	0.0	0.0	0.0	0.0	0.0	0.0	0.0
breast cancer	0.0	0.0	0.0	3.33333	0.0	0.0	0.0
renal carcinoma	0.0	0.0	0.0	0.0	3.22580	0.0	0.0
renal carcinoma	0.0	0.0	0.0	3.22580	0.0	0.0	0.0

Figure 10. Example of FCGR processing by DeepInsight.

Note that for data normalization, we adopt in all this work the logarithmic scale to place the feature values between 0 and 1. As for the dimensionality reduction technique, we utilize t-SNE.

Figure 11 shows DeepInsight results of the $FCGR_3$ for three different cancer types which are: breast cancer, skin cancer, and brain cancer. Here, we illustrate the DeepInsight output for three examples of different genes that engender each disease. We remark that even with different genes, we have almost a similar representation by the DeepInsight method of the disease; Which renders the combination of the FCGR and the DeepInsight techniques a powerful tool for the disease characterization. Also, the difference between these diseases is clear in these images. Hence the usefulness of our approach in predicting the diseases.

After this step of data preparation, we will move to classification in the next section; where we will use the DeepInsight images for cancer disease multi-classification with a transfer learning algorithm.

3.3 Cancer disease multi-classification

In this section, we compare three ResNet variants on DeepInsight outputs generated from $FCGR_k$ with k = {4,5,6,7}.

In the ResNet architecture, the role of convolutional layers is the feature selection. The classification is done by the fully connection (FC) layer.

To improve our results, we use SVM classifier combined with convolutional layers of ResNet model. For each proposed model, genetic algorithms are used to tune hyperparameters.

Genetic algorithms (19) automate the process of hyperparameter tuning by using a fitness function to evaluate the performance of different hyperparameter configurations, and evolving a population of potential solutions through a process of selection, crossover, and mutation.

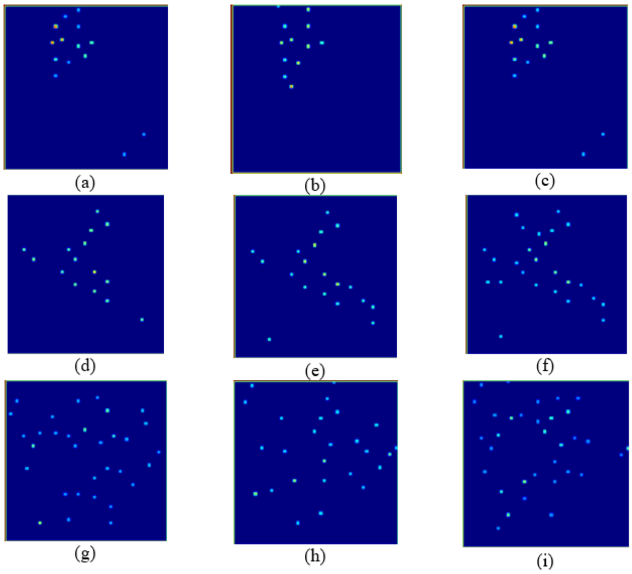


Figure 11. DeepInsight result by cancer type (a,b,c) reast cancer, (d,e,f) Skin cancer, (g,h,i) Brain cancer.

This can help to optimize the performance of machine and deep learning models and reduce the time and resources required for manual hyperparameter tuning.

Best results are obtained with 30 epochs, the used optimizer is Adam with learning rate=0.01. The batch size is 64 and the activation function of the fully connected layers is Relu.

Table 2 presents the obtained results when we used Resnet combined with a fully connected layer for classification, and thus for the four FCGR orders.

The results of a comparative study are presented in Table 3, where we considered SVM as classifier instead of the FC layer in the previous architecture.

5. Discussion

From Table 2, we can see that using the FCGR$_5$ data and the first architecture ResNet50 and FC gives the best result with an accuracy rate of 0.97 and F1-score of 0.961. However, with the second architecture that combines ResNet50 and SVM, we register the values of 0.986 as accuracy and 0.974 as F1-score using 6-mers nucleotide as features. For the two models, going deeper doesn't improve obtained performances. Since we have complex data, using a complex model leads to an overfitting problem.

Considering that best combination of architecture and data representation, we show in Figure 12 the corresponding confusion matrix of the model.

From this matrix, we notice that we have good performances for almost the 18 cancer diseases. Indeed, the least accuracy value is obtained for endometrial cancer which is 0.78; Whereas the highest value is of 0.99 for cervix carcinoma.

Table 2. Classification results with ResNet+FC layer for different FCGR orders

k	4		5		6		7	
	accuracy	F1score	accuracy	F1score	accuracy	F1score	accuracy	F1score
ResNet50	0.931	0,958	**0.970**	**0,961**	0.931	0,958	0.964	0,951
ResNet152	0.874	0.831	0.885	0.792	0.813	0.794	0.89	0.796
ResNet101	0.906	0.884	0.931	0.919	0.898	0.914	0.911	0.923

Table 3. Classification results with ResNet+SVM for different FCGR orders

k	4		5		6		7	
	accuracy	F1score	accuracy	F1score	accuracy	F1score	accuracy	F1score
ResNet50	0.948	0.932	0.969	0.964	**0.986**	**0.974**	0.952	0.956
ResNet152	0.893	0.792	0.928	0.905	0.925	0.917	0.945	0.933
ResNet101	0.906	0.885	0.909	0.927	0.896	0.918	0.899	0.906

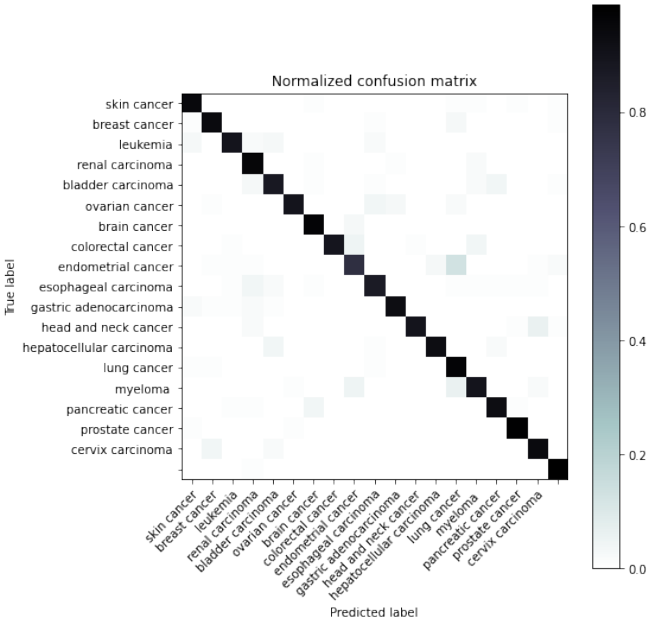


Figure 12. Confusion matrix obtained with Resnet50+SVM classifier and FCGR6.

Considering always the best model that consists of FCGR$_6$ data representation and the Resnet50+ SVM classifier, we tried to modify the dimension reduction option of Deepnsight. We recall that we opted for the logarithmic method to clamp feature values between 0 and 1 aiming at maintaining the consistency.

In Table 4 we evaluate the effect of varying the dimension reduction method in DeepInsight in terms of the classification accuracy. This inner parameter can be t-SNE or kPCA. The results show here that the best dimension reduction model to use in DeepInsight is t-SNE.

Table 4. Classification result depending on the DeepInsight hyperparameters.

	Accuracy	F1-score
t-SNE	0.986	0.974
kPCA	0.945	

5.1 related literature works

In this section, we present a comparative analysis of our model in relation to several state-of-the-art results mentioned in the introduction. The performance evaluation is summarized in Table 5, where our proposed model achieves an impressive accuracy of 0.97. Out of the referenced models, our approach outperforms six of them. Two models, specifically those based on Instantaneously Trained Neural Networks (DITNN) and 2D CNN, yield marginally higher results compared to ours. However, it's important to note that these models are designed exclusively for detecting a single type of cancer. In contrast, our model demonstrates its efficacy across 18 different cancer types. Additionally, we utilize generated sequences altered by miRNA, which differ from the use of existing patient sequences. Furthermore, the data we consider here enables very early screening of cancers even before the apparition of symptoms.

4. Conclusion

Cancer is one of the most common diseases in the world. However, most of the time it is detected in the final stage. Therefore, much research has been done to detect earlier this disease and begin the therapy on the time. Those researches are based on medical images and biological data which needs a lot of execution time and many pre-processing costly steps. Even if deep and machine learning are used to detect diseases, they are still expensive due to the large size of the used datasets.

The objective of this work is to early detect a cancer disease, not based on the classical data types (such as medical images and biological samples) but by scanning a patient’s genome. For this, we present a novel method of different cancer disease detection by analysing the gene sequence composition.

For the data collection, we started by downloading mutations caused by different cancer disease types from the DoCM website. This site presents alteration caused by 18 cancer disease type on 111 gene. Then we created a python script to generate altered genes

Table 5. Comparative results with some of the state-of-the-art approaches.

Work	Data type	Disease	Architecture	Accuracy
Alanazi et Al. (5)	Histology Images	Breast Cancer	CNN	87%
Shakeel et Al. (2)	Cancer imaging Archive	Lung cancer detection	DITNN	98%
Asuntha et Al. (3)	Lung cancer scan	Lung cancer detection	FPSOCNN	95%
Nawaz et Al. (8)	Histopathological images	Lymph node breast cancer	DensNet	95%
Mallick et Al. (9)	Brain MRI Image	Brain cancer	deep wavelet autoencoder and DNN	93%
Lopez-Garcia et al. (20)	Gene Expression Data	Lung Cancer	Transfer Learning + CNN	73%
Das et al. (21)	Gene Sequences	Liver cancer	Deep Transfer Learning	1D CNN model = 80% 2D CNN model = 98%
Our model	Altered gene sequences	skin cancer, breast cancer, leukemia, renal carcinoma, bladder carcinoma, ovarian cancer, brain cancer, colorectal cancer, endometrial cancer, esophageal carcinoma, gastric adenocarcinoma, head and neck cancer, hepatocellular carcinoma, lung cancer, myeloma, pancreatic cancer, prostate cancer, and cervix carcinoma	$FCGR_5$, DeepInsight, Resnet50+SVM	97%

sequences. After eliminating redundant data, the final database we obtained, contains 472 samples.

For the feature extraction step, we used the Frequency Chaos Game Representation (FCGR) method. Hence, the Features are the k-mers composition of a gene. For a given order k, 4^k is the number of the extracted features, which is considerably big. In this work, we used FCGR with order k=4, 5, 6, 7. Therefore, we thought of using DeepInsight as a dimension-reduction technique. Through this work, we noted that the combination of FCGR and DeepInsight allowed a good characterization of the mutated genes sequences.

As a final step, we applied Resnet: a transfer learning model for the feature extraction and multinomial classification of cancer disease types. This model is a pre-trained deep learning algorithm that helps in reducing execution time. Here, we adopted three Resnet architectures: ResNet50, ResNet101, and ResNet152.

In the ResNet models, the convolutional layers are used for feature selection, while classification is done by the fully connected layer (FC).

To further improve our results, we used SVM for classification. In this case, we combined SVM with the convolutional layers of the ResNet model. We tested both of the models with 30 epochs and a learning rate of 0.01.

Comparing the obtained results, we found that the best model consists of the combination of ResNet50 and SVM using 6-mers. This model provided 0.986 as accuracy and 0.974 as F1-score against an accuracy

rate of 0.97 and F1-score of 0.961 with ResNet50 and FC, using $FCGR_5$.

We also varied the dimension reduction method in DeepInsight, considering our best classification model, and we found that t-SNE is best suited to our work.

In this work, we use transfer learning algorithm ResNet with 50 layers combined with SVM to classify images generated with DeepInsight model. The created model shows its efficiency in cancer disease identification. Used datasets are the gene sequences altered by miARN which means that we don't need a complex biological treatment to detect cancer disease.

Based on these promising results, we expect to extend this work in such a way as to cover a larger range of diseases.

References

1. Jiang L, Zhu J. Review of MiRNA-disease association prediction. Current Protein and Peptide Science. 2020; 21(11): 1044–1053.
2. Shakeel PM, Burhanuddin MA, Desa MI. Lung cancer detection from ct image using improved profuse clustering and deep learning instantaneously trained neural networks. Measurement. 2019; 145: 702-712.
3. Asuntha A SA. Deep learning for lung cancer detection and classification. Multimedia Tools and Applications. 2020; 79(11): 7731–7762.
4. Allugunti VR. Breast cancer detection based on thermographic images using machine learning and deep learning algorithms. International Journal of Engineering in Computer Science. 2022; 4(1): 49--56.
5. Alanazi, Saad Awadh and Kamruzzaman, MM and Islam Sarker, Md Nazirul and Alruwaili, Madallah and Alhwaiti, Yousef and Alshammari, Nasser and Siddiqi, Muhammad Hameed. Boosting breast cancer detection using convolutional neural network. Journal of Healthcare Engineering. 2021; 2021.
6. Begum, Almas and Kumar, V Dhilip and Asghar, Junaid and Hemalatha, D and Arulkumaran, G. A Combined Deep CNN: LSTM with a Random Forest Approach for Breast Cancer Diagnosis. Complexity. 2022; 2022.
7. Mambou, Sebastien Jean and Maresova, Petra and Krejcar, Ondrej and Selamat, Ali and Kuca, Kamil. Breast cancer detection using infrared thermal imaging and a deep learning model. Sensors. 2018; 18: 2799.
8. Nawaz, Majid and Sewissy, Adel A and Soliman, Taysir Hassan A. Multi-class breast cancer classification using deep learning convolutional neural network. Int. J. Adv. Comput. Sci. Appl. 2018; 9(6): 316--332.
9. Mallick, Pradeep Kumar and Ryu, Seuc Ho and Satapathy, Sandeep Kumar and Mishra, Shruti and Nguyen, Gia Nhu and Tiwari, Prayag. Brain MRI image classification for cancer detection using deep wavelet autoencoder-based deep neural network. IEEE Access. 2019; 7: 46278--46287.
10. Alharbi, Fadi and Vakanski, Aleksandar. Machine learning methods for cancer classification using gene expression data: A review. Bioengineering. 2023; 10: 173.
11. Ainscough, Benjamin J and Griffith, Malachi and Coffman, Adam C and Wagner, Alex H and Kunisaki, Jason and Choudhary, Mayank NK and McMichael, Joshua F and Fulton, Robert S and Wilson, Richard K and Griffith, Obi L and others. DoCM: a database of curated mutations in cancer. Nature methods. 2016; 13(10): 806--807.
12. Jeffrey, H Joel. Chaos game representation of gene structure. Nucleic acids research. 1990; 18(8): 2163--2170.
13. Sharma, Alok and Vans, Edwin and Shigemizu, Daichi and Boroevich, Keith A and Tsunoda, Tatsuhiko. DeepInsight: A methodology to transform a non-image data to an image for convolution neural network architecture. Scientific reports. 2019; 9(1): 1--7.

14. Laurens van der Maaten and Geoffrey Hinton. Visualizing Data using t-SNE. Journal of Machine Learning Research. 2008; 9(86): 2579--2605.

15. Schölkopf, Bernhard and Smola, Alexander and Müller, Klaus-Robert. Kernel principal component analysis. In International conference on artificial neural networks; 1997: Springer. p. 583--588.

16. Sain, Stephan R. The nature of statistical learning theory. 1996..

17. Simonyan, Karen and Zisserman, Andrew. Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556. 2014.

18. He, Kaiming and Zhang, Xiangyu and Ren, Shaoqing and Sun, Jian. Proceedings of the IEEE conference on computer vision and pattern recognition. In ; 2016. p. 770--778.

19. Kramer, Oliver and Kramer, Oliver. Genetic algorithms: Springer; 2017.

20. Lopez-Garcia, Guillermo and Jerez, Jose M and Franco, Leonardo and Veredas, Francisco J. Transfer learning with convolutional neural networks for cancer survival prediction using gene-expression data. PloS one. 2020; 15.

21. Das, Bihter and Toraman, Suat. Deep transfer learning for automated liver cancer gene recognition using spectrogram images of digitized DNA sequences. Biomedical Signal Processing and Control. 2022; 72: 103317.