IDENTIFICATION OF CIRCULATING TUMOR DNA (ctDNA) BY USING A CONVOLUTIONAL NEURAL NETWORK AND SUPPORT VECTOR MACHINE

[1*]Raghi K.R
Departmen of Computer Science and Engineering
College of Engineering,Guindy
Chennai,TamilNadu

Abstract**-- Cell-free DNA (cfDNA) refers to short fragments of acellular nucleic acids detectable in almost all body fluids, including blood, and is involved in various physiological and pathological phenomena such as immunity, coagulation, aging, and cancer. In cancer patients, a fraction of hematogenous cfDNA originates from tumors, termed circulating tumor DNA (ctDNA), and may carry the same mutations and genetic alterations as those of a primary tumor. Thus, ctDNA potentially provides an opportunity for noninvasive assessment of cancer over the past few decades, cancer-specific mutations in ctDNA have been detected using a variety of untargeted methods such as digital karyotyping, personalized analysis of rearranged ends (PARE), whole-genome sequencing of ctDNA, and targeted approaches such as conventional and digital PCR-based methods and deep sequencing-based technologies. In this study, a new approach was proposed for the detection of cancer genes, which is an important step for the prediction of cancer. In the proposed approach, the gene sequences were digitized by mapping techniques. Following digitization, these DNA sequences were initially examined in two different ways as two-dimensional spectrogram images. Firstly, the digitized sequences were examined with the designed CNN model as a one-dimensional signal. Secondly, DNA signals were converted to 2D spectrogram images and examined with two different 2D CNN models. The proposed method indicated that effective features were extracted with 'Convolutional Neural Network' (CNN) for ctDNA detection and 'Support Vector Machine'(SVM) algorithm for classification of ctDNA and normal cell-free DNA (cfDNA). 'ctDNA' could provide clues for growth factor of cancerous cells, type of tumors, location, etc. Therefore, our work provides a new way for early detection and a new prospect for early cure. The application results showed that the system is ready to be tested with a larger dataset and different cancer types.**

Keywords: Cell-free DNA, Circulating Tumor DNA, Deoxyribonucleic acid, Convolutional Neural Network, Support Vector Machine, pattern recognition.

I. INTRODUCTION

The size of DNA pieces gives imperative data for the development of physical genome maps and genotyping. In particular, by realizing the DNA piece length and the DNA atom profile. It is conceivable to explore the properties of single DNA particles or DNA-protein cooperation. Cell-free, for instance, to recognize distinctive DNA optional structures and to set up if and in what way a ligand ties to DNA. When a cell in the body dies, it releases cell-free DNA (cfDNA) into the bloodstream. The term cfDNA is that broadly describes the different types of DNA freely circulating in the bloodstream at any given time. When the small fragments of DNA are released into the bloodstream, Researchers can capture the fragments and sequence them to look for mutations known as Next Generation Sequencing (NGS). ctDNA tells us if cancer is present, it can also tell, the best-individualized treatment selection. DNA sequence classification plays a vital role in computational biology. When a patient is infected by the virus, the samples collected from the cancer patient and the genomes are sequenced.

Adenine (A), cytosine (C), guanine (G), and thymine (T) are the four nucleotides that make up DNA Sequence .These are called the building blocks of cfDNA. The DNA of each affected virus is unique, and the pattern of arrangement of the each ucleotides determines the unique characteristics of a virus. Each form of nucleotide binds to its complementary pair on the opposite strand in double-stranded DNA. In this four types Adenine (A) and thymine (T) form a pair, while cytosine(C) and guanine (G) form a pair. Ribonucleic acid (RNA) may be single-stranded or double-stranded and they differ very much Therefore, the genome is the sequence of nucleotides (A, C, G, and T) for DNA virus. The DNA sequence is very long, having a length of around 32,000 nucleotides maximum, and it is challenging to understand and interpret. This raw DNA sequence cannot give as input to the CNN for feature extraction. It has to be converted into numerical representation before it is processed in the CNN and then classified using SVM. As completely risk-free surgery still not guaranteed, a comparison is analyzed in our work.

II. RELATED WORK

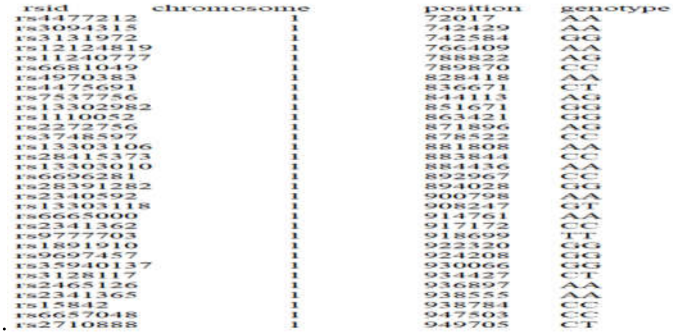
The segment extraction is done in Deep transfer learning for automated liver cancer gene recognition using spectrogram images of digitized DNA sequences in the 2021 strategy utilized for early detection of cancer. Leung [1] et al. (2020) identified genomic markers in Hepatitis B Virus (HBV) associated with liver cancer by comparing gene sequences of liver cancer patients and healthy individuals. In that study, Genotype B and C group HBV DNA sequences were collected from more than 200 patients and these authors extracted rules using the Evolutionary Algorithm based on the Rule-Learning algorithm and developed a new classification method using the Nonlinear Integral. The method provided an accuracy of >70% in the diagnosis of liver cancer. In the study, the author has proposed an ensemble method and compared several classification methods including Naive Bayes, Generalized Linear Model (GLM), the k-nearest neighbors (KNN), Support Vector Machine (SVM), and C5.0 Decision Tree (DT). Kaishan Tao [2] (2020) proposed a novel model: DNAs released from tumor cells into blood (circulating tumor DNAs, cDNAs) carry tumor-specific genomic aberrations, providing a non-invasive means for cancer detection. In this study, we aimed to leverage somatic copy number aberration (SCNA) in ctDNA to develop assays to detect early-stage HCCs. Methods: It conducted low-depth whole-genome sequencing (WGS) to profile SCNAs in 384 plasma samples of hepatitis B virus (HBV)-related HCC and cancer-free HBV patients, using one discovery and two validation cohorts. To fully capture the robust signals of WGS data from the complete genome, it developed a machine learning-based statistical model that is focused on detection accuracy in early-stage HCC. Findings: It built the model using a discovery cohort of 209 patients, achieving an overall area under the curve (AUC) of 0.893, with 0.874 for early-stage (Barcelona clinical liver cancer [BCLC] stage 0-A) and 0.933 for advanced-stage (BCLC stage B-D). Dimitrios Mathios [3] (2021) proposed a novel model for e an opportunity for cancer detection and intervention. Here, we use a machine learning model for detecting tumor-derived cfDNA through genome-wide analyses of cfDNA fragmentation in a prospective study of 365 individuals at risk for lung cancer, validate the cancer detection model using an independent cohort of DNA from non-cancer individuals and cancer patients. Combining fragmentation features, clinical risk factors, and CEA levels, followed by CT imaging, detected 94% of patients with cancer across stages and subtypes, including 91% of stage I/II and 96% of stage III/IV, at 80% specificity. Genome-wide fragmentation profiles across ~13,000 ASCL1 transcription factor binding sites distinguished individuals with small cell lung cancer from those with non-small cell lung cancer with high accuracy (AUC = 0.98). A higher fragmentation score represented an independent prognostic indicator of survival. This approach provides a facile avenue for tor non-invasive detection of lung cancer.

III. MATERIAL AND METHODS

Cell-free DNA (cfDNA) is short, extracellular, fragmented double-stranded DNA found in the plasma Plasma of data solid tumor in Cancer patients has been found to show significantly increased quantities of cfDNA. Due to the protection of genotype information from the patients, which is not needed for the fragmentation analysis, most cfDNA datasets are deposited in the controlled-access repositories in research labs. The data access in these repositories requires special and lengthy application processes and sometimes data transfer agreements that may take several weeks between the two organizations legal departments. Moreover, the cfDNA fragmentation patterns are inferred from the mapping locations of paired-end short-read sequencing, which are highly affected by the reads quality, length (h), and choices of the mapping strategy. These batch effects will significantly affect the downstream computational inference and data analysis. Currently, a centralized database with uniformly processed cfDNA datasets from a variety of physiological conditions is still not publicly the community.

- Source of data: UCF 101 (https://github.com/OpenGene/CfdnaPattern)
- Size of dataset: 3,755 DNA images from 67kbs to 500kbs
- Description: It includes DNA images like chromosomes, ctDNA, body fluids, like urine, pleural effusion, and cerebrospinal fluid.

Table 1 The dataset used for analysis



rsid	chromosome	position	genotype
rs4477212	1	72017	AA
rs3094315	1	742429	AA
rs3131972	1	742584	GG
rs12124819	1	766409	AA
rs11240777	1	788822	AG
rs6681049	1	789870	CC
rs4970383	1	828418	AA
rs4475691	1	836671	CT
rs7537756	1	844113	AG
rs13302982	1	851671	GG
rs1110052	1	863421	GG
rs2272756	1	871896	AG
rs3748597	1	878522	CC
rs13303106	1	881808	AA
rs28415373	1	883844	CC
rs13303010	1	884436	AA
rs6696281	1	892967	CC
rs28391282	1	894028	GG
rs2340592	1	900798	AA
rs13303118	1	908247	GT
rs6665000	1	914761	AA
rs2341362	1	917172	CC
rs9777703	1	918699	TT
rs1891910	1	922320	GG
rs9697457	1	924208	GG
rs35940137	1	930066	GG
rs3128117	1	934427	CT
rs2465126	1	936897	AA
rs2341365	1	938555	AA
rs15842	1	938784	CC
rs6657048	1	947503	CC
rs2710888	1	949705	CT

IV. PROPOSED METHOD

A. Outline of the proposed issue

With certain features regarding the condition of cancer, patients are put forward for consideration for MRI scans and biopsies. Certain drawbacks of MRI the time and expense of the procedure and it may produce claustrophobia, some people having a pacemaker, cannot be scanned safely. After having a lot of research perspective on MRI-based based brain tumor diagnosis and detection, considering the mentioned issues, we have proposed a way of early dnon-invasive technique, liquid biopsy, less expensively and with less time. The implementation of artificial intelligence makes the way easier and helps to get the desired result by the following solution.

B. Implemented Solution

In the first stage, CNN models were utilized to identify genomic markers of ctDNA and useful clinical information in predicting the recurrence of ctDNA and response to treatment. To achieve this, we first digitized the DNA gene sequence mapping techniques.

Subsequently, the digitized DNA sequences were initially exhausting using one-dimensional signal in the designed CNN model, and then the DNA signals were converted to 2D spectrogram images and examined with two different 2D-CNN models. In any case, the DNA design scattering requires pair-end pattern sequencing, so this part will be not open for single-end pattern sequencing data. Other than the DNA spread can be impacted by the plan cutting action, which is generally associated before the data pre-processing stage.

The block diagram of the proposed approach is shown in the figure below.

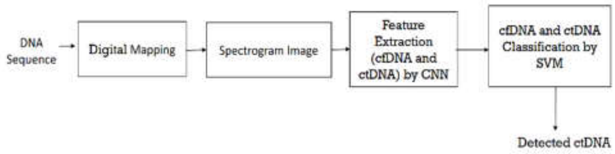

DNA Sequence
Digital Mapping
Spectrogram Image
Feature Extraction (cfDNA and ctDNA) by CNN
cfDNA and ctDNA Classification by SVM
Detected ctDNA

Fig. 1. Flowchart of ctDNA detection process

The models are compared considering the dataset size, it is highly difficult to perform training effectively with small datasets, particularly when training spectrogram images. Therefore, the method proposed was weaker to overcome this problem, either the dataset should be increased or another method that can classify with smaller datasets should be used to detect a technique for DNA pattern recognition using Convolutional Neural Networks and Support Vector Machine. To find ctDNA with genetic differences aids in cfDNA detection. DNAs can be detected and identified. After their identification DNAs As cfDNA were classified us using the ort vector machine from normal DNAs for canon monitoring and a prognosis. Before discovering certain therapeutic agents ctDNA can be a feasible biomarker for early cure. Better results can be obtained in CSF in plasma.This challenge of oncology can be solved with the clues into the mechanism underlying resistance to epidermal growth factor, provided by the biomarker. This method is based on spectrogram comparisons. A spectrogram gives a combined view of the local periodicity throughout the nucleotide sequence, and it was introduced as a tool for the visualization of DNA sequences. In spectrograms, some patterns, which are often related to the sequence function or structure, are observed. A periodicity of 10 bp reflects the DNA folding of bacteria. The first spectrum reducing method maps thymine (T), cytosine C, and adenine (A) occurrences into red, green, and blue (RGB) layers of the color spectrogram, respectively.

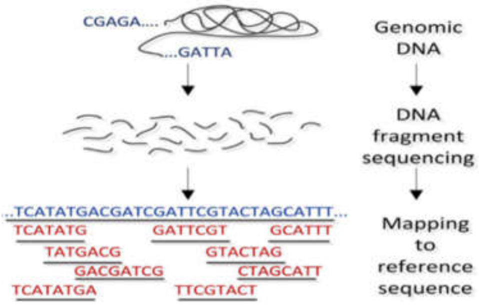


Fig. 2 Digital mapping

Let the DNA string be represented by $[n]$ of length k, where it is made of A, T, C, and G.

Consider

$$D\,[n] = \{\mathrm{A, T, C, G}\} \qquad (1)$$

For example, consider the DNA string of length 10 as in

$$D\,[n] = \cdots \mathrm{TTCACTAGCA} \qquad (2)$$

The color of the strip depends on the nucleotide composition and not on the quantity of the nucleotide; if the occurrence of thymine (A), cytosine (C), and adenine (A) is represented in the red, green, and blue layer, respectively, the codons AAT and ATT appear in the spectrogram as the violet strip.

C. Spectrogram Image

The digitized DNA sequences were converted to spectrogram images to obtain a 12-ms window width (Hamming windowing), a 8-ms overlap value, and a Fourier transform of 512 points. Spectrogram images were obtained using the Viridis color palette. It shows the signal samples that were digitized by mapping technique and then obtained spectrogram images. Given a DNA sequence S of length L, where S[j] = A, T, C, G, this is mapped into a numeric representation ŝ using the Voss representation which consists of generating four-vectors Si with i ∈ [A, T, G, C], where

$$\hat{s}_A[j]\begin{cases} 1, if\ S[j] = A, \\ 0, otherwise \end{cases} \qquad (3)$$

$$\hat{s}_G[j]\begin{cases} 1, if\ S[j] = G, \\ 0, otherwise \end{cases} \qquad (4)$$

$$\hat{s}_C[j]\begin{cases} 1, if\ S[j] = C, \\ 0, otherwise \end{cases} \qquad (5)$$

$$\hat{s}_T[j]\begin{cases} 1, if\ S[j] = T, \\ 0, otherwise \end{cases} \qquad (6)$$

Since each of the vectors ŝ can be seen as a digital signal that represents the patterns of occurrence of its corresponding nucleotide type, it is possible to perform a frequency analysis of each of those signals by estimating the digital signal spectra of the signal.

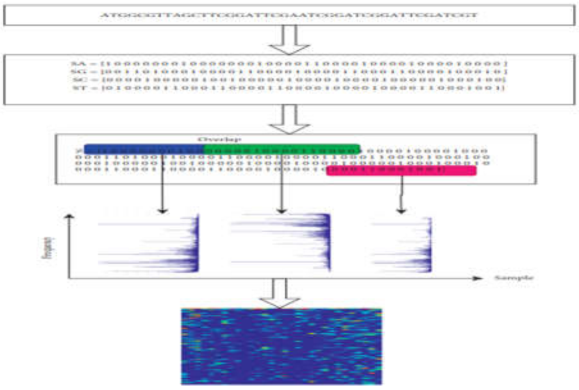


Fig.3 Spectrogram image

D. Feature Extraction by CNN

In this module, after the spectrogram image, A Convolutional Neural Network (CNN) is a Deep Learning algorithm and regularized version of the multilayered perceptron, where, each neuron in one layer is connected to all neurons of the next layer. It can take an image as input, assigns weights and biases to various objects in the, image, and differentiates one from the other. Here, the requirement of preprocessing is less than in other algorithms. Convolution multiplies two arrays of numbers with different sizes but with the same dimension and creates the third array with the same dimension. Here, each element of the image is added to its local neighbors, weighted by the kernel. For symmetric kernel, the center of the kernel is placed on the cur, rent pixel, and for the non-symmetric kernel; it needs to be flipped around the horizontal and vertical axis. Then kernel elements are multiplied with the overlapped pixel values and the obtained values are added, thus convolution works. This is a very interesting and useful tool for different signal processing, image processing application, and complex mathematical problems as well. A convolutional neural network can be formed by input and an output layer, as well as multiple hidden layers. A series of convolutional layers are used to form the hidden layers of CNN. The hidden layers also consist of additional convolutions such as pooling layers, fully connected layers, and normalization layers, and their inputs and outputs are masked by the activation function and final convolution. The activation function is a Rectified Linear Unit (ReLU) layer.

E. Convolutional Layer

Before programming a CNN, we have to make sure that each convolutional layer of the neural network should have the following features:

- Input is an array with shape (number of images) x (image width) x (image height) x (image depth).
- Convolutional kernels whose width and height are hyper-parameters, and whose depth must be equal to that of the image.

Convolutional layers convolve the input and pass its output to the next layer.

F. Pooling Layer

The Pooling layer is also responsible for reducing the spatial size of the convolved feature as a convolution layer. It is a very important step for capturing dominant features which are rotational and positional invariants. Five max-pooling lay a 2 × 2 window and stride In the last layer, Softmax is used to classify the input data from the previous layer. Rectified Linear Unit (ReLu) is used for the activation function

G. Fully Connected Layer

Fully connected layers connect every neuron of one layer with another layer. Its principle is the same as the traditional multi-layer perceptron neural network. An image is nothing but a matrix of pixel values, so we just flatten the image (e.g. a 3x3 image matrix into a 9x1 vector) and feed it to a fully connected layer for classification purposes. This layer uses a probabilistic neural network to generate the percentage of accuracy for matching.

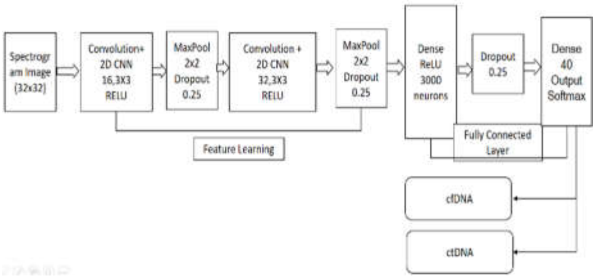


Fig.3 Feature extraction by CNN

In this module, Until or unless the result becomes convincing, they could repeat those layers. Then finally the output is flattened and passed to a fully connected layer. Here SVM to use as a classifier for the fully connected layer and the top 3 results will be taken. All the detected DNAs can be classified into ctDNA and normal DNA by using SVM. It proposed solution takes the SVM classifier because it's a powerful supervised classifier. In the field of machine learning, a confusion matrix is a specific table layout to visualize the performance of the algorithm. Support Vector Machine is a binary classifier, which in this case classifies normal DNA and ctDNA. But for any classification problem, correct or incorrect prediction and the accuracy can be measured correctly by a confusion matrix. That's why after SVM classification and then plot here confusion matrix. Detected ctDNA in this module, As a result of the said method, we have detected circulating tumor cells as mentioned before, and circulating tumors are using DNA can be detected by our approach

$$i_{Corrected}(\mathrm{i,j}) = l_{inp}(i,j) \text{- } l_{background}(i,j) \qquad (7)$$

Where $i_{Corrected}$ is the corrected image and (i,j) is the image coordinates and $l_{background}$ is the background image. $l_{inp}($ is the input image.

V. CLASSIFICATION AND PERFORMANCE EVALUATION

As SVM is a supervised classifier, we need to specify some feature labels. Based on this, it will classify those classes (ctDNA and normal DNA). Here we have given some feature labels i.e., length (for ctDNA it will be approximately 150 bp and fragmented), specificity (for ctDNA it will be high), and sensitivity (for ctDNA it will be high). In the field of machine learning, a confusion matrix is a specific table layout to visualize the performance of the algorithm. Support Vector Machine is a binary classifier, which in this case classifies normal DNA and ctDNA. But for any classification problem, correct or incorrect prediction and the accuracy can be measured correctly by a confusion matrix. That's why after SVM classification we would plot here confusion matrix. From this we would get specificity and sensitivity, by using this:

$$\text{Accuracy} = (TP+TN)/(TP+FN+TN+FP)$$

Where
True Positive TP (number of correctly identified Circulating free DNA (cfDNA))
False Negative FN (number of incorrectly defined Circulating free DNA (cfDNA))
True Negative TN (number of correctly identified Circulating Tumor DNA (ctDNA))
False Positive FP (number of incorrectly identified circulating Tumor DNA(ctDNA))
P = Positive, Observation is positive
N = Negative, Observation is negative

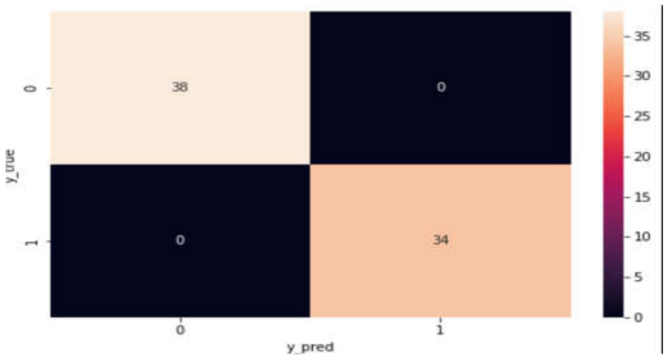


Fig.4 Confusion matrices of classification accuracy according to feature extraction performed with CNN

A. Comparison of models

When the models are compared considering the dataset size, it is highly difficult to perform training effectively with small datasets, particularly when training spectrogram images. Therefore, the first method proposed was weaker at this point. In order to overcome this problem, either the dataset should be increased or another method that can classify with smaller datasets should be used. Accordingly, the use of pre-trained networks known as transfer learning provides highly effective and successful results. The second method, CNN, which is a known model, achieved an effective classification with a successful training process. This method was a model with 1000 class output. In the third method, going one step further, a two-class dataset was utilized to yield a two-class output (fine-tuning) and it was revealed that the classification performance of the model increased further and reached 100%. The results were quite revealing in terms of understanding the classification capabilities of the models. In the comparison of the models with regard to training times, pre-trained models were found to be faster than the other models. A 100 × 1-dimensional signal was given as input in the first method and 224 × 224 × 3 images were given in the second and third methods. Although the input size was remarkably small, the retraining process was highly time-consuming. Taken together, these findings indicate that a pre-trained system is both faster and more successful.

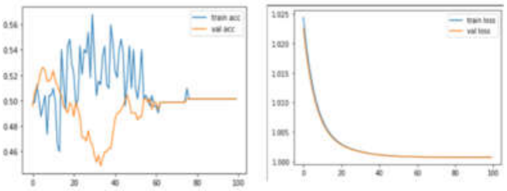


Fig.5 Accuracy and loss graphs of a fold in the training process

Permanent alteration of DNA sequence caused gene mutation; some specific genetic mutations can be found in the case of malignant tumors tumor. After ctDNA detection gene mutation can be detected by using machine learning in bioinformatics.

VI. CONCLUSION AND FUTURE WORK

The determination of sequence alteration can reveal more information about cancer. From the correct allelic sequence, we can predict the mutation hotspot and types of tumor, and tumor location can be found. Thus, early treatment would diminish the chances of death due to cancer and would save lives. In this proposed work, DNA design acknowledgment is a vital issue in bioinformatics and biomedical informatics in this paper. In this paper, the problem by utilizing the likelihood strategy and metric rather than customary recurrence metric are taken care of. And after that, it set forward neural system, which has preferable execution on time intricacy over some grouping arrangement calculations in a similar field. The aftereffects of the complexity tests demonstrate that Neural Network calculation can perceive DNA arrangements accurately and viably with no ambiguities. Prepared with examples, the system effectively grouped pictures given as information. DNA sequences with deep learning, which has become a popular method over the last few years. Unlike traditional machine learning techniques, in which data on feature selection, number of selected features, and experience of the expert designing the system are effective parameters for classification performance, deep learning models do not require manual feature extraction and extract features from the raw data. In this study, the classification performances of three different deep learning architectures were compared and the results demonstrated the superiority of the architecture that performs the fine-tuning to the final layers of the VGG16. We consider that the proposed method can be applied with different deep learning architectures and different DNA types. Future studies are warranted to increase the validity and reliability of the proposed model by using larger datasets

VII.References

[1] Bihter Das a,*, Suat Toraman,' Deep transfer learning for automated liver cancer gene recognition using spectrogram images of digitized DNA sequences', Elsevier- Biomedical Signal Processing and Control(2021).

[2] Hong Yang, Senior Member, IEEE, Kuo-Chuan Wu, Li-Yeh Chuang, and Hsueh-Wei Chang. (2020), 'DeepBarcoding: Deep Learning for Species Classification using DNA BarcodingCheng'-IEEE-ACM Transactions On Computational Biology And Bioinformatics.

[3] Otezatu I, Serdyuk O, Potapova G, Shelepov V, Alechina R, Molyaka Y, Anan'ev V, Bazin I, Garin A, Narimanov M, Melkonyan H, Umansky S, Lichtenstein AV. Genetic analysis of DNA excreted in urine: a new approach for detecting specific genomic DNA sequences from cells dying in an organism. Clin Chem 46:1078-1084 (ICON).:0576-0581, 2020.

[4] W. Seo, Y. Gao, Y. He, J. Sun, H. Xu, D. Feng, S.H. Park, Y.-E. Cho, A. Guillot, T. Ren, R. Wu, J. Wang, S.-J. Kim, S. Hwang, S. Liangpunsakul, Y. Yang, J. Niu, B. Gao, ALDH2 deficiency promotes alcohol-associated liver cancer by activating oncogenic pathways via oxidized DNA-enriched extracellular vesicles, J. Hepatol. 71 (5) (2019) 1000–1011.

[5] Y.-L. Chen, C.-J. Ko, P.-Y. Lin, W.-L. Chuang, C.-C. Hsu, P.-Y. Chu, M.-Y. Pai, C.- C. Chang, M.-H. Kuo, Y.-R. Chu, C.-H. Tung, T.-M. Huang, Y.-W. Leu, S.-H. Hsiao, Clustered DNA methylation changes in polycomb target genes in early-stage liver cancer, Biochem. Biophys. Res. Commun. 425 (2) (2012) 290–296.

[6] S. Bose, D.M. Tripathi, P. Sukriti, S.N. Sakhuja, S.K.S. Kazim, Genetic polymorphisms of CYP2E1 and DNA repair genes HOGG1 and XRCC1: Association with hepatitis B related advanced liver disease and cancer, Gene 519 (2013) 231–237.

[7] J. Atamaniuk. (2012), 'Apoptotic cell-free DNA promotes inflammation in hemodialysis patients', Nephrology Dialysis Transplantation, vol. 27, pp. 902–905.

[8] Jahr S, Hentze H, Englisch S, Hardt D, Fackelmayer FO, Hesch RD. (2018),'DNA fragments in the blood plasma of cancer patients: quantitations'.

[9]Liimatainen SP, Jylhv J, Raitanen J, Peltola JT, Hurme MA(2020),'T he concentration of cell-free DNA in focal epilepsy', Epilepsy Res 105(3):292-8, 2019, 49, 227–239.

[10] M.W. Snyder, M. Kircher, A. J. Hill, R. M. Daza, and J. Shendure(2016), 'Cell-free DNA comprises an in vivo nucleosome footprint that informs its tissues-of origin, Cell, vol. 164, no. 1/2, pp. 57–68.

[11] Otezatu I, Serdyuk O, Potapova G, Shelepov V, Alechina R, Molyaka Y, Anan'ev V, Bazin I, Garin A, Narimanov M, Melkonyan H, Umansky S, Lichtenstein AV. (2020), 'Genetic analysis of DNA excreted in urine: a new approach for detecting specific genomic DNA sequences from cells dying in an organism', Clin Chem 46:1078-1084 (IEMCON).:0576-0581.