International Journal of Engineering and Advanced Technology (IJEAT)
ISSN: 2249-8958 (Online), Volume-8 Issue-5, June 2019

ML Approach for Breast Cancer Detection using DNA Sequence Recognition

P. Sabitha, Kartik Gupta, Tejas Sharma, Ravi Kumar Singh, Jugnu Kumar

Abstract- Current approaches for DNA pattern recognition rely on traditional methods and algorithms of machine learning. Machine learning is a process that makes a computer or a program learn from experience. When this DNA sequence gets recognized, it can be utilized to scrutinize numerous applications in the field of bioinformatics and biomedical informatics. In this paper, we solve the problem of pattern recognition by the use of probability method and metric, then we put forward the concept of neural networks in place of some sequence alignment algorithms which intensifies the performance on time complexity. The use of neural networks will assist to recognize the sequences without any ambiguities. This result will serve as a utility for detecting breast cancer in the DNA by matching number of DNA's (sequences) which possess cancer and DNA's with no cancer cells, to diagnose whether cancer cells possess the human body or not. This method of prior detection of the disease using DNA overcomes the problem of complex techniques for diagnosis. This matching will involve another machine learning algorithm which will comprise of logistic regression, random forests. This method to recognize and match the DNA sequence to put forward as a application to detect cancer is just a sample model which has a lower accuracy rate, and does not solely depend on computer technologies. Medical research in deep levels is necessary to implement in daily life.

Keywords-Bioinformatics; DNASequencing; Classification; Machine Learning ; Neural Networks; Probability Method and metric ; Logistic Regression ; DNA Matching; Cancer Detection
Index Terms: About four key words or phrases in alphabetical order, separated by commas.

I. INTRODUCTION

If you put all the DNA molecules in your body end to end, the DNA would reach from the Earth to the Sun and back over 600 times,; so one would just wonder about its amount of complexity and the range of its storing genetic instructions for the development and functioning of living being. DNA being a double helix structure has two strands which store the same biological information. This. information is replicated at the time when strands get separated. Maximum part of DNA, specially for humans is non-coding which denote that these sections do not serve as patterns for protein sequences. A gene is a physical unit of heredity which is made up of DNA, and some of these genes act as instructions. Genes are short segments of DNA, but not all DNA is genes. Genes are only about 1-3% of your DNA.Of the 3 billion base pairs in the human genome, only 0.1% are unique to us. While that 0.1% is still what makes us unique, it means we're all more similar than we are different. DNA is a lot more complex than we know, and still the scientists are researching to completely study it. DNA sequencing is a method used to determine the precise order of the four nucleotide bases – adenine, guanine, cytosine and thymine - that make up a strand of DNA. These DNA are oftenly fragmented into small DNA strands either intentionally or spontaneously. These DNA sequence patterns need to be recognized so as to be utilized in number of applications in the field of bioinformatics. By using this biological technique, medical science becomes capable of investigating various diseases and genetic illnesses. Many mutations occur by faulty DNA sequencing. Modification of DNA is called as mutation. Mutations become very essential for evolution of our human body, otherwise with which a human cannot evolve with the growing needs of the body. Its advantages include the therapeutic discovery used in genome engineering, sociology, and the most important for forensic department for law enforcements. In short, The advantage of DNA Sequencing is being unlocking your profile genetically, but also help us to know about the spread of diseases from our DNA irrespective of being genetically or non genetically. It also tells what type of nutrition is most effective for different types of our work Every human being shares 99% of their DNA with the other human being, and even a parent-child share 99.5% of the same DNA. This fact can be widely used to examine whether the child which has inherited his/her ancestral properties will be prone to any disease. In this paper we look through identifying breast cancer in the DNA of the children at a certain age. One of the most important fact to be noted is, that only 10% of the cancers are hereditary, and we would work only on that 10% population. Rest 90% cancers occur due to our silly lifestyles and being not cautious about our health.. Therefore, the detection of cancer is just a wake up call for a individual to take proper measures for the future. If the body goes undetected, even medical science wouldn't guarantee his/her future health as it wholly depends on the individual's lifestyle and how he/she take cares. Cancer is the leading cause of rising mortality rate, accounting for large proportion of global deaths. We have more than 1 million breast cancers reported in a year only in India, taking with the fact that large amount of female population is still not aware of its early symptoms. The diagnosis of cancer becomes complicated because cancer is conglomeration of many related diseases that all involve uncontrolled cellular growth and reproduction.

Manuscript published on 30 June 2019.
***** Correspondence Author (s)
Kartik Gupta, CSE, SRM University, Chennai, India.
Tejas Sharma, CSE, SRM University, Chennai, India..
Ravi Kumar Singh, CSE, SRM University, Chennai, India.
Jugnu Kumar, CSE, SRM University, Chennai, India.
Mrs P. Sabitha, CSE,AP,ME, SRM University, Chennai, India..

The early diagnosis and prognosis of a cancer type have become a necessity in cancer research, as it can facilitate the subsequent clinical management of patients. Numerous diagnostic techniques have been researched and patented for early breast cancer detection. The best example being the imaging technology and the use of trending technologies like artificial intelligence and analytics for efficiency in cancer imaging, and also maintaining accuracy and cost of medical imaging for the early diagnosis of cancer as well as increasing patient confidence in crucial preventive screening programs.Screening can identify cancer in its early stages before symptoms become apparent. It's well established early detection leads to better outcomes, such as increased likelihood of a positive response to treatment, a greater probability of survival and less expensive treatment.

The WHO (World Health Organization) and other major bodies are encouraging the doctors, medical professionals, and all the people related with healthcare to educate patients about the value of early diagnosis,screening and spread awareness to the people about the basic symptoms and make them to participate in regular health care and checkup camps.WHO organizes many healthcare programmes throughout the world to connect with various NGO's, research institutes, multinational hospitals, and many doctors motivate and educate people,and uplift them from traditional mindsets. The NCCP (National Cancer Control Programmes) run by WHO is also a public health body specially proposed to decrease the number of cancer cases and deaths and improve quality of life of cancer patients. NCCP has a particular planning structure to reach the proposed target. This is done by implementing systematic, equitable and evidence-based strategies for prevention, early detection, diagnosis, treatment and palliation using available resources. This strategy and structure to reduce the cancer risks become resistant in some countries and areas due to factors like lack of resources, lack of capitals, lack of education. These situations are taken care by NCCP (run by WHO) and it makes sure that no matter what resource constraints a country faces it should help that country or an individual to reduce the cancer burden and improve services for cancer patients and their families. The basic structure of their planning involves three major checkpoints- 1. Where are we now ? 2. Where do we want to be ? 3. How do we get there ?, after they pass all the checkpoints of planning, they focus on implementing them using accurate data, including reliable cancer registries, monitoring and evaluation programmes to ensure programmes are appropriately prioritized and quality assured.

One of the best screening method for breast cancer is mammography, but surprisingly some evidences state that some women avoid that due to silly reasons, even though they know that mammography reduces breast cancer mortality. The false-positive test result is the primary factor behind the silly reason. The number of false positives can be high, one study showed that 23% patients experienced false positives on image readings for certain cancer screening and often lead to unnecessary invasive procedures and followed-up scans that increased anxiety for patients. This high rate of false positive tests in cancer screenings is due the part of fact that images are traditionally evaluated visually by radiologists, who are often very cautious of missing something. This paper highlights how advancements in medical science and technology have resulted in high end cancer detection with high accuracy. Technologies like artificial intelligence, analytics, machine learning, image processing, deep learning algorithms can overcome challenges in medical science such as false-positive test results. Advanced genomic testing is designed to help in identifying the DNA alterations that may drive the growth of a specific tumor. Information about genomic mutations are unique which may help doctors to identify treatments designed to targetthosemutations

The approach of machine learning makes computer to improve the experience that can be widely used for numerous applications. This paper deals with matching of DNA sequence of parents with that of children, to partially detect the probability of possessing cancer. Once the person is detected, he/she can take protective measures for not developing it. As stated earlier it wholly depends on the lifestyle of the individual which would decide whether the detected or undetected person will possess cancer or not.

II. LITERATURE SURVEY

1. The research paper from Research Gate states machine learning approaches in cancer detection by accurately distinguishing between benign and malignant tumors. Traditionally statistical methods have been used for classification of high and low risk cancer, to overcome the drawbacks of traditional methods machine learning is used to handle high dimensional data with increasing application in clinical decision support. This paper also discusses new research directions and highlights the main challenges related to machine learning approaches in cancer detection.
2. Other research paper throws light on the significance of DNA sequence, and how recognizing DNA sequence can lead us to several applications, one of the being species identification. In this work, effectiveness of supervised machine learning methods has been analyzed to classify species with DNA barcode.It used algorithms like simple logistic functions, random forests, k-nearest neighbour to identify species with partial DNA sequence with huge accuracy.
3. The objective from study of application from DNA is to use machine learning to predict suicidal and non-suicidal deaths from DNA methylation data. It used algorithms like support vector machines.This research constitutes a baseline study for classifying suicidal and non-suicidal deaths from DNA methylation data. Future studies with larger sample size may reduce the bias and improve the accuracy of the results.
4. Bioinformatics is the interdisciplinary science of interpreting biological data using information technology and computer science. Machine learning (ML) focuses on automatic learning from data set. Machine learning includes the learning speed, the guarantee of convergence, and how the data can be learned incrementally.This paper gives a review on the mechanisms of gene sequence classification using Machine Learning techniques, which includes a brief detail on bioinformatics, literature survey and key issues in DNA Sequencing using Machine Learning.
5. Another research paper focuses on distinguishing cell free DNA from other DNA samples using simple machine learning techniques like k- nearest neighbour, random forests, and support vector machines.

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication (BEIESP)

It works on the patterns and found that it could determine whether a sample is cfDNA or not by just looking into the first 10 cycles of its base content curves. The result of 1000 iteration and bootstrapping shows that all these classifiers can give an average accuracy higher than 98%, indicating that the cfDNA patterns are unique and can make the dataset highly separable.

III. ARCHITECTURE DIAGRAM

Architecture Diagram is portrayal of set of all components and elements which on giving a glimpse should give a person a basic outline and also interpret skeleton of the project. This diagram exclusively designed to understand the fundamental structure of our project clearly depicts the flow, and how the four modules are been implemented. The diagram is crystal clear and explains how the project is being done. It also briefs about the working of some algorithms. We will go through a quick description of the architecture and try to end up with some results.

The project starts with the preprocessing module where the DNA is taken as input in form of images and then processed with the help of Median filter, Canny edge detection, and DWT. These three techniques reduce the noise(variation in color and brightness), extract structural information and extract wavelet information from the image respectively. After we produce necessary data, we move through Feature Extraction Module where we use probability metric and method for DNA sequence recognition after fragmentation of DNA occur. This recognition comes to use when we utilize this for some application. This comes the way for the last two modules, Collaboration and Knowledge Management Module is responsible for algorithms like Logistic Regression, Random Forests, and also the use of Neural Networks. Logistic Regression is used to classify the DNA's into cancerous and non-cancerous, whereas Random Forests is used for increasing the performance of time complexity, decreasing time consumption and memory allocation. Neural Networks is used as a additional technology so that if the problem turns big, which when can be converted into small nodes and these small nodes produce us with the desired result. Finally we move to Detection and Analysis Module where we matched many cancer causing DNA's to normal DNA's to detect whether breast cancer cells are present in the DNA or not. Hence we complete with the basic architecture of the project. This diagram clearly signifies us importance of using algorithms at specific places. After this we move to the part of Modules, Implementation and try to show some healthy results from this research. We will follow with the conclusion and future enhancement part after we are done with implementation part.

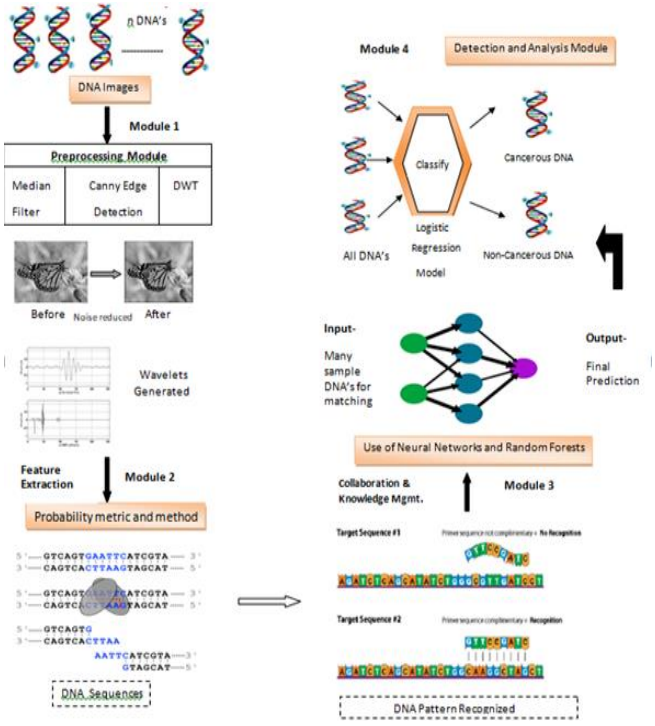

Fig 1:-Architecture diagram

IV. MODULES

The work or the paper can be classified into four different modules that define the flow and structure of our research. Each module contains a keen approach towards steps to accuracy, and each module is designed to be interlinked with the other. Some modules involves the implementation of algorithms, and the algorithms has been briefly drafted in uncomplicated manner so that even a naive person gets at least a rough idea about the algorithm. The modules are 1. Preprocessing Module; 2. Feature Extraction Module- Pattern Recognition; 3.Collaboration and Knowledge Management Module - Algorithms; 4. Detection and Analysis Module. Now these modules are illustrated as follows-

1. Preprocessing Module

This module can also be called the initiation module, but as this involves primary aim to preprocess the input it is recognized as the preprocessing module. As the main component of our research involves DNA, we take the digital image of the same and process it with various techniques to extract necessary data from it. The extraction of data is done with methods like Median filter, Canny edge detection and DWT etc., Median filter is a non-linear digital filtering technique usually utilized when we need to eradicate noise from an image. Noise refers to the scattered pixels randomly which leads to variation in brightness and color of the image. This noise reduction step is a must step for preprocessing and improves the result of later processing. This technique preserves the edges while removing noise simultaneously so that we effectively apply Canny edge detection. Canny edge detection smoothens the image with Gaussian filter to remove unwanted details and textures from the image and then applies some gradient operations to detect wide range of edges in images.

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication (BEIESP)

Canny edge detection is also edge detection operator that uses a multi stage algorithm that extracts some useful structural information from different angles and reduce the amount of data to be processed.Discrete Wavelet Transform abbreviated as DWT is a signal processing tool based on wavelet transformations. DWT extracts wavelets from the image by allowing both time and frequency analysis of signals simultaneously. Its use is more as a advantage as compared to other wavelet transforms as it captures both frequency and location information. These three techniques completely extricate vital data from the DNA (image), and we become ready to utilize this data so as to proceed to upcoming modules.

2. Feature Extraction Module- Pattern Recognition

As stated in the prior described module, the data extracted from the image can be now used for further process. This module works on how DNA sequence pattern can be recognized using probability method and metric instead of frequency metric. The use of probability method overcomes many drawbacks which includes decreased rate of accuracy, increased consumption of time. Using frequency metric, one cannot detect the pattern for multiple images in a short interval of time. To overpower this drawbacks, probability method and metric intensifies the accuracy rate and also exact and accurate matching can be attained. And then, we put forward the concept of neural network , about which we will study in the coming module. Neural networks is being used for better performance of time complexity, or it would have been far lower due to use of some sequence alignment algorithms in the same field. The use of Neural networks can recognise the DNA Sequences correctly and effectively without any ambiguities. In the forthcoming module we will give a brief instance about neural networks and also get to know about some algorithms used for matching of DNA's.

3. Collaboration and Knowledge Management Module- Algorithms

Until now we recognized the DNA sequence using probability method and metric and also introduced the concept of neural networks. This module will take us through a long tour with many checkpoints, where each checkpoint would mean a algorithm or a concept. We will start with the first checkpoint, understanding the notion of neural networks. Neural networks are inspired by biological neurons, where biological neurons refers to our neural network of our brain which contain n number of neurons which communicate with each other and thereby generating a electronic pulse as a output. Neural networks also works the same way, where it is made up of n number of nodes, each node is responsible for solving a small part of problem , and when all these nodes work together, we get a desired result. Linear Regression is the most basic algorithm practised by beginners who want to pursue their future in machine learning. Regression is a method of modelling a target value based on independent predictors. It is used for forecasting and finding cause and relationship between variables of a problem. Whereas Linear Regression uses a single independent variable and produces a relationship between the dependent and independent variable. Linear Regression is used for prediction, where as logistic regression is used primarily for classification. Here in our project we use the algorithm of logistic regression to classify the DNA sequences as cancerous or non-cancerous by matching number of DNA's which influence cancer with normal DNA's to detect whether the normal DNA possess any cancer development.

Random Forest algorithm is often used to reduce the complexity when we have huge amount of samples. This algorithm can be easily understood when we take our own situation as a context. Suppose we have 1000 DNA's for observation and matching. Random Forest algorithm will take any random 100 DNA's for matching and repeat the process for 10 times and then it makes a final prediction which is the mean of each prediction. This way the algorithm did run for mere 10 times rather than 1000 times thus reducing the time complexity and memory utilized. This algorithm finishes all our checkpoints of our long path and we are well equipped to move to the last and final module. modules where we summarize and try to produce a healthy output. We analyzed many sequences and tried to detect whether the normal human body will develop cancer or not in the future. Well as stated earlier, it is only a insignificant way, and it wholly relies on the individual's lifestyle, his daily habits and his/her awareness about symptoms, and proper measures.The modules put down here are explained in very simple language so that a person with a basic understanding about algorithms will get a brief idea about this. There we complete all our modules and finally boast about our research. Now we move to perceiving the architecture of the project.

4. Detection and Analysis Module

Thus we successfully grasped enough about the algorithms and how they work out, although we didn't dig deep into their implementation.The first two modules which included preprocessing and feature extraction were ended up with a result of recognized DNA pattern with the help of some processing tools and a simple algorithm. The last module told us how we tried to match the different DNA sequences to detect whether the normal DNA possess any cancer quality using many algorithms, thereby called as Collaboration and Knowledge Management Module. So this module called as Detection and Analysis Module is just a mere conclusion of all the So to fulfill this task we will use the algorithm called logistic regression algorithm,this algorithm will help in segregating the recognised DNA sequences into cancerous and the non cancerous DNA[Fig 3] ,to apply this algorithm we will be needing huge amount of data which will increase the time complexity of the process,so to decrease its time complexity and increase the efficiency of the result we use the random forest algorithm after this process we will be able to identify the DNA that are affected with cancer and the DNA which are normal.Then after our last process will be to summarize and give a healthy output.

V. IMPLEMENTATION & RESULT

We have completed our whole project and its research by using the four modules first the image of DNA will be taken ,the image would have lot of unnecessary data in it so we will extract the useful data and the unnecessary data will be removed with the help of of multistage algorithm,Then after getting the extracted material we start the process of recognition.In the next step with the help of probability method and metric algorithm the DNA patterns are recognised and are matched with there correct sequences this part is also told as DNA sequencing[Fig 2].Our aim is to tell that

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication (BEIESP)

whether the DNA which we have recognised has cancer or not.We will try to explain some of the concepts in our project algorithmically. We take each primary concept from each module. Starting from first module, we look through how actually median filter works. Median Filter just as normal filtering technique considers each pixel of the image, and looks its nearby surroundings i.e., its neighbors. It takes all the pixel values of its neighbors, and a median is found by first sorting all the pixel values. Now that pixel value is compared with median pixel value. The median value is a more robust than mean as it doesn't create unrealistic pixel value. This approach makes median filter better at preserving edges of the image for canny edge operator to work. The second module is about DNA Sequencing. When DNA is fragmented either spontaneously or intentionally, its sequence needs to be recognized so as to be utilized. The third module goes through Logistic Regression. The graph [Fig 3] clearly tells us how a result is obtained after we apply this algorithm. This algorithm doesn't try to predict the value given a input value, but the output obtained is in the form of probability, that the given input belongs to certain class. We also know that P(+) + P(-) = 1. Therefore, the output of Logistic Regression always lies in [0,1]. The dividing line in between is called as linear discriminant because it helps the model discriminate between the different points. The final module briefs us about the result.

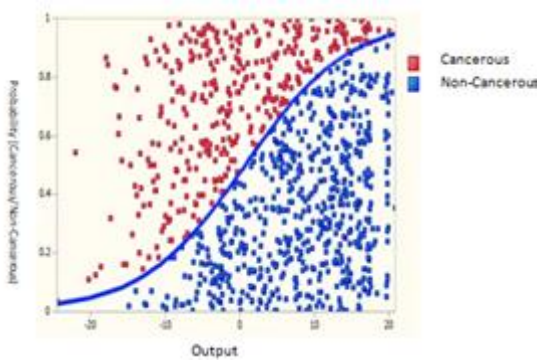

Fig 2:- Result after logistic regresssion

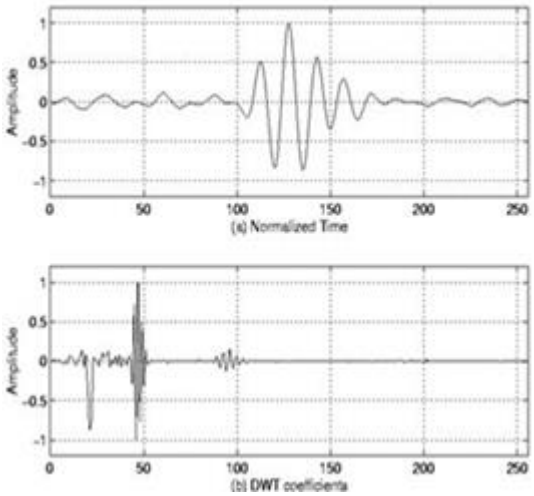

Fig 3:-Wavelet generated

VI. CONCLUSION

Deoxyribonucleic acid, DNA is a double helical molecular structure which carries the genetic instruction which is responsible for growth,development,functioning and reproduction of all living organisms. Genes are only about 1-3% of your DNA. Out Of the 3 billion base pairs in the human genome there are only 0.1% are unique to us.. DNA is a lot more complex than we know, and still the scientists are researching to completely study it. DNA sequencing is a method used to determine the order of four nucleotide bases of DNA that is adenine,guanine,cytosine and thymine these four makes a strand of DNA.With the help of DNA sequencing we can study the DNA and check for the various things like diseases,genetic illness,allergies,with the help of this we can also predict problems that can be carried from parent DNA to child DNA.The concern of our project is to predict or to find the perpetuation of breast cancer in humans.Cancer is still one of the leading cause of rise in mortality rate there are lot of people who are dying globally because of cancer. We have around 1 million breast cancer reported in a year only in india.There are large number of female population who are still not aware of the early symptoms of the breast cancer.There are various technologies such as artificial intelligence which can be used for the early detection of the cancer.So we are using the method of DNA sequencing to predict the chances of cancer cells.Our method of sequencing is to find or predict the chances of cancer cells in a body but if the the body goes undetected that does not means that he/she cannot have cancer in the future because that wholly depends on the individuals lifestyle and his/her habits.Our research is wholly based on that 10% of cancer which comes from genetic transfer.The work is being classified in to four modules. The first module will be the preprocessing module the main aim of this module is to preprocess the input,in this we take the digital image of the DNA and extract the necessary data from it this is done with the help of methods like Median filter,Canny edge detection and DWT.In this process the noise is eradicated and useful information is taken.The second module is a feature extraction module in which different patterns are recognized with the help of probability method and metric we recognised the DNA and found there exact matching. The third module is collaboration and knowledge management module in which we have used various algorithms.In this first the DNA sequence is classified into cancerous and non-cancerous by matching the normal DNA with cancerous DNA with the help of logistic regression algorithm and then we use Random forest algorithm to reduce the complexity this algorithm is highly beneficial when we have huge amount of samples.The fourth module was detection and analysis module, the cancer cell is detected and the analysis of cancer cell is done and its level is checked.Through these modules we complete our research.

FUTURE ENHANCEMENT

In the coming or near future there will be lot of scope of our project because as everybody knows that the era which is now coming is the era of technology, and this project will create a lot of ease in detecting the breast cancer and the patients will not need to go through long procedures and wait for long time.Through this project the detection will be done in less time.In the future there are chances that patients of the breast cancer will increase if the situation remains the same ,so our project will be very helpful because it will give the result in very short period of time.

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication (BEIESP)

REFERENCES

1. "Machine Learning Approaches in Cancer Detection and Diagnosis: Mini Review", ResearchGate, 2017.
2. "Machine Learning in Bioinformatics: A Novel Approach for DNA Sequencing",Fifth International Conference on Advanced Computing & Communication Technologies,2015.
3. "Automated DNA Fragments Recognition and Sizing Through AFM Image Processing", IEEE TRANSACTIONS ON INFORMATION TECHNOLOGY IN BIOMEDICINE, 2005.
4. "Species Identification using Partial DNA Sequence: A Machine Learning Approach", IEEE 18th International Conference on Bioinformatics and Bioengineering, 2018.
5. "Learning from Longitudinal Data in Electronic Health Record and Genetic Data to Improve Cardiovascular Event Prediction", Scientific Reports, 2018.
6. "DNA Methylation Data to Predict Suicidal and Non-Suicidal Deaths: A Machine Learning Approach", IEEE International Conference on Healthcare Informatics, 2018.
7. "https://towardsdatascience.com/introduction-to-machine-learning-algorithms-linear-regression-14c4e325882a"
8. "https://hackernoon.com/introduction-to-machine-learning-algorithms-logistic-regression-cbdd82d81a36"
9. "https://www.analyticsvidhya.com/blog/2014/06/introduction-random-forest-simplified/"
10. "https://softwareengineering.stackexchange.com/questions/72093/what-is-a-neural-network-in-simple-words

AUTHORS PROFILE

Kartik Gupta, Computer Science And Engineering, SRM IST,Ramapuram

Tejas Sharma, Computer Science And Engineering, SRM IST,Ramapuram

Ravi Kumar Singh, Computer Science And Engineering, SRM IST,Ramapuram

Jugnu Kumar, Computer Science And Engineering, SRM IST,Ramapuram

Mrs.P. Sabitha Asst Professor, Department of Computer Science and Engineering, SRM IST, Ramapuram