IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Vol. 10, Issue 4, April 2021

DOI 10.17148/IJARCCE.2021.10475

Microarray Based Disease Diagnosis Using Deep Learning Techniques

M. Ananda Kumar[1] M.E., MBA., A. Manju[2] and S. DilipKumar[3] M.Tech., (Ph.D)

Assistant Professor, Department of Computer Science & Engineering, Arasu Engineering College, Kumbakonam, India[1]

PG Scholar, Department of Computer Science & Engineering, Arasu Engineering College, Kumbakonam, India [2]

Assistant Professor, Department of Computer Science & Engineering, Arasu Engineering College, Kumbakonam, India[3]

Abstract: Clinical time has created with a terrible degree of accomplishments in illness design expectation, anticipation and fix with the progressions of information mining procedures. Among different mining strategies highlight determination possesses a key part for improving precision in any sort of forecasting or fix of sicknesses. This paper presents SeQual, a versatile device to productively perform quality control of enormous genomic datasets. Our tool currently supports in excess of 30 distinct tasks (e.g., separating, managing, designing) that can be applied to DNA/RNA peruses in FASTQ/FASTA organizations to improve subsequent downstream investigations, while giving a straightforward and easy to use graphical interface for non-master clients. Hence it is treated as a fundamental before work of any sort of mining methods. Microarray information has a high element of factors and a little example size. In microarray information examinations, two significant issues are the ways to pick qualities, CNN give solid and great forecast to illness status, and how to decide the last quality set that is best for grouping. Relationship among genetic markers mean one can misuse data excess to possibly diminish order cost as far as time and calculation cost. So in this undertaking, CNN can actualize the structure to foresee the sicknesses utilizing advancement and order calculation like Genetic calculation and Semi-administered profound learning calculation with improved exactness rate. This paper presents an overview of different sickness order techniques and assesses these proposed strategies dependent on their arrangement precision, computational time and capacity to uncover quality data. We have likewise assessed and presented different proposed quality determination technique.

Keywords: next-generation sequencing (NGS), SeQual, DNA/RNA, FASTQ/FASTA, microarray, Genetic algorithm, deep learning.

I. INTRODUCTION

The advancement of Next-Generation Sequencing (NGS) technologies [1], [2] has reformed biological research in the course of the most recent decade by definitely diminishing the expense of DNA/RNA sequencing and essentially expanding the throughput of produced information. Large information examination is the regularly complex interaction of analyzing huge information to uncover data like secret examples, connections, market patterns and client preferences that can help associations settle on educated business choices. On a wide scale, information investigation innovations and methods give a way to break down informational collections and remove new data. Which can help associations settle on educated business choices? The nature of NGS information is viewed as vital for different downstream examinations, for example, gene expression studies and genome arrangement gathering [3]. The partner editorial manager organizing the audit of this original copy and approving it for publication was Juan Wang. The pipeline. For example, changing the info information from FASTQ to FASTA organization might be fundamental if any bioinformatics application can just work with information put away in the last configuration. Right now, there are a few tools to perform quality control and preprocessing of raw NGS information to guarantee the vital quality for additional handling [4], [5]

A. *Big Data Analytics*

Large information investigation applications frequently incorporate information from both inside frameworks and outer sources. Big data has become progressively valuable in production network investigation. Big supply chain analytics examination uses large information and quantitative strategies to improve dynamic cycles across the store network. In particular, large inventory network examination extends datasets for expanded investigation that goes past the customary inward information found on big business asset arranging (ERP) and production network the executives (SCM) frameworks. Likewise, enormous inventory network examination executes exceptionally powerful measurable techniques on new and existing information sources. The bits of knowledge assembled encourage better educated and more compelling choices that profit and improve the inventory network. Likely traps of enormous information examination activities incorporate an absence of inward investigation abilities and the significant expense of recruiting experienced information researchers and information designers to fill the holes.

B. *Microarray Technology*

Microarray innovation has gotten one of the fundamental devices that numerous scholars use to monitor genome wide articulation levels of qualities in a given creature. A microarray is normally a glass slide on to which DNA atoms are fixed in a methodical way at explicit areas called spots (or highlights). A microarray may contain a large number of spots and each spot may contain a couple million duplicates of indistinguishable DNA atoms that particularly correspond to a gene. The DNA in a spot may either be genomic DNA or short stretch of oligo-nucleotide strands that relate to a quality. Quite possibly the most famous instruments for quality control which has been broadly utilized in numerous new natural examinations [11], [12]. Our device additionally gives this usefulness (and surprisingly more) yet in a fundamentally lower runtime by completely misusing the equal handling abilities of Spark. Despite the fact that there are a couple of equal apparatuses to eliminate copy DNA/RNA groupings (one explicit activity that can be utilized for quality control) on circulated memory frameworks [13], [14], up as far as anyone is concerned, SeQual is the main openly accessible instrument planned for this kind of equal frameworks that gives full usefulness (in excess of 30 tasks) rather than just permitting to eliminate copy reads.

II. RELATED WORK

"Review of current methods, applications, and data management for the bioinformatics analysis of whole Exome sequencing" This paper manages the issue of feature subset selection (FSS) in the feature subset selection of directed characterization. Directed grouping is probably the information mining/AI strategy that is most regularly utilized practically speaking. numerous bioinformatics devices carried out on top of Big Data preparing systems, for example, Hadoop [15] and Spark [9] Feature subset determination is a critical issue in the datamining characterization task that assists with getting more minimal and reasonable models without debasing (or in any event, improving) their presentation. In this work we center on FSS in high-dimensional datasets, that is, with an enormous number of prescient ascribes. For this situation, standard refined covering calculations can't be applied as a result of their intricacy, and computationally lighter filter wrapper calculations have as of late been proposed [17]. In this work we propose a stochastic calculation dependent on the GRASP meta-heuristic, with the principle objective of accelerating the component subset choice interaction, fundamentally by lessening the quantity of covering assessments to complete. Handle is a multi-start helpful strategy which builds an answer in its first stage, and afterward runs an improving stage over that arrangement. Handle would then be able to be seen as a worldwide enhancement technique, which normally improves the nature of the arrangement acquired by explicit heuristics.

"Quality control and preprocessing of metagenomics datasets" SeQual attempts to combine the functionality and ease of use of PRINSEQ along with the presentation of PRINSEQ++ yet in a distributed manner way depending on Big Data advancements. Truth be told, the misuse of Big Data groups to speed up the capacity, handling and representation of huge NGS datasets has been as of late investigated in various past works. For example, Time arrangement is an arranged grouping of perceptions that are equitably divided at uniform time stretches and estimated progressively. Expectation of time arrangement utilizes a succession of verifiable qualities to build up a model for determining future qualities [18] [19]. The nature of burden estimates essentially affects the financial activity of the electric utility since numerous choices depend on these gauges. Exact forecast will decrease the danger of the executives' dynamic in government and industry. As of late, time arrangement forecast has gotten significant applications in different documented, like financial matters and business, science, medication, substance, and designing. Burden figure has been generally worried in current force framework arranging, activity, and control.

"Classification of genes for disease identification using data mining techniques" Assessing their significance scores to express a cut-off schema where an upper/lower bound is forced to pick highlights with the best scores. Different apparatuses (FaQCs, FastProNGS) don't uphold FASTA as info design, while additionally give essential UIs simply restricted to order line communication. Besides, there are instruments that simply appear to be as of now inaccessible as their sites don't longer work (NGS-QC, QC-Chain) Wrapper strategies utilize a classifier to track down the most discriminant include subset by limiting a mistake forecast work. Especially, GS from microarrays is routed to distinguish/find those qualities which are communicated differentially as per a decided objective infection (in particular instructive qualities) [6]. GS strategies have been isolated into the accompanying four classifications: channels, coverings, installed and group.

III. PROPOSED METHODOLOGY

Microarray innovation has spread the word about the cutting edge natural examination sets of information will be utilized to prepare the profound learning conventions to arrange infections as per their gene example by allowing synchronous study of gene expression data including huge part of the genome. that right now gives a full set of 33 activities for performing quality control and preprocessing on raw NGS datasets. It can get as input either single-end or matched end DNA/RNA groupings, which can be put away either in FASTA or FASTQ documents, as these are the most popular unaligned sequence formats designs. The result of this study will give data with respect to the effectiveness of the AI methods, Convolutional Neural network strategy. In combination with classification techniques, this process can be helpful to help clinical management decision for individual patients, for example in oncology. Standard measurement approaches in grouping or expectation to predict don't function well when the number of variables quantity of factors p (qualities) extremely far too exceeds the number of testing sample n which is the situation in gene microarray expression data.

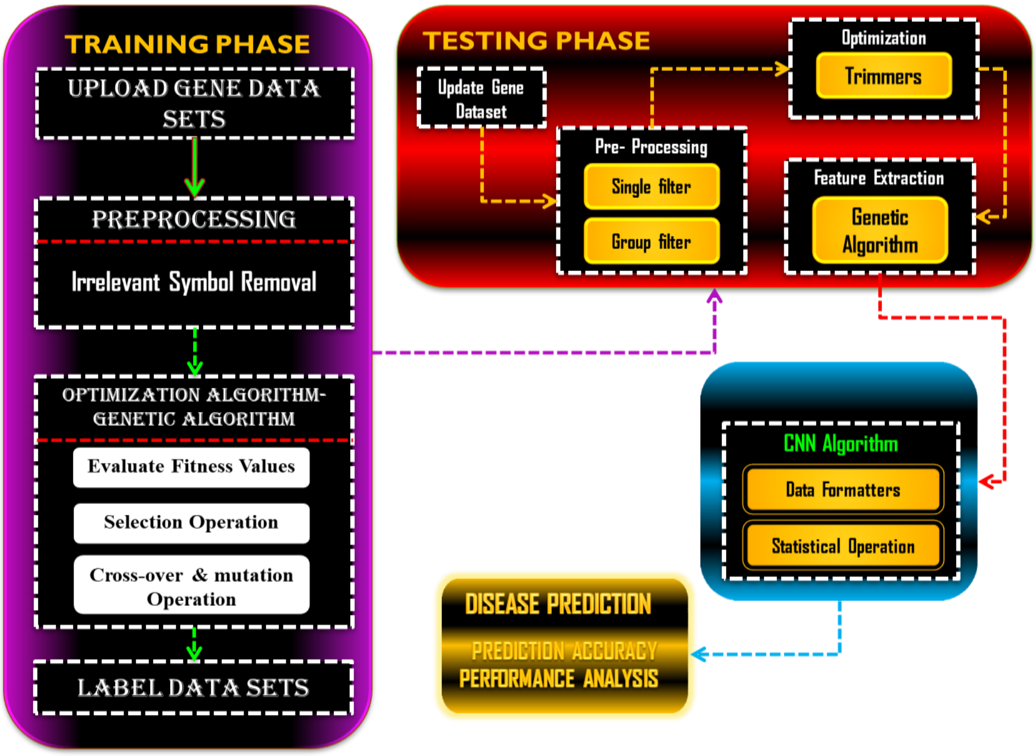


Fig. 1 System architecture

The objective of our proposed task will be to utilize supervised learning algorithm to classify and to predict the sicknesses, based on gene expression gathered from microarrays. Because of the quick advancement of DNA Micro cluster innovation, classifying strategies and quality determination procedures are being processed for better utilization of algorithm in microarray gene expression data. Microarrays are fit for deciding the articulation levels of thousands of gene simultaneously.

Benefits

- Dimensionality can be decreased
- Handle huge number of gene pattern
- Similar genes are gathered
- Predict the infection with exactness accuracy.

A. *Datasets Acquisition*

The vital employments of a microarray data set are to store the estimation of measurable data, deal with an accessible record; what's more, make the dataset is accessible to different applications for examination and understanding. A microarray database is a vault containing microarray quality articulation information. The vital employments of a microarray data set are to store the estimation information, deal with an accessible file, and make the information accessible to different applications for examination and interpretation. In this module, transfer the datasets. The dataset may be microarray dataset. A microarray data set is a store containing microarray gene expression data. At that point actualize pre-handling steps to take out the insignificant images.

B. *Preprocessing*

The median is the value separating to isolate the higher half from the lower half of an data test (a population or a probability distribution). For an data set index, it could be considered as the "middle" esteem. The activities given by SeQual can be isolated into the accompanying four primary functionalities: 1) Filter Channels: These activities dispose of those input reads that don't satisfy a specific models indicated by the client. Channels are partitioned into two classes, depending on the number of sequences involved groupings engaged with the channel rule

Single Filters: which assess peruses individually. SeQual incorporates 12 single channels. For example, successions can be shifted by their length, quality or the nonappearance/presence of a certain example in their bases.

Group Filters: which think about peruses by sets and dispose of those that are equivalent. SeQual contains 5 gathering channels that permit, for example, to analyze the arrangements as supplement or opposite supplement. The user can also specify a certain number of allowed mismatches to discard those sequences that are almost equal.

C. *Optimization*

In Genetic calculation, can examine coverage of the data prior before to grouping starts. What's more, propose an calculation, which changes the closest sorting arranging and the exchange transfer calculation, of the spatial medians grouping. It has two particular stages: one of moving an article starting with one clustering group then onto another the next and the other of amalgamating the single part bunch with it's the closest group.

Trimmers: SeQual includes 10 tasks for request to manage the start or finishing of the groupings by eliminating those bases that are not fascinating for the client. The client can determine the number of bases that should remain, or the quality needed for the managed arrangement sequence.

D. *Classification:*

Classifiers dependent on gene expression on quality are large probabilistic, that only predict that a certain percentage level of the people that have a given articulation profile will likewise have the aggregate, or then again result, of interest. Hence, factual approval is essential before models can be utilized, particularly in clinical settings. In this module execute convolutional neural network algorithm to characterize the different kinds of sicknesses from gene expression

Data Formatters: Three functions to convert from DNA to RNA read (and vice versa) or from FASTQ to FASTA formats are also provided by our tool.

Statistical Operations: Finally, SeQual provides three additional functions to obtain some statistics about the initial and/or final data. For instance, these operations can be used to count the number of input sequences, or to calculate their average length/quality.

With the assistance of CNN classifier we can undoubtedly isolate out the vessels into corridors and veins. The CNNs exhibit different appealing highlights like great speculation capacity contrasted with different classifiers. To be sure, there are generally any free boundaries to change and few free parameters to adjust and it is not required to find the architecture experimentally. The CNNs calculation isolates the classes of information designs with the maximal edge hyperplane. This hyper plane is developed as:

$$f(x) = \langle w, x \rangle + b \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots.\ldots\ldots\ldots. \text{ (1)}$$

Where x is the element vector, w is the vector that is opposite to the hyper plane and $b\|w\| -1$ determines the balance from the start of the arrange framework. To profit by non-straight choice limits the division is acted in an element space F, which is presented by a nonlinear planning φ the info designs. This planning is characterized as follows:

$$\langle \varphi\ (x1\ ), \varphi(x2) \rangle = K(x1, x2\ )\ \forall (x1, x2) \in X \ldots\ldots (2)$$

E. *Disease Prediction*

Utilizing multi class grouping calculation algorithm to arrange the seriousness level of illnesses utilizing grouped classification data check. If count is more than check limit implies, give seriousness as high and tally is less than edge implies, consider as typical. At that point give remedy to patients as indicated by the illnesses. And furthermore assess the presentation of the framework regarding exactness rate. With respect to the use of the instrument, SeQual gives two execution modes

Through the command-line interface by specifying:

- The way to the dataset(s) as input arguments;
- The tasks to be performed on these datasets utilizing a Java Properties document.
- Through a graphical interface given by SeQual to improve on its utilization to non-computer science specialists
- This graphical interface has been executed upon the open-source JavaFX project [40], which permits worked in partition between the application rationale and the visual piece of SeQual.

IV. GENETIC ALGORITHM PROCESSING STAGE

The nature of each assignment is being handled by the client prerequisites except if the client is being fulfilled. It doesn't deal with that task as entire rather it assesses every boundary of that task on premise of wellness esteem. A basic hereditary calculation incorporates three fundamental hereditary activities: choice, hybrid, and change.

Five phases are considered in a genetic algorithm.

- Initial population
- Fitness function
- Selection
- Crossover
- Mutation.

A. *Initial Population*

An individual is described by a set of parameter boundaries (factors) known as Genes. The cycle starts with a bunch of people which is known as a Population. Genes are joined into a string to shape a Chromosome (arrangement). Every individual is an answer for the difficult you need to solve. In a genetic algorithm, the arrangement of genes of an individual is addressed utilizing a string, as far as a alphabet set.

B. *Fitness Function*

The fitness function work decides to gives a wellness score to every person. How fit an individual is (the capacity of a person to contend with others). The probability that an individual will be chosen for reproduction depends on its wellness score.

C. *Selection:*

Two sets of individual people (guardians) are chosen dependent on their fitness scores. The possibility of choice stage of selection phase is to choose the fittest people and allowed them to pass their qualities to the future. People with high fitness have more opportunity to be chosen for generation.

D. *Crossover*

For each pair of guardians to be mated, a hybrid point of crossover is picked at inside the genes; Crossover is the main stage in significant phase in a genetic algorithm.

E. *Mutation*

This implies that a portion of the bits in the string can be flipped. In certain new offspring framed, a portion of their genes can be exposed to a change with a low irregular likelihood, random probability, within the population and prevent premature convergence Mutation occurs to maintain diversity.

V. PROPOSED ALGORITHM

A. *K-Nearest Neighbour Algorithm*

It named as CNNs (Convolutional Neural Networks) address feed-forward neural network which comprise of different combination of the convolutional layers, max pooling layers, and completely related layers. This organization differs standing to the spatial channel size and the quantity of output classes of input gene data.

IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Vol. 10, Issue 4, April 2021

DOI 10.17148/IJARCCE.2021.10475

B. *Performance Parameters*

The parameters are taken as execution boundaries to validate the proposed framework. Where, TN (number of correct predictions forecasts that an occasion is negative), FP(number of inaccurate expectations that a case is positive), FN(number of mistaken of forecasts that an occurrence negative), TP(number of correct predictions forecasts that an instance is positive).

- Accuracy = TP + TN / TP+TN+FP+FN
- Sensitivity = TP / TP + FN.
- Specificity = TN / TN + FP

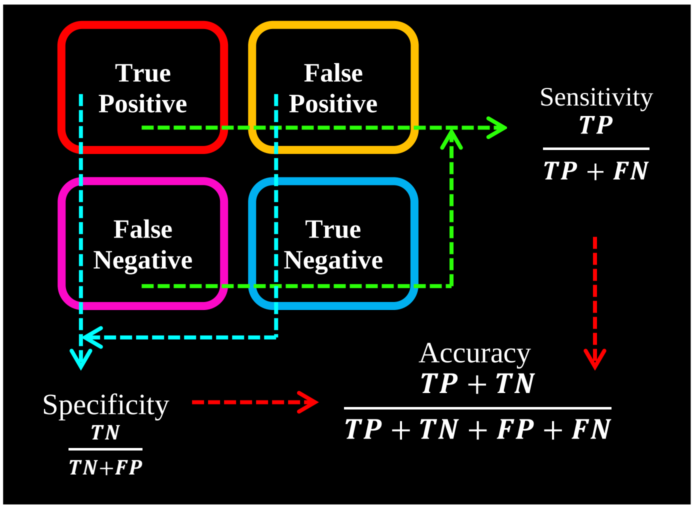

True Positive
False Positive
False Negative
True Negative
Sensitivity $\frac{TP}{TP+FN}$
Accuracy $\frac{TP+TN}{TP+TN+FP+FN}$
Specificity $\frac{TN}{TN+FP}$

Fig. 2 Performance parameter

VI. IMPLEMENTATION

At most elevated level of reflection, the general work process of SeQual is isolated into three primary stages:

- Reading of the input dataset(s) indicated by the client, to specify by a couple of user consisting FASTQ/FASTA text-based sequence documents when working in single-or matched end mode, separately.
- Processing of the input file records as per the quality-control tasks selected by the client in the graphical interface determined in Properties of document when utilizing the order to command line interface.
- Writing of the prepared dataset(s) to their relating output text file

TESTING PHASE MODEL	TRAINING PHASE MODEL

Constructing the CNN Model
function INITCNNMODEL (θ, [n1–5]) layer Type = [convolution, max-pooling, fully-connected, fully-connected]; layer Activation = [tanh(), max(), tanh(), softmax()]
model = new Model(); for i=1 to 4 do
layer = new Layer(); layer.Type = layer Type[i]; layer.inputSize = ni
layer. Neurons = new Neuron [ni+1];
layer.params = θi; model.addLayer(layer);
end for return model; end function
	Initialize learning rate α, number of max iteration ITERmax, min error ERRmin, training batch BATCHES training, batch size batch,
	Compute $n2$,$n3$,$n4$,$k1$,$k2$, according to $n1$ and $n5$;
	for batch = 1 to BATCHEStraining do [$\nabla\theta J(\theta)$, $J(\theta)$] = CNN Model. Train
	(Training Data's, Training Labels), as (4) and (8); Update θ using (7);
	err = err + mean($J(\theta)$);
	end for err = err/BATCHEStraining;
	iter++; end while
	Save parameters θ of the CNN

Table 1 CNN phase model

C. *Results and Discussion*

From the analysis, it is concluded that the CNN model is more accurate to precise when contrasted with different models. To improve the precision of this model, and compute the sensitivity and specificity to shows the outcomes insights for Table 2.

MODEL NO.	MODEL NAME	ACCURACY	SENSITIVITY	SPECIFICITY
1	CNN	94.5	82.56	87.27

Table 2 Single model results

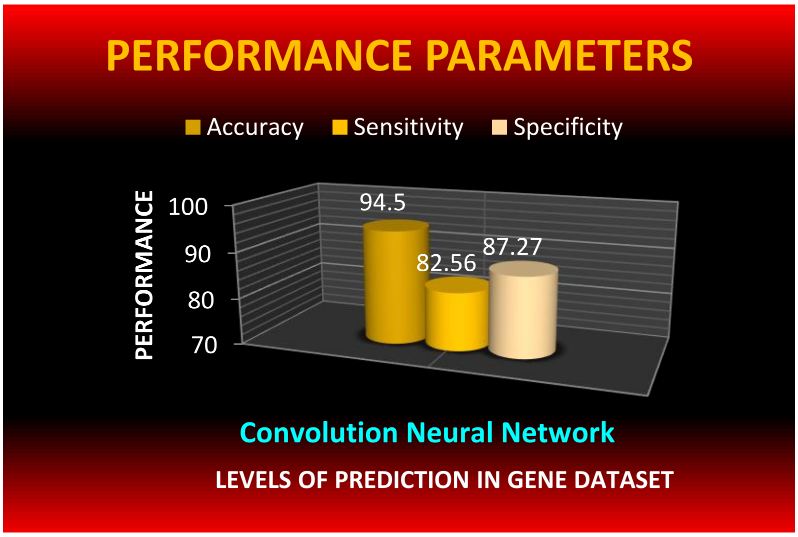


Fig. 3 Performance parameters of single model

VII. Experimental Results

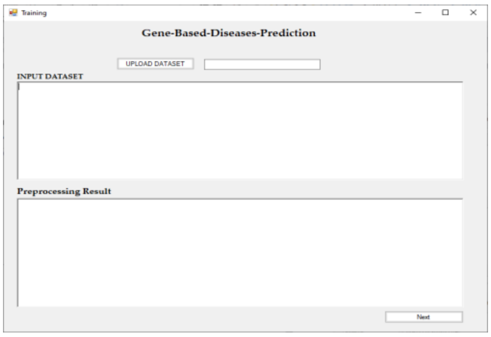
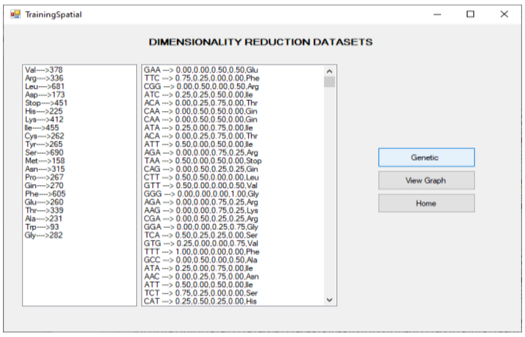


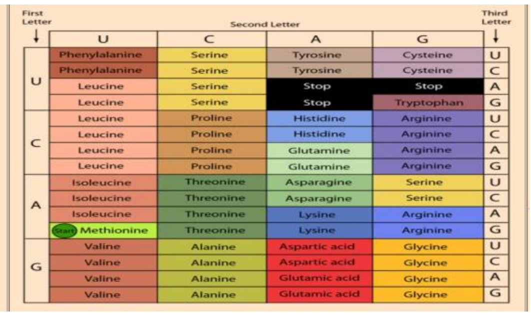
Fig. 4 Upload the data



```
Val---->378
Arg---->336
Leu---->681
Asp---->173
Stop---->451
His---->225
Lys---->412
Ile---->455
Cys---->262
Tyr---->265
Ser---->690
Met---->158
Asn---->315
Pro---->267
Gin---->270
Phe---->605
Glu---->260
Thr---->339
Ala---->231
Trp---->93
Gly---->282
```

```
GAA ---> 0.00,0.00,0.50,0.50,Glu
TTC ---> 0.75,0.25,0.00,0.00,Phe
CGG ---> 0.00,0.50,0.00,0.50,Arg
ATC ---> 0.25,0.25,0.50,0.00,Ile
ACA ---> 0.00,0.25,0.75,0.00,Thr
CAA ---> 0.00,0.50,0.50,0.00,Gin
CAA ---> 0.00,0.50,0.50,0.00,Gin
ATA ---> 0.25,0.00,0.75,0.00,Ile
ACA ---> 0.00,0.25,0.75,0.00,Thr
ATT ---> 0.50,0.00,0.50,0.00,Ile
AGA ---> 0.00,0.00,0.75,0.25,Arg
TAA ---> 0.50,0.00,0.50,0.00,Stop
CAG ---> 0.00,0.50,0.25,0.25,Gin
CTT ---> 0.50,0.50,0.00,0.00,Leu
GTT ---> 0.50,0.00,0.00,0.50,Val
GGG ---> 0.00,0.00,0.00,1.00,Gly
AGA ---> 0.00,0.00,0.75,0.25,Arg
AAG ---> 0.00,0.00,0.75,0.25,Lys
CGA ---> 0.00,0.50,0.25,0.25,Arg
GGA ---> 0.00,0.00,0.25,0.75,Gly
TCA ---> 0.50,0.25,0.25,0.00,Ser
GTG ---> 0.25,0.00,0.00,0.75,Val
TTT ---> 1.00,0.00,0.00,0.00,Phe
GCC ---> 0.00,0.50,0.00,0.50,Ala
ATA ---> 0.25,0.00,0.75,0.00,Ile
AAC ---> 0.00,0.25,0.75,0.00,Asn
ATT ---> 0.50,0.00,0.50,0.00,Ile
TCT ---> 0.75,0.25,0.00,0.00,Ser
CAT ---> 0.25,0.50,0.25,0.00,His
```

Fig. 5 Dimensionality reduction



First Letter ↓	Second Letter				Third Letter ↓
	U	C	A	G	
U	Phenylalanine	Serine	Tyrosine	Cysteine	U
	Phenylalanine	Serine	Tyrosine	Cysteine	C
	Leucine	Serine	Stop	Stop	A
	Leucine	Serine	Stop	Tryptophan	G
C	Leucine	Proline	Histidine	Arginine	U
	Leucine	Proline	Histidine	Arginine	C
	Leucine	Proline	Glutamine	Arginine	A
	Leucine	Proline	Glutamine	Arginine	G
A	Isoleucine	Threonine	Asparagine	Serine	U
	Isoleucine	Threonine	Asparagine	Serine	C
	Isoleucine	Threonine	Lysine	Arginine	A
	Start Methionine	Threonine	Lysine	Arginine	G
G	Valine	Alanine	Aspartic acid	Glycine	U
	Valine	Alanine	Aspartic acid	Glycine	C
	Valine	Alanine	Glutamic acid	Glycine	A
	Valine	Alanine	Glutamic acid	Glycine	G

Fig. 6 Genetic code

DOI 10.17148/IJARCCE.2021.10475

VIII. CONCLUSION

In this paper, the enormous measure of information created by present day NGS innovations supports the requirement for adaptable instruments with the capacity to perform equal calculations Microarray is a significant apparatus for disease arrangement at the sub-atomic level. In this paper we have we have proposed a mixture quality determination technique, which consolidates a Genetic strategies and CNN characterization to accomplish high order execution. SeQual, a Big Data device that completely abuses the highlights of Apache Spark to diminish the runtime required for the quality control and preprocessing of DNA/RNA groupings. The outcomes on different infection datasets shows the significance of a similar classifier utilized for both the quality determination and arrangement can improve the strength of the model. The undertaking zeroed in on promising exactness results with not very many number of quality subsets empowering the specialists to anticipate the different infections.

REFERENCES

[1] Bermejo, Pablo, Jose A. Gámez, and Jose M. Puerta. "A GRASP algorithm for fast hybrid (filter-wrapper) feature subset selection in high-dimensional datasets." Pattern Recognition Letters 32.5 (2011): 701-711.
[2] Lee, Cheng-Ming, and Chia-Nan Ko. "Time series prediction using RBF neural networks with a nonlinear time-varying evolution PSO algorithm." Neurocomputing 73.1-3 (2009): 449- 460.
[3] Yu, Jianbo, Shijin Wang, and Lifeng Xi. "Evolving artificial neural networks using an improved PSO and DPSO." Neurocomputing 71.4-6 (2008): 1054-1060.
[4] Bansal, Jagdish Chand, and Kusum Deep. "A modified binary particle swarm optimization for knapsack problems." Applied Mathematics and Computation 218.22 (2012): 11042-11061.
[5] Ramos, Juan, et al. "An agent-based clustering approach for gene selection in gene expression microarray." Interdisciplinary Sciences: Computational Life Sciences 9.1 (2017): 1- 13.
[6] Booma, P. M., and S. Prabhakaran. "Classification of genes for disease identification using data mining techniques." Journal of Theoretical and Applied Information Technology 83.3 (2016): 399.
[7] Genetic Algorithm Gene Feature Selection For Microarray Classification." International Journal of Applied Engineering Research 11.4 (2016): 2761-2770.
[8] Bennet, Jaison,. "A hybrid approach for gene selection and classification using support vector machine." Int. Arab J. Inf. Technol. 12.6A (2015): 695-700.
[9] Nagpal, Rashmi, and Rashmi Shrivas. "Cancer Classification Using Elitism PSO Based Lezy IBK on Gene Expression Data." Journal of Scientific and Technical Advancements 1.4 (2015): 19-23.
[10] Thangaraju, Mr P., and R. Mehala. "Novel Classification based approaches over Cancer Diseases." system 4.3 (2015).
[11] K. A. Phillips, ''Assessing the value of next-generation sequencing technologies: An introduction,'' Value Health, vol. 21, no. 9, pp. 1031–1032, Sep. 2018.
[12] W. R. McCombie, J. D. McPherson, and E. R. Mardis, ''Next-generation sequencing technologies,'' Cold Spring Harbor Perspect. Med., vol. 9, no. 11, p. a036798, 2019.
[13] Alkan, S. Sajjadian, and E. E. Eichler, ''Limitations of next-generation genome sequence assembly,'' Nature Methods,vol. 8, no. 1, pp. 61–65, Jan. 2011.
[14] R. Bao, ''Review of current methods, applications, and data management for the bioinformatics analysis of whole Exome sequencing,'' Cancer Informat., vol. 13, no. 2, pp. 67–82, 2014.
[15] S. Pabinger, ''A survey of tools for variant analysis of next-generation genome sequencing data,'' Briefings Bioinf., vol. 15, no. 2, pp. 256–278, 2013.
[16] J. Luo, M. Wu, D. Gopukumar, and Y. Zhao, ''Big data application in biomedical research and health care: A literature review,'' Biomed Inf. Insights, vol. 8, pp. 1–10, Jan. 2016,
[17] Smowton, A. Balla, D. Antoniades, C. Miller, G. Pallis, M. D. Dikaiakos, and W. Xing, ''A cost-effective approach to improving performance of big genomic data analyses in clouds,'' Future Gener. Comput. Syst., vol. 67, pp. 368–381, Feb. 2018.
[18] M. Zaharia, ''Apache spark: A unified engine for big data processing,'' Commun. ACM, vol. 59, no. 11, pp. 56–65, 2019.
[19] R. Schmieder and R. Edwards, ''Quality control and preprocessing of metagenomic datasets,'' Bioinformatics, vol. 27, no. 6, pp. 863–864, Mar. 2019.

DOI 10.17148/IJARCCE.2021.10475

BIOGRAPHY

Name : Mr. M. Ananda Kumar **M.E., MBA.,**

Designation : Assistant Professor,
Department of Computer Science and Engineering,
Arasu Engineering College,
Kumbakonam

Specialization : Big Data, Cloud Computing, Cryptography, Data Mining, Distributed Computing, Machine Learning and Neural Networks.

Name : A. Manju **(M.E – CSE)**

College : Arasu Engineering College, Kumbakonam

Specialization : Big Data, Data Mining, Machine Learning.