Periodicals of Engineering and Natural Sciences ISSN 2303-4521
Vol. 9, No. 2, March 2021, pp.462-470

Evaluation patterns and algorithm for cancer identifications using dynamic clustering

Waleed Hadi Madhloom Kurdi[1], Hussein Ali Rassool[2], Aqeel Hamza Al-fatlawi[3]
[1] Department of Medical Laboratory Techniques, Altoosi University College, Najaf, Iraq
[2] Department of Nursing, Altoosi University College, Najaf, Iraq
[3] Department of Computer Techniques Engineering, Imam Kadhum University College, Iraq

ABSTRACT

The domain of knowledge discovery and deep data extraction is quite prominent and used in assorted domains including engineering, mathematics and even in medical diagnosis. A number of benchmark datasets are available in which huge research work is going on with the enormous aspects of genomics that is associated with the medical data analytics. In this research manuscript, the work presents the evaluation patterns and the approaches which are used for the cancer identification with the use of dynamic clustering and deep data analytics. The work is having the elements with the medical datasets and their key features by which the training of data in the data mining algorithm can be integrated and then the overall predictions can be done on assorted parameters.

Keywords: Bioinformatics, Dynamic Clustering, Data Mining, Medical Data Analytics

Corresponding Author:

Waleed Hadi Madhloom Kurdi
Department of Medical Laboratory Techniques
Altoosi University College
Najaf, Iraq
E-mail: waleedalkurdi@altoosi.edu.iq

1. Introduction

Bio-informatics refers to the field of biological data understanding methods and software tools, especially where the data sets are large and complicated, is an interdisciplinary domain. Bioinformatics combines biology, computer science, information technology, mathematics and statistics to analyze and interpret biological data as an interdisciplinary field of science. Bioinformatics has been used with mathematical and statistical techniques to analyze biological research in silicon [1].

The biology field includes bioinformatics research using computer programming as part of its methodology, and a specific "pipeline" analysis used on a number of occasions in the field of genomics, in particular. Popular uses in bioinformatics include candidates and individual nucleotide polymorphisms recognition [2]. Such identification is often done to better understand the genetic basis of disease, unique adjustments, desirable properties (particularly on agricultural species) or population differences. The organizational principles within nucleic acid and protein sequences known as proteomics are also being sought in a less formal manner by bioinformatics [3].

In many fields of biology, bioinformatics has become a major component. Bioinformatics techniques like image and signal processing allow us to derive valuable results from vast quantities of raw data in experimental molecular biology. In the field of genetics, genomes and their mutations are sequenced and annotated [4]. It plays a role to organize and query biological data in the text mining of biological literature and in the development of biological and gene ontology. It also contributes to the study and control of gene and protein expression. Tools of bioinformatics help to compare, analyze and interpret genetic and genomic data and to understand the evolving aspects of molecular biology more generally. It helps to examine and index biological pathways and networks, a central aspect of system biology, at a more inclusive basis. It helps in the simulation and modeling of DNA, RNA and proteins, in structural biology. Bioinformatics and Medical Data Analytics is one of the key domains of research in the data mining, machine learning and knowledge discovery in which the medical datasets are analyzed for the prediction of diseases and diagnostic features [5]. The medical datasets

are classically available in assorted formats including BLAST, GENOME structures, FASTA and many others [6].

1. Bioinformatics and datasets

Datasets Taxonomy in the international formats for medical diagnosis can be grouped as follows:

- Clustalw
- PubMed
- FASTA
- Medline
- UniGene
- GenBank
- GenBank
- Blast

These datasets are available on benchmark portals of research whereby key feature points are extracted and then trained using specific algorithms. The work is presenting the case analytics in which the datasets of bioinformatics related to cancer are evaluated.

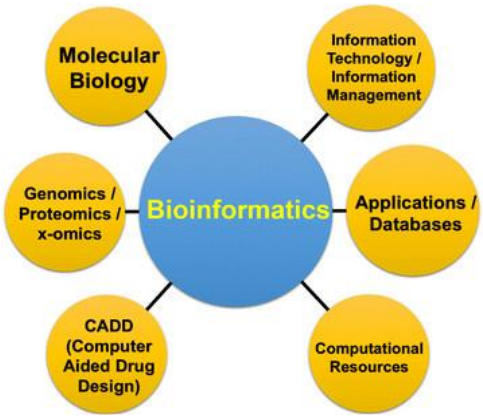

Figure 1. Key Elements of bioinformatics data analytics

Figure 1 presents the key elements of bioinformatics for the data analytics and other problem solving related to medical datasets. There are enormous elements including computational resources, CADD, genomics, protenomics, molecular biology drug designing and many others.

In spite of the massive assessment machines in clinical sciences, the software tools and libraries are in like way utilized. These software tools and applications assess the regular information which are gotten from the electronic confirmation machines. Here, the chance of Bioinformatics goes to the situation wherein the software tools and applications are utilized to welcome the common and clinical information [7]. These software suites utilize pervasive programming vernaculars at the back-end to process and assess the characteristic dataset with the goals to locate the human body parameters for inconceivable treatment [8].

Computer applications and the Internet are the key instruments of a bio-informatician. A key activity is the sequence analysis of DNA and proteins using different web-based programmes and databases. All those with access to the Internet or to relevant websites can now freely discover the composition of biological molecules such as nucleic acids and proteins using basic bio-informatics instruments, from clinicians to molecular biologists. This does not make it convenient for anyone to manage and interpret raw genomic data. Bioinformatics is an emerging science, with the use of sophisticated software to capture, sort, interpret, forecast and store data on DNA and protein sequences by specialists in the field of bioinformatics [9].
Bioinformatics growth has been a global undertaking, developing computer networks that make biological knowledge easy to access and allow software for smooth analysis to be created. The entire research community is publicly accessible over the Internet with many international programs aimed at supplying gene and protein databases.

Today, though it remains a center of genomics and genetics, bioinformatics has become an arena for a broader scope of biological science studies examining, structuring, systemizing, annotating, searching, mine-projecting and visualizing biological knowledge available and a number of biomedical text documents. Even if it is not possible to draw a fine line between bioinformatics and other similar areas since computers, statistics and mathematics are increasingly used to address scientific problems and to experiment with life sciences, there should be no confusion of bioinformatics definition and aims. Biometric and biostatistics, creation of DNA machines or computerized production and recording of imaging data, for example, should not be mixed with biometry or biostatistics [10].

For example, bioinformatics techniques like genomic and genetic analysis and/or signal processing allow users to view and understand molecular and evolutionary processes and interactions in the field of wet-bench experimental molecular biology from vast quantities of crude data [11]. In application for system biology, bioinformatics is a crucial method in modeling and cataloguing biochemical / genetic roads and networks to combine pieces of studied data in order to represent and model a detailed image of life processes [12]. The use of bio-informatics tools for reconstruction, pattern recognition, folding, simulation and molecular modeling will detect structural peculiarities and molecular sequences interactions that are essential to structural biology and medicinal drugs design [13]. It is difficult to analyze all these large-scale, genome-based, molecular "big data" sequence analyzes manually. In combination with modern computing knowledge, this led the biology science research community to apply interdisciplinary methods and tools for "Big Data" analysis, resulting in the formation of new interdisciplinary bioinformatics sciences. Let us look first at the historical developments in the field of bioinformatics [14].

Bioinformatics is a computer science field related to the sequence analysis of biological molecules. It normally refers to genes, DNA, RNA or protein and is especially useful for comparing genes and other protein and other sequences within or between organisms in order to determine their functions through patterns across DNA and protein sequences. Bioinformatics can primarily be thought of as the linguistic aspect in genetics. That is, people of linguistics look at language trends and bioinformatics look at patches inside sequences of DNA or protein. This is what people look at.

2. Benchmark medical datasets

They include the following viruses and diseases:

a. Cancer
- ICCR Cancer Dataset [15]
- World DataSet [16]
- Kaggle [17]
- UCI Machine Learning Datasets [18]
- BIOGPS [19]
- DataHub [20]

b. Wuhan CoronaVirus
- NCBI Portal [21]

c. Viruses
- NCBI Portal [22]

d. HIV
- IANL Dataset [23]
- RCSB Dataset [24]

e. Bacteria Viral Data
- NCBI Dataset [25]

f. Dengue
- NCBI Portal [26]

The knowledge discovery from assorted datasets is quite important and thereby the need arise to integrate the suitable algorithms so that the effectual usage patterns can be implemented [26].

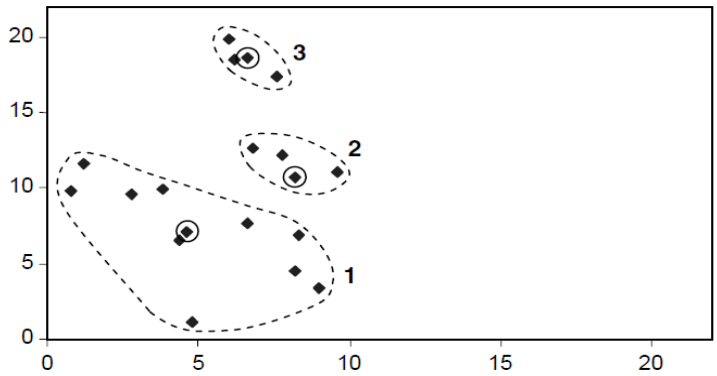

Figure 2. Data clustering

Figure 2 and Figure 3 presents the illustration of data clustering and outlier respectively in which the similar data elements are grouped.

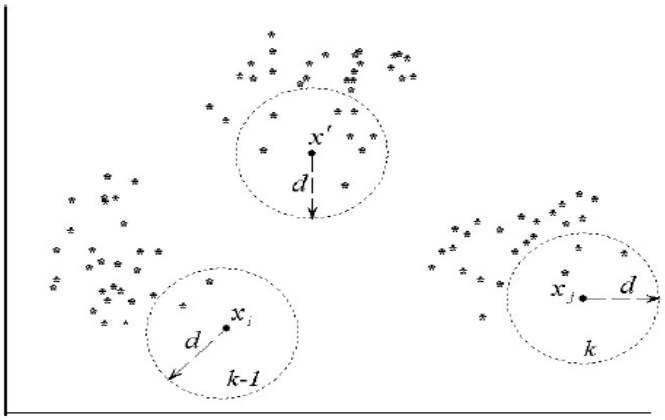

Figure 3. Illustration of outlier in data

3. Dynamic clustering for cancer identification

The dynamic clustering for cancer identification is done using the advanced algorithm that makes use of the soft computing-based implementation using which the interior components can be effectively clustered using fitness function [27].

The fitness function is required so that the appropriate elements and the medical points can be put in the specific cluster and others are identified as the outliers. The outliers here are associated with the non-cancerous points and thereby the predictions can be done.

4. Mathematical formulation

It can be summarized as follows:
DataSet Integration: $\sum (DS_i.\{RS(TS_i)\}\text{->}FSF) \Rightarrow CS_R$ ϑ $RS(PSCT\ (RS):PSTG\ (RS)) \Rightarrow CF_m m <= i$
DS_i= Dataset repository
$RS(T_i)$=Record or Data Set for the Analytics
FSF= Effective Fitness Function
CS_R=Eligibility of Cluster for Inclusion
CSF_m=Integrated Eligibility for the Cluster
PSTG = Percentage Level
PSCT = Effective Percentile Level towards the Cluster
OUTPUT: $(\text{Ɛ}\ CS_i) \infty (TS_i - TS_{i-1})$

CS_i = Medical Cluster set
OL_i=Outlier that is non-cancerous

The presented formulation is having the numerous elements and factors and out of these the fitness function and formulation is the key point that forms the base of the clustering.

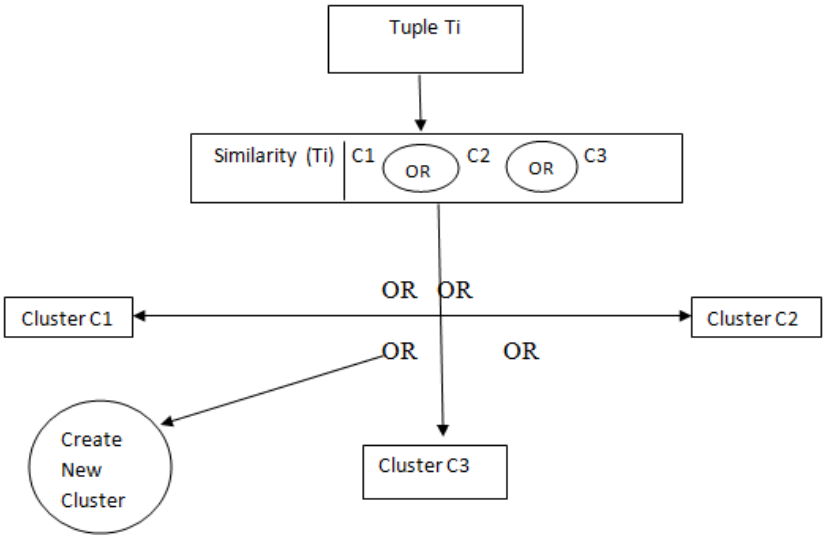

Figure 4. Cluster association pattern

Figure 4 illustrates the cluster association pattern using similarity-based analysis and inclusion in the specific cluster.

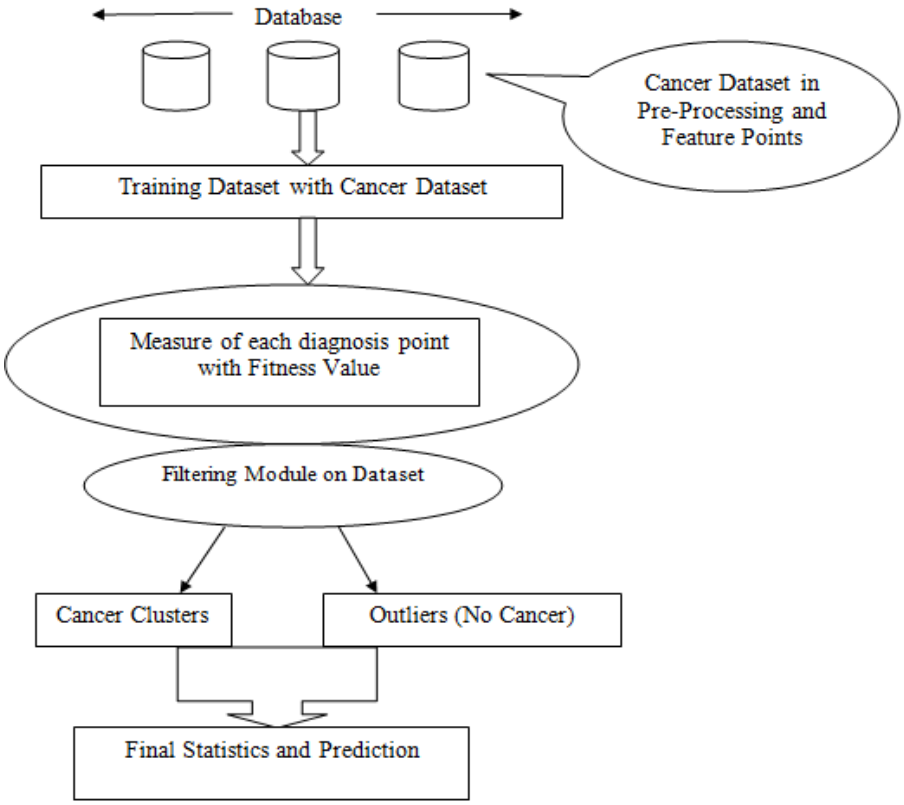

Figure 5. Projected clustering approach

The projected clustering approach in Figure 5 presents the cluster formation for the cancer disease by which the cancerous elements are grouped.

5. Results and implementation outcomes

Following are the results and outcomes from the evaluation of dataset with the analytics and prediction-based approach.

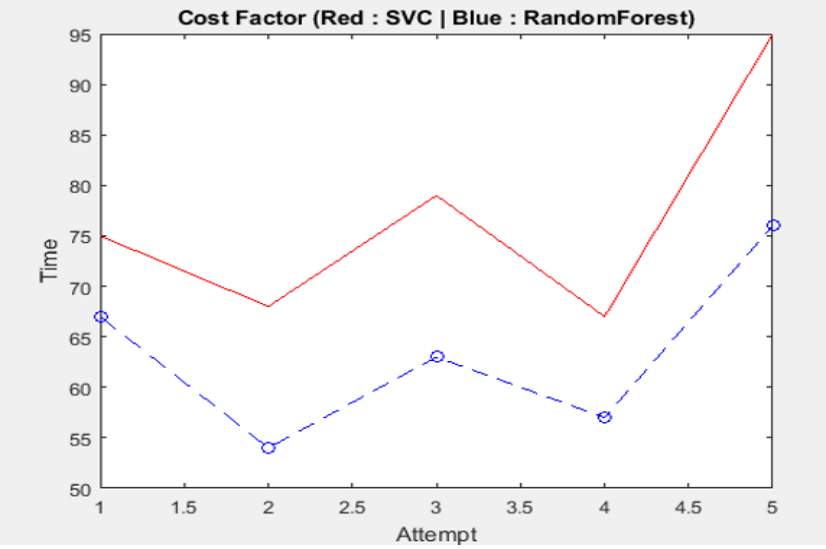

Figure 6. Evaluation of cost factor

The results and outcomes from the Support Vector Machine and Ensemble Learning based approach and the analytics patterns are depicted in Table 1.

Table 1. Evaluation of performance

Execution Scenario	Ensemble Learning	Support Vector Machine
1	75.5	66
2	72.5	56.5
3	78	64.5
4	67.6	56
5	78.5	65.5

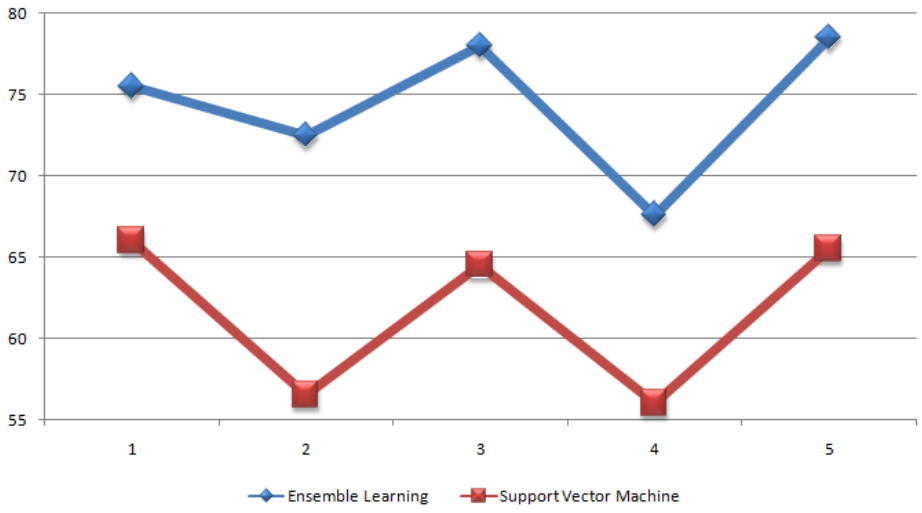

Figure 7. Analytics of performance

Table 2. Accuracy levels in the training and testing data

Features	Epoch	Accuracy	MSE
10	10	98.7	3.47196e-3
20	20	95.8	2.84596e-3
30	23	96.8	2.57884e-3
40	26	98.8	1.53845e-3
50	29	98.7	1.02626e-3
250	**56**	**99.9**	**6.37337e-4**

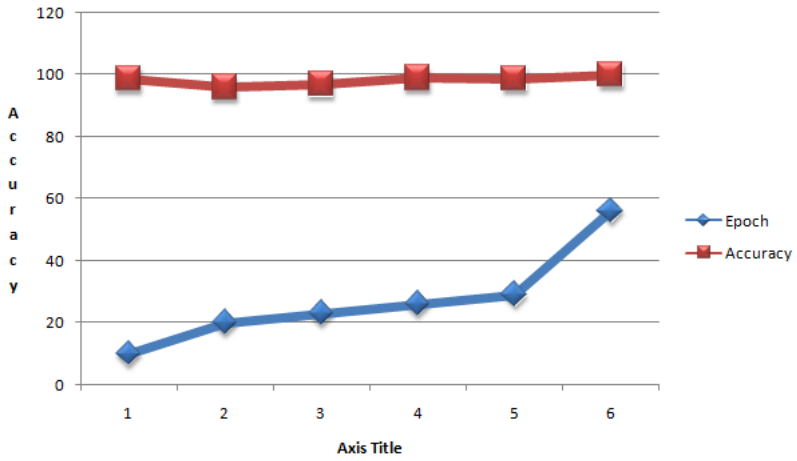

Figure 8. Outcomes on accuracy on the implementation aspects

The FASTA File for Analytics is based on the following:

>sp|P25730|FMS1ECOLI CS1 fimbrial precursor subunit A (CS1 pilin)
MKLKKTIGAMALATLFATMGASAVEKTISVTASVDPTVDLLQSDGSALPASVALTYSPAV
AAFEAKTIATVVKTADSDKGVVVKLSADPVLSAVLAPTLQIPVSVAFAGKPLSTTGITID
SADLAFASSGVAKVSSTQKLSIKADATRVTGGALTAGQYQGLVSIILTKSTTTTTTTKGT
>sp|P15488|FMS3ECOLI CS3 fimbrial subunit A precursor (CS3 pilin)
MLKIKYLLIGLSLSAMSSYSLAAAGPTLTKELALAVLSPAALDATWAPQDALTLSATGVS
ATLVGVLTLSATSIDTVSIASTAVSDTSKAGTVTFAKETAASASFATTISTDAAAITLDK
AAGATIVKTTAGSQLPTALPLKFITTEGAEKLVSGAYRAAITITSTIKGGGTKKGTTDKK
Id: sp|P25730|FMS1ECOLI
sp|P25730|FMS1ECOLI
sp|P25730|FMS1ECOLI CS1 fimbrial subunit A precursor (CS1 pilin)
Annotations: {}
Sequenced Data:

MKLKKTIGAMALATLFATMGASAVEKTISVTASVDPTVDLLQSDGSALPASVALTYSPAVAAFEAKT
IATVVKTADSD
KGVVVKLSADPVLSAVLAPTLQIPVSVAFAGKPLSTTGITIDSADLAFASSGVAKVSSTQKLSIKADAT
RVTGGALTA
GQYQGLVSIILTKSTTTTTTTKGT
Sequenced Alphabet: SingleLetterAlphabets()
Id: sp|P15488|FMS3ECOLI
sp|P15488|FMS3ECOLI
sp|P15488|FMS3ECOLI CS3 fimbrial subunit A precursor (CS3 pilin)
Annotations: {}
Sequenced Data:
MLKIKYLLIGLSLSAMSSYSLAAAGPTLTKELALAVLSPAALDATWAPQDALTLSATGVSATLVGVL
TLSATSIDTVS

IASTAVSDTSKAGTVTFAKETAASASFATTISTDAAAITLDKAAGATIVKTTAGSQLPTALPLKFITTEG
AEKLVSGA
YRAAITITSTIKGGGTKKGTTDKK
Sequenced Alphabet: SingleLetterAlphabets()

Additional prospective work can be employed to develop cancer identifications using cloud computing [28-29] or Internet of Thing (IoT) [30], or by using genetic algorithm for results optimization [31].

6. Conclusion

The domain of knowledge disclosure and profound information extraction is very conspicuous and utilized in grouped domains including building, arithmetic and even in clinical determination. Various benchmark datasets are accessible in which immense research work is going on with the tremendous parts of genomics that is related with the clinical information examination. Right now, the work presents the assessment designs and the methodologies which are utilized for the disease recognizable proof with the utilization of dynamic grouping and profound information investigation. The work is having the components with the clinical datasets and their key highlights by which the preparation of information in the information mining algorithm can be incorporated and afterward the general forecasts should be possible on arranged parameters.

References

[1] M. Ovesný, P. Kˇrížek, J. Borkovec, Z. Švindrych, and G. M. Hagen, "ThunderSTORM: a comprehensive ImageJ plug-in for PALM and STORM data analysis and super-resolution imaging," *Bioinformatics*, vol. 30, no. 16, pp. 2389–2390, 2014. [Online]. Available: 10.1093/ bioinformatics/btu202;https://dx.doi.org/10.1093/bioinformatics/btu202

[2] W. Torres-García, S. Zheng, A. Sivachenko, R. Vegesna, Q. Wang, R. Yao, M. F. Berger, J. N. Weinstein, G. Getz, and R. G. Verhaak, "PRADA: pipeline for RNA sequencing data analysis," *Bioinformatics*, vol. 30, no. 15, pp. 2224–2226, 2014. [Online]. Available: 10.1093/ bioinformatics/btu169;https://dx.doi.org/10.1093/bioinformatics/btu169

[3] P. Jiang, K. Sun, F. M. F. Lun, A. M. Guo, H. Wang, K. C. A. Chan, R. W. K. Chiu, Y. M. D. Lo, and H. Sun, "Methy-Pipe: An Integrated Bioinformatics Pipeline for Whole Genome Bisulfite Sequencing Data Analysis," *PLoS ONE*, vol. 9, no. 6, pp. e100 360–e100 360, 2014. [Online]. Available: 10.1371/journal.pone.0100360;https://dx.doi. org/10.1371/journal.pone.0100360

[4] R. J. Carroll, L. Bastarache, and J. C. Denny, "R PheWAS: data analysis and plotting tools for phenome-wide association studies in the R environment," *Bioinformatics*, vol. 30, no. 16, pp. 2375– 2376, 2014.

[5] Q. Zou, J. Zeng, L. Cao, and R. Ji, "A novel features ranking metric with application to scalable visual and bioinformatics data classification," *Neurocomputing*, vol. 173, pp. 346– 354, 2016.

[6] M. U. Ali, S. Ahmad, and J. Ferzund, "Harnessing the potential of machine learning for bioinformatics using big data tools," *International Journal of Computer Science and Information Security*, vol. 14, no. 10, pp. 668–668, 2016.

[7] L. Mak, D. Marcus, A. Howlett, G. Yarova, G. Duchateau, W. Klaffke, A. Bender, and R. C. Glen, "Metrabase: a cheminformatics and bioinformatics database for small molecule transporter data analysis and (Q)SAR modeling," *Journal of Cheminformatics*, vol. 7, no. 1, pp. 31–31, 2015.

[8] A. D. Baxevanis, G. Bader, and D. Wishart, *Bioinformatics*, 2020.

[9] A. R. Wattam, D. Abraham, O. Dalay, T. L. Disz, T. Driscoll, J. L. Gabbard, J. J. Gillespie, R. Gough, D. Hix, R. Kenyon, D. Machi, C. Mao, E. K. Nordberg, R. Olson, R. Overbeek, G. D. Pusch, M. Shukla, J. Schulman, R. L. Stevens, D. E. Sullivan, V. Lonstein, A. Warren, R. Will, M. J. Wilson, H. S. Yoo, C. Zhang, Y. Zhang, and B. W. Sobral, "PATRIC, the bacterial bioinformatics database and analysis resource," *Nucleic Acids Research*, vol. 42, no. D1, pp. D581–D591, 2014.

[10] R. J. Urbanowicz, R. S. Olson, P. Schmitt, M. Meeker, and J. H. Moore, "Benchmarking relief-based feature selection methods for bioinformatics data mining," *Journal of Biomedical Informatics*, vol. 85, pp. 168–188, 2018.

[11] R. S. Olson, W. L. Cava, Z. Mustahsan, A. Varik, and J. H. Moore, 2017.

[12] I. Merelli, H. Pérez-Sánchez, S. Gesing, and D. Agostino, 2014.

[13] T. R. Connor, N. J. Loman, S. Thompson, A. Smith, J. Southgate, R. Poplawski, M. J. Bull, E. Richardson, M. Ismail, S. Elwood- Thompson, C. Kitchen, M. Guest, M. Bakke, S. K. Sheppard, and M. J. Pallen, "CLIMB (the Cloud Infrastructure for Microbial Bioinformatics): an online resource for the medical microbiology community", Microbial *Genomics*, vol. 2, no. 9, 2016. [Online]. Available: 10.1099/mgen.0.000086;https://dx.doi.org/10.1099/mgen.0.000086

[14] R. Couronné, P. Probst, and A.-L. Bluestein, "Random forest versus logistic regression: a large-scale benchmark experiment," BMC *Bioinformatics*, vol. 19, no. 1, pp. 270– 270, 2018.
[15] U. I. Dataset. [Online]. Available: http://www.iccr-cancer.org/datasets
[16] *World Bioinformatics Dataset URL*.
[17] K. Dataset and Url. [Online]. Available: https://www.kaggle.com/uciml/ breast-cancer-wisconsin-data
[18] U. Dataset. [Online]. Available: https://archive.ics.uci.edu/ml/datasets/ Lung+Cancer
[19] Dataset URL. [Online]. Available: http://biogps.org/dataset/tag/cancer/
[20] Dataset URL. [Online]. Available: https://datahub.io/machine-learning/breast- cancer
[21] U. N. P. Dataset. [Online]. Available: https://www.ncbi.nlm.nih.gov/ genbank/2019-ncov-seqs/
[22] NCBI Portal Dataset. [Online]. Available: https://www.ncbi.nlm.nih.gov/genome/ viruses/variation
[23] U. Dataset. [Online]. Available: https://www.hiv.lanl.gov/content/sequence/HIV/USER_ALIGNMENTS/keele.html
[24] U. R. Dataset. [Online]. Available: https://www.rcsb.org/pages/help/ advancedsearch/sequence
[25] U. N. P. Dataset and U. Dataset. [Online]. Available: https: //www.ncbi.nlm.nih.gov/genome/?term=Klebsiella%20pneumoniae
[26] U. N. P. Dataset. [Online]. Available: https://www.ncbi.nlm.nih.gov/ nuccore/LC379197.1?report=fasta
[27] A. A. Ferdaus and M. A. Khan, "A Genetic Algorithm Approach using Improved Fitness Function for Classification Rule Mining," *International Journal of Computer Applications*, no. 23, pp. 97–97, 2014.
[28] Y. S. Mezaal, H. H. Madhi, T. Abd, S. K. Khaleel, "Cloud computing investigation for cloud computer networks using cloudanalyst," Journal of Theoretical and Applied Information Technology, vol. 96, no. 20, 2018.
[29] T. Abd, Y. S. Mezaal, M. S. Shareef, S. K. Khaleel, H. H. Madhi, & S. F. Abdulkareem," Iraqi e-government and cloud computing development based on unified citizen identification", *Periodicals of Engineering and Natural Sciences*, vol.7, no.4, pp.1776-1793, 2019.
[30] Y. S. Mezaal, L. N. Yousif, Z. J. Abdulkareem, H. A. Hussein, S. K. Khaleel, "Review about effects of IOT and Nano-technology techniques in the development of IONT in wireless systems," International Journal of Engineering and Technology (UAE), vol. 7, no. 4, 2018.
[31] Y. S. Mezaal, S. F. Kareem, "Affine Cipher Cryptanalysis Using Genetic Algorithms," JP Journal of Algebra, Number Theory and Applications, vol. 39, no. 5, pp. 785-802, 2017.