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

Engineering, mathematics, and even medical diagnostics all use deep data extraction for knowledge discovery and extraction. Many benchmark datasets exist in which a large amount of research is taking place in relation to genomics and medical data analytics. Data analytics and dynamic clustering are utilized to identify cancer in this research publication, which outlines the evaluation patterns and methodologies employed for this purpose. Working with medical datasets and their important properties, a data mining procedure may be trained, and thus a predictions variety can be made on various parameters.

Keywords: Bioinformatics, Data Mining, Dynamic Clustering, Therapeutic 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

When dealing with vast and complex datasets, bioinformatics is an interdisciplinary field that requires a wide range of expertise. Interdisciplinary bioinformatics is the study of biological data through the use of computer science, information technology (IT), mathematics, and statistical methods. Analysis of biological studies in silicon using bioinformatics and mathematical and statistical tools has been done [1].

An analysis known as "pipeline" has been employed on numerous occasions in the genomics field, as well as in other areas of biology that use computer programming. Candidates and individual nucleotide polymorphisms are two of the most commonly used bioinformatics applications [2]. To better comprehend a genetic basis of disease, distinctive adaptations, anticipated features (specifically on agricultural species), or population differences, this identification is commonly done. Bioinformatics is also looking for the same organizational principles in nucleic acid and protein sequences known as proteomics [3].

Bioinformatics has been an imperative part of numerous branches of biology. Experiments in molecular biology can benefit greatly from the use of bioinformatics tools such as image and signal processing. Genes are sequenced and annotated as part of the study of genetics. It is used in a text mining of biological literature and the construction of biological and gene ontologies to start up and query biotic data. There are many other uses for this technology, including the study of gene and protein expression and control. Using bioinformatics tools, one may compare, analyze and understand the developing features of molecular biology and genetics and genomic data. To better comprehend and index the central features of system biology, such as pathways and networks, this helps. DNA, RNA, and proteins can be simulated and modeled with this software. There are many areas of research in bioinformatics, medical data analytics, and machine learning that focus on the analysis of medical datasets for predicting diseases and diagnostic characteristics [5]. BLAST, GENOME structures, FASTA, and many other formats are commonly used for medical datasets [6].

Using benchmark portals of study, these datasets can be used to extract and train certain feature points. The case analyses presented in this study are based on bioinformatics datasets linked to cancer.

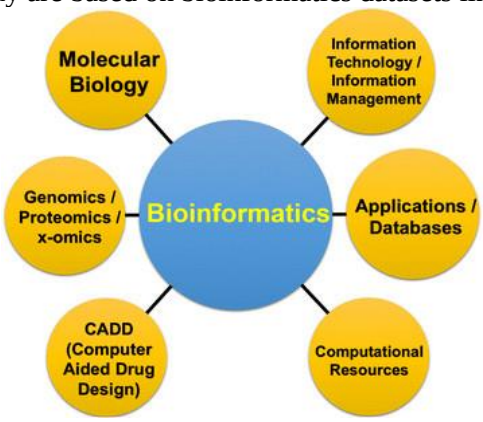

Figure 1. Significant Components of bioinformatics data analytics

The usage of bioinformatics is required for data analysis and other problem solving activities involving medical information, as seen in Figure 1. CADD, genomics, protenomics, molecular biology pharmaceutical design, and a host of additional characteristics are among those available.

Despite the vast amount of assessment equipment used in clinical sciences, software solutions and resources are used equally as much. For the purpose of analyzing the routine data that is collected from electronic verification equipment, software and apps are utilized. [7] This is an area where Bioinformatics has the ability to make a difference, since programs are employed for receiving data from both ordinary and medical sources. When it comes to the backend of these software sets, prevalent programming vernaculars are used to process and evaluate the unique data set in order to pinpoint human body factors for unimaginable therapy [8], pervasive programming vernaculars are used to handle and evaluate the differentiated set of data in order to find humanoid body parameters for unimaginable therapy.

Computers and the Internet are the key tools used by bioinformaticians in their work. The examination of DNA and protein sequences using a variety of web-based programs and databases is a significant activity in biotechnology. From doctors to molecular scientists, anyone with access to the Internet may now utilize simple bioinformatics tools to determine the structure of organic components like nucleic acids and proteins, regardless of their training or experience. Management and analysis of raw genomic data are not made any easier as a result of this development, though. With the use of advanced software, specialists in the field of bioinformatics are able to record data on DNA and protein sequences, sort them, evaluate them, forecast them, and store them [9].

It has taken a global effort to advance bioinformatics research, beginning with the establishment of computerized systems that make biological information available and tolerate for the development of software that makes smooth analysis possible. Over the efforts of various international projects, the entire research community can now be reached through the Internet.

Bioinformatics is used to investigate, structure, systemize, annotate, search, mine, and illustrate a broad range of biomedicine text documents as well as publicly available biological information. It is a hub for more than just genetics and genome analysis anymore; it is also a hub for more than just genetics and genome analysis. The concept and purpose of bioinformatics should not be in doubt, despite the fact that computing, statistics, and math and science are increasingly being used to address scientific problems and conduct research in the field of life sciences. It is not a good idea to use biometrics and biostatistics in conjunction [10], and it is even worse to use DNA tools or computerized imaging data creation and recording.

To better visualize and comprehend evolutionary and molecular processes and interactions, wet-bench experimental molecular biology makes use of bioinformatics technologies like genetic and genomic analysis and/or signal processing to examine enormous volumes of raw data. System biology relies on bioinformatics, which is used to develop precise models of biochemical and molecular processes and networks [12]. Bioinformatics is one of the most essential approaches in system biology.

In structural biology and drug development, bioinformatics technologies including as pattern recognition, folding, simulation, and molecular modeling can help uncover structural idiosyncrasies and molecular sequence relationships [13]. Big data sequence analyses are challenging to analyze manually because of their size and complexity. New bioinformatics sciences were born out of this mix of modern computing expertise and the use

of multiple tools and methods for "Big Data" investigation. To begin, let's take a look at the evolution of bioinformatics over time [14].
Biological sequence analysis is the focus of bioinformatics, which stands for a division of computer science. Genome sequencing is typically used to identify the roles of genes, RNA, or protein sequences, but it can also be used to compare genes and other protein and other arrangements within or among organisms. Genetics' linguistic component is best represented by bioinformatics. As a result, people in linguistics and bioinformatics look at language trends and DNA or protein sequences, respectively. This is what most individuals see when they pass by.

2. Benchmark medical datasets

They include the following viruses and diseases:

a. Malignant cells
- ICCR Cancer data set [15]
- World data set [16]
- Kaggle [17]
- UCI Machine Learning data set [18]
- BIOGPS [19]
- DataHub [20]

b. Wuhan CoronaVirus
- NCBI Gateway [21]

c. Viruses
- NCBI Gateway [22]

d. HIV
- IANL data set [23]
- RCSB data set [24]

e. Bacteria Viral Records
- NCBI data set [25]

f. Dengue
- NCBI Gateway [26]

It is critical to integrate the appropriate algorithms for effective use patterns so that knowledge discovery can take place from a wide range of data sources [26, 27].

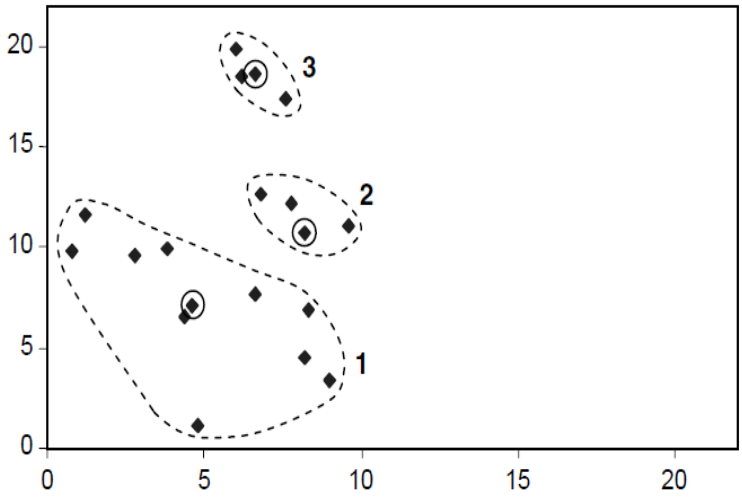

Figure 2. Data clustering [27]

Similar data components are classified into clusters as seen in Figure 2 while outliers are depicted in Figure 3.

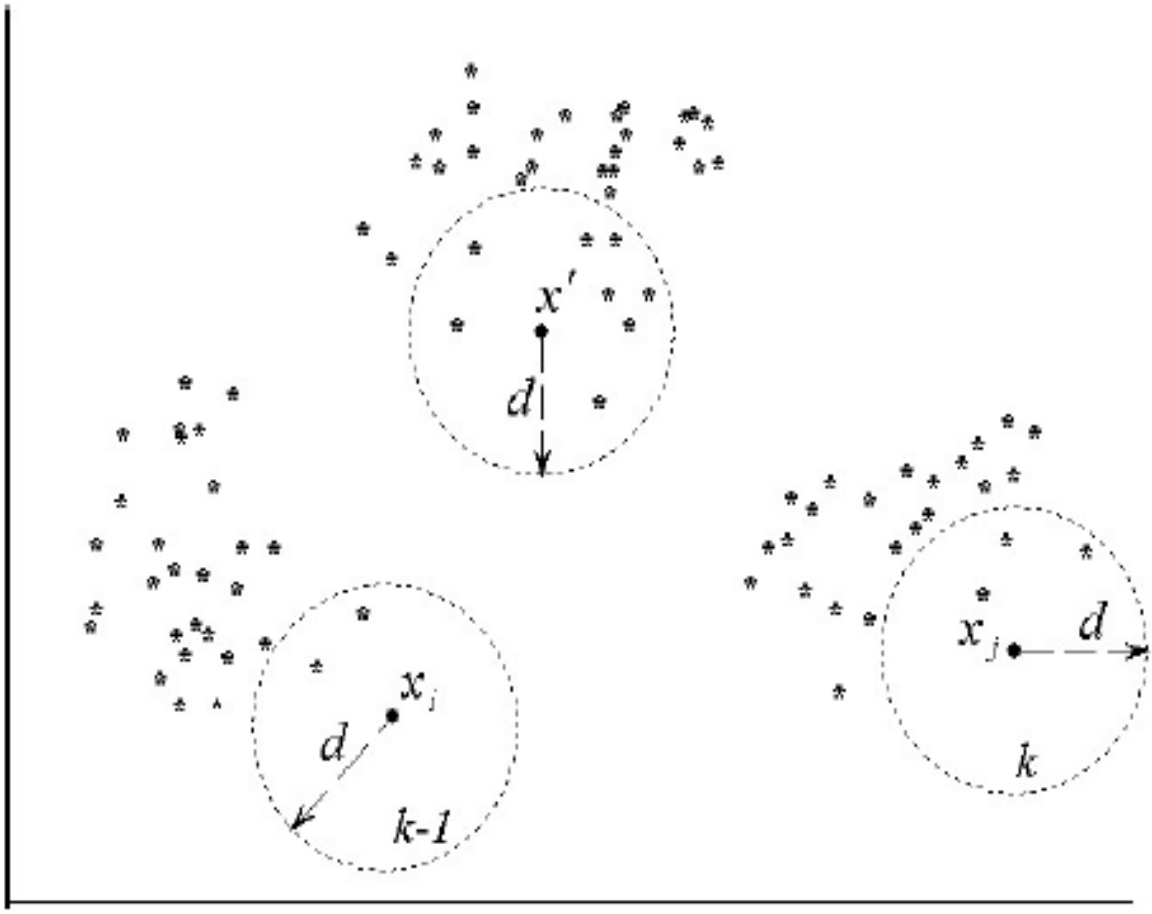

Figure 3. Outlier depiction in data [27]

3. Cancer identification based on dynamic clustering

In the dynamic clustering for cancer recognition, the innovative method uses a soft computing-based application, that efficiently clusters the internal constituents using a fitness function [27].
Using a fitness function, the correct components and medical points are assigned to a correct cluster, while outliers can be found using the fitness function. The non-cancerous locations are linked to the outliers in this dataset, making predictions possible.

4. Methodologies

Important mathematical equations can be summarized as follows [27]:
Data Set Incorporation: $\sum (DS_i.\{RS(TS_i)\}\text{->}FSF) \Rightarrow CS_R$ ϑ RS(PSCT (RS):PSTG (RS)) => CF_mm <=i
DS_i= Data set source
$RS(T_i)$=Recorded Data set for the Analytics
FSF= In effect Fitness Function
CS_R= Cluster Suitability for Enclosure
CSF_m=Integrated Suitability of a Cluster
PSTG = Percentage Level
PSCT = In effect Percentile Level headed for a Cluster
OUTPUT: $(\mathcal{E}\, CS_i) \infty (TS_i - TS_{i-1})$
CS_i = Therapeutic Cluster set
OL_i= Non-Cancerous Outlier

Clustering is based on a number of parameters, including a fitness function and origination stated in the presented formulation.

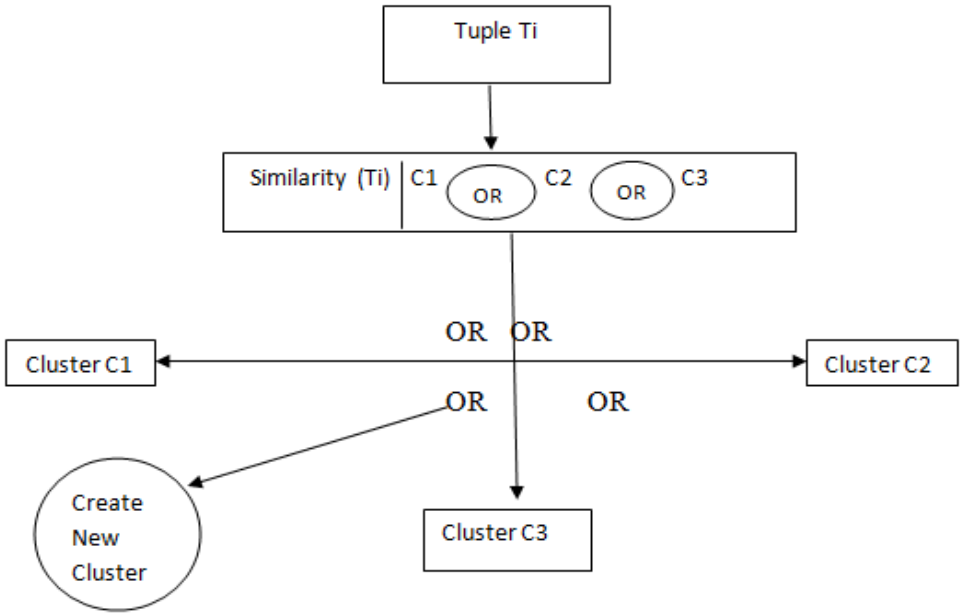

Figure 4. Cluster association pattern [27]

According to the similarity-based investigation in a certain cluster, a cluster association pattern is depicted in Figure 4.

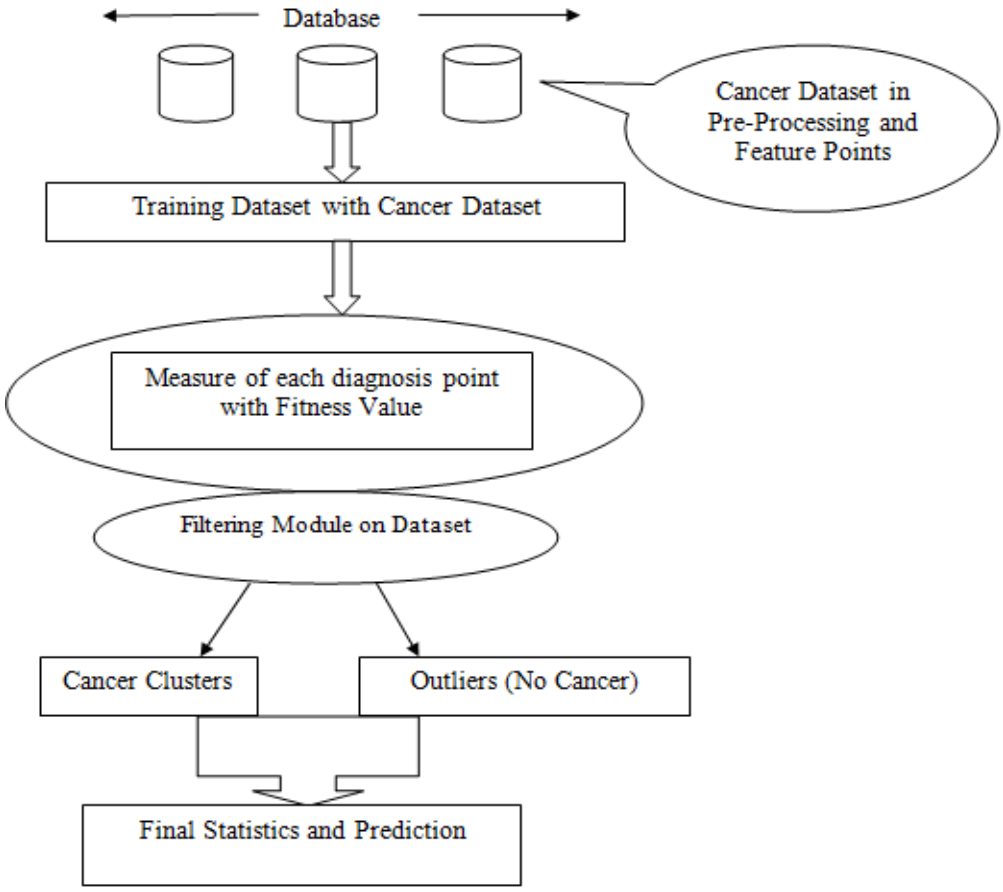

Figure 5. Clustering approach sample [27]

Figure 5 depicts a cancer clustering method in which the disease's malignant parts are grouped.

Thus, bioinformatics will be extremely relevant in an identifying and validation of biomarkers explicit to symptomatology associated with initial diagnosis, measurements to screen the progression of illness and reaction to therapy, as well as predictor variables for the enhancement of patient life quality, among other things, in the future. Biomarkers in cancer have been researched extensively, ranging from a single marker to a large number of indicators, from expression to function signal, and from network to dynamic network. Protein-protein interactions have been investigated as a potential new type of biomarker by including information regarding protein annotations, interactions, and signal transduction pathways into the research process. Using dynamic network biomarkers, which monitor and evaluate changes in network biomarkers at various stages and intervals during the development of diseases, is one of the novel approaches being used. It was projected that clinical informatics, such as diagnoses, history, treatments, clinical syndromes, physician checkups, biological analyses, imaging summaries, pathologies, and other measurements, will be linked with dynamic network biomarkers in the future. Systems medical medicine is a new-fangled approach to the development of cancer biomarkers that is gaining popularity. Integrating systems biology with clinical phenotypes and high-throughput technologies such as bioinformatics and computational science has the potential to develop medical diagnoses, treatment, and prognosis. Cancer biomarkers must comprise properties such as network dynamics, interactions, and specificities that are useful for disease identification, treatment, and prediction. Grasp the interface between clinical informatics and bioinformatics is the first and most critical step in identifying and creating new diagnostics and treatments for diseases. An approach like this has been documented in a variety of illnesses, including acute rejection following kidney transplantation and lung disease, among others. To put it another way, human samples from clinical research were obtained and harvested with a thorough profile of clinical informatics that is translated from the descriptions of the patients who are participating in the research. It is possible to employ bioinformatics and systems biology to identify dynamic networks and interaction among genes and/or proteins in specified samples using these techniques. Patients' disease-particular systems and dynamic networking of genes and/or proteins can be identified using computational methods, and disease-specific biomarkers are evaluated and optimized using computational methods. Finding the most effective system for converting clinical characterizations into clinical informatics, implementing bioinformatic methods to analyze disease severity and location as well as sensitivity to therapies, and integrating each component from clinical and high-throughput data for exact inferences are just a few of the challenges systems clinical medicine faces. Identifying how distinct molecular networks interact with one another, and therefore how gene and/or protein interconnections affects gene and/or protein expression, is even more complicated. Several studies published in Thematic Series on Cancer Bioinformatics in BMC Bioinformatics have found that combining protein network and interacted data improves an ability to deduce gene signatures under investigation to stratify breast cancer patients. These studies found that R weighted Recursive Feature Elimination and average sequence expression were the highly efficient in creating interpretable signatures among the tested methods [28].

Systems cancer medicine has been advocated as a means of advancing the implementation of forecasting, defensive, adapted, and participating medicine (P4 medicine). Theorists lately projected that a virtual cloud of billions of data produced by extraordinary throughput technology solutions in patients could be figured out, with one or more illness-perturbed networks in cells of an appropriate organ in a disease, and that this cloud would be used to figure out the disease. It is possible that disease-perturbed molecular networks will show irregularities in information provided and functioning, paving the way for the ultimate application of P4 medicine in cancer therapy. A major step toward achieving systems clinical medicine, clinical bioinformatics is a critical component of this process. It involves joining clinical measurements and signs with humanoid malignant cells tissue-generated bioinformatics to better comprehend disease progression, treatment response, and mapper connections that incorporate discrete elements that jointly initiate universal function within a particular genomic category. The optimizing methods for clustering and classifying developed by Ren and colleagues from the Thematic Series on Cancer Bioinformatics and published in BMC Bioinformatics includes data types from proteomics and next-generation sequencing. Using this method, it is feasible to ensure efficient

and effective predict the kind of cancer in breast cancer and leukemia benchmark datasets, among other things [28].

Oncology bioinformatics has a number of important applications, including evaluating how effective and safe precision medicine is based on the specific genetic and protein changes found in each patient. Because of a semantic heterogeneity of produced data by proteomics, epigenetics microarrays, and subsequent generation sequencing, an ontology-based solution for enquiring scattered databases via service-oriented, model-driven infrastructures was developed as a result of this research. After conducting further research with mouse models and RNAi, it was observed that a frequent genetic mutation in human liver cancer (resulting in instigation of FGF19), resulted in discriminating sensitivity to FGF19 inhibition, which was discovered via succeeding investigations with mice models and RNAi. In order to build precise tools for treating each patient's tumor based on the molecular network characteristics of their tumor, it is intended to analyze the molecular network characteristics of every patient's cancer. Through the development of new medicines, cancer analytics and systems biology are predictable to increase cancer avoidance, identification, and treatment. As it comes to analyzing the genomes, physiological sequences, large-scale "omics" benchmark datasets, and protein three-dimensional structure, computational cancer research is mainly reliant on statistical and bioinformatics approaches [28].

Additional prospective work can be employed to develop cancer identifications using cloud computing [29-30] or Internet of Thing (IoT) [31], or by using genetic algorithm for results optimization [32].

5. Conclusion

The field of datamining is highly visible and used in a variety of groups, including construction, arithmetic, and even medical determination. Numerous benchmark datasets are available in which enormous genomics research is taking place in relation to clinical data analysis. For the time being, the effort focuses on the evaluation methods and designs used to demonstrate the presence of a disease using dynamic grouping and in-depth information analysis. An important part of this project is to gather data from many sources, such as clinical datasets, so that information mining algorithms may include it, and then make general forecasts based on those factors.

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] S. K. Mandala, N. Gurrapu. "Advanced machine learning and data science based approach for prediction of cancer using dynamic clustering", Materials Today: Proceedings, 2021.

[28] D. Wu, C.M. Rice and X. Wang, "Cancer bioinformatics: A new approach to systems clinical medicine", *BMC Bioinformatics, Vol.* 13, no.71, 2012.

[29] 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.

[30] 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.

[31] 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.

[32] 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.