IJFANS INTERNATIONAL JOURNAL OF FOOD AND NUTRITIONAL SCIENCES
ISSN PRINT 2319 1775 Online 2320 7876
Research paper Volume 12, Iss 1, 2023

AN OVERVIEW TO BIOINFORMATICS AND ITS USES IN DIGITAL SIGNAL PROCESSING

[1]Dr BALASUBRAMANIAN B, [2]Dr SAROJ KUMAR SHAH, [3] T.SASIKALA
Department of Biomedical Engineering, Excel Engineering College
(AUTONOMOUS)Komarapalayam, Namakkal, Tamil Nadu.

Abstract

Recent developments in technology have significantly advanced our understanding of the genetic basis of phenotypes. In keeping with these developments, genomics has transformed the way that biological questions are approached on a whole genome scale, providing a wealth of data and creatinga plethora of new opportunities. Conversely, the massive volume of data generated highlights the obstacles that need to be addressed for the storage and processing of biological data (Moore's law). Computational biology and bioinformatics have worked to address these issues in this setting. A description of bioinformatics and its application to biological data analysis is provided in this article, which also looks at new techniques, tools, and approaches that can be used to give the data generated biological meaning..

Key words:
Databases; Data analysis; Systems biology, Genomics

Introduction

Recent technological developments have produced an abundance of "omic" data, leading to an unparalleled revolution in science. For experts from various fields, the increasing development of this information and its accessibility in public databases posed, and continues to pose, a difficulty. But what really is the problem? Making sense of the massive amount of structural information and sequences that have been created at various levels of biological systems is the fundamental problemin biology. However, in the field of bioinformatics, the creation of statistical and computational tools that can help comprehend the mechanisms underlying the study's biological issues is still required.

Moreover, this is an extremely reductionist perspective when taking into account the complexity of science. Other sciences that have an integrated molecular biology interface, such bioinformatics and computation in biology, are born or evolve alongside the "new biology" era. Bioinformatics and genomics have developed in tandem and had a historical influence on the body of knowledge, despite their recent consideration. As a result, the purpose of this review is to give a succinct synopsis of these fields and offer guiding principles for bioinformatics that cover the following areas: databases and forms of

biological data; ii) molecular modelling and sequence analysis; iii) genomic analysis; and iv) systems biology.

These are so vast topics, and our goal is to draw attention to important aspects of using novel approaches while also offering resources for data analysis and the interpretation of the outcomes produced by these technologies.

Bio what? Ahistoricaland conceptualvision of bioinformatics

A decade prior to the practicality of DNA sequencing, bioinformatics emerged. The discovery of the DNA structure by Watson and Crick in 1953 is one historical event that may be emphasised for its development, in addition to the body of information and understanding that biochemistry and protein structure have amassed via the research of Coren, Ramachandran, and Pauling in the 1960s.

Margaret O. Dayhoff is regarded as the mother of bioinformatics since she was a pioneer in organising the body of knowledge regarding the three-dimensional (3-D) structure of proteins (Hunt, 1984). This is because it played a part in the creation of peptide sequence-determining computers, programmes that identified and displayed structures for use in X-ray crystallography, and computational techniques for protein sequence comparison that let us deduce the evolutionary relationships between kingdoms (Hagen, 2000; Verli, 2014). Dr. Dayhoff, among other writers, produced a small book titled "Atlas of Protein Sequence and Structure," which is regarded as a turning point in the organisation and exchange of data.

In addition to them, a large number of additional researchers have made significant contributions to the field of bioinformatics development, which would not have been feasible without the advancement of computers. Therefore, the major advancements of today are primarily attributable to the genome projects and increases in processing power. An enormous amount of data could be obtained in the 1990s because to the introduction of large-scale capillary DNA sequencers and the fluorescence-marking of dideoxynucleotides. [1] However, the number of full genomes and the amount of data being generated are both increasing with the introduction of next-generation sequencing technology (NGS). [2] Consequently, the understanding of genetic variation and the evolutionary and functional mechanisms behind the genetic architecture requires the use of computers in study (Ritchie et al., 2015).

.

The field of bioinformatics and genomics can be summed up from three perspectives: "the application of computational tools to organise, analyse, understand, visualise, and store information associated with biological macromolecules," [3] which speaks to the multidisciplinary nature of the field. i) The molecular biology core concepts and the cell. [4] The organism, which exhibits variations between the many developmental phases and bodily regions, is the second focus of this attention. [5] The author concludes by highlighting

ISSN PRINT 2319 1775 Online 2320 7876
Research paper Volume 12, Iss 1, 2023

three global perspectives: iii) the tree of life, which divides millions of species into three evolutionary branches.

[6] The objectives of bioinformatics are outlined by these authors as follows: i) 2rganizati data so thatresearchers may access it and add new entries; ii) developing tools and resources to aid in data analysis; and iii) using these tools to analyse and interpret data in meaningful ways. [7] We can divide the problems associated with bioinformatics into two categories: sequence-related problems and biomolecular structure-related problems (Figure 1).

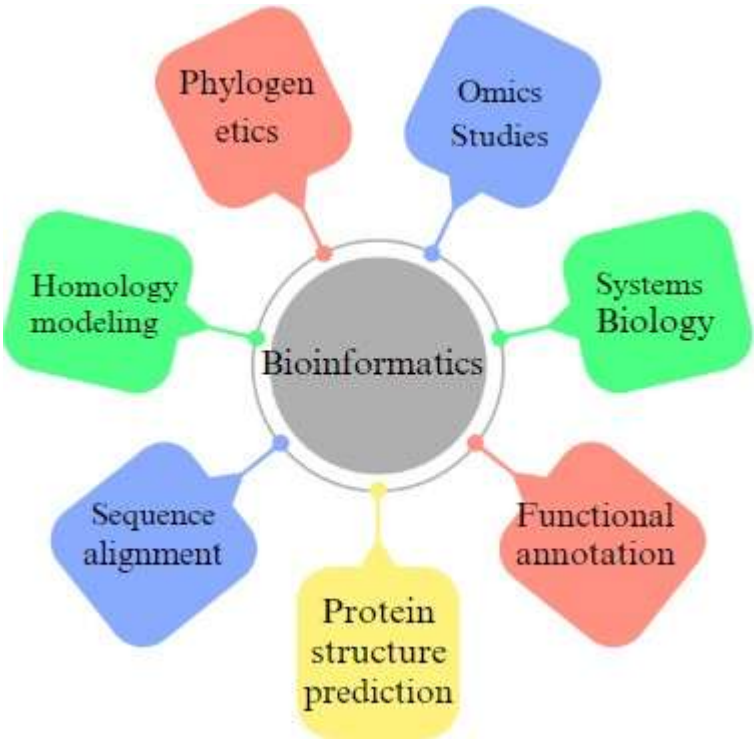

Figure 1. A selection of applications in bioinformatics.

Information organisation: information categories and databases

Owing to the massive amount of data generated, data 3rganization and storage are now essential. [8] As a result, databases were built, including a vast amount of biological data that has been processed and archived for use by the scientific community. The Nucleic Acids Research journal has been responsible for the compilation, modifying, and distribution of the growing number of biological databases that have been created in tandem with the growth in data. The most recent version, [9] which was released in January2017, lists 1739 biological databases. Bioinformatics uses a variety of information sources, including raw DNA sequences, sequences of proteins, macromolecular structures, and genome sequencing.

ISSN PRINT 2319 1775 Online 2320 7876
Research paper Volume 12, Iss 1, 2023

Large volumes of information are stored in public databases, which are divided into primary and secondary databases. The experimental data results that are released without a thorough study pertaining to earlier publications make up the main databases. [10] However, a process known as content curation involves compiling and interpreting data found in secondary databases. In addition to these, [11] there are useful databases that enable metabolic map evaluation and interpretation, such as Reactome and the Kyoto Encyclopaedia of Genes and Genomes (KEGG).

Analysis of biological sequences

The increased availability of data produced by NGS technology has facilitated alignment, which is widely utilised and crucial for biological sequence comparison [12]. In this method, two or more nucleotide sequences (DNA, RNA), or amino acid sequences (peptides, proteins), are compared in search of a set of distinct features or patterns that are likewise ordered in the sequences. Applications of this process include learning about the relationships between genes, persons, animals, and structures as well as prediction functions and structures. [13] Alignment techniques are also essential for whole genome analysis, which compares genomes from the same species or from different ones in order to find sequence differences and link them to certain traits.

Based on the number of sequences that are compared, alignment can be divided into two types: 1) simple alignment and 2) multiple alignment. By definition, a multiple alignment takes into account a value larger than three sequences, [14] whereas a simple alignment expressly shows the similarity relationship between two sequences. These can nevertheless be categorised as global in terms of alignment degree. [15] It is possible to categorise the employed algorithm as heuristic or optimum. The best alignment achievable is the optimum result, but the best alignment during the duration of the study is presented by the heuristic, even though it does not yield an optimal result.

An overview of the primary algorithms and alignment techniques is shown in Figure 2. Table 1 lists theprimary alignment programmes and their attributes in relation to these.

ISSN PRINT 2319 1775 Online 2320 7876
Research paper Volume 12, Iss 1, 2023

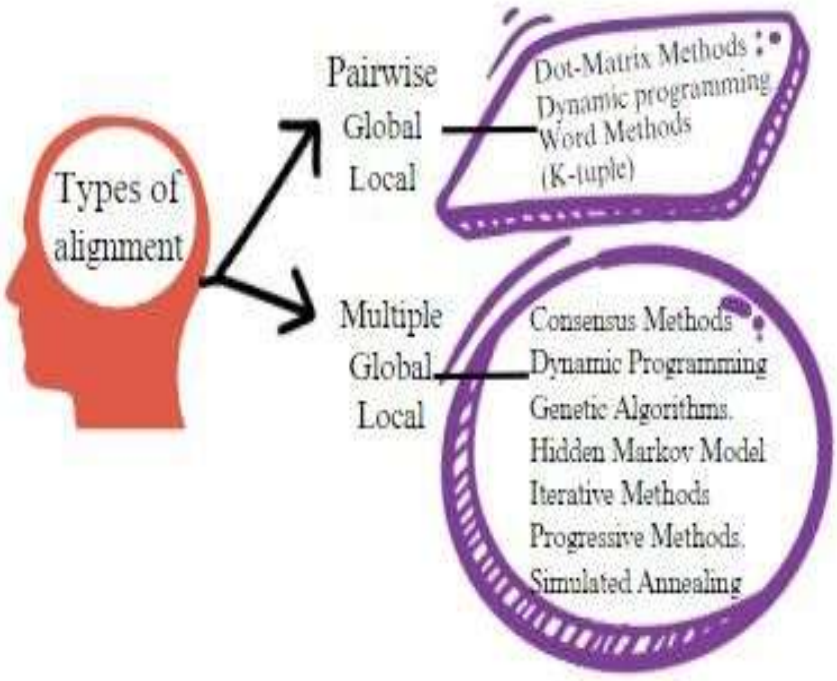

Figure 2. forms of alignment and implemented algorithms.

Table 1 Principal programmes of alignment and their attributes.

Sequence number	Program	Accuracy of alignment	Type of alignment
N	ClustalW	Heuristic	Global
N	Multalin	Heuristic	Global

ISSN PRINT 2319 1775 Online 2320 7876
 Volume 12, Iss 1, 2023

2	Needleman -Wunsch	Optimu m	Globa l
2	BLAST2 Sequences	Heuristic	Local
2	Smith-Waterma n	Optimu m	Local

Simple alignment

This method emphasises the k-tuple method, dot matrix analysis, and dynamic programming algorithms.
[16] The foundation of the dynamic programming approach is the Bellman's optimality principle, which states that a complicated problem's numerous subproblems can solve it. This approach can be used to generate local alignments using the Smith-Waterman algorithm and global alignments using the Needleman-Wunsch algorithm. [17] A score system for amino acids or nucleotides, matches and mismatches, and a penalty value for gaps are necessary for alignment. The algorithm will determine the best alignment between the sequences in this manner. Indels and repeats can be efficiently detected using the theoretically straightforward dot matrix approach . The regions of similarity can be visually represented through the use of an identity matrix. This method involves aligning one sequence vertically and the other horizontally, and then signalling sections that have the same characters to indicate potential matches. The areas of similarity will be shown as a line on the diagonal, with random correspondences represented by theother dots.

Multiple alignments

In global alignment, the dynamic programming method is typically used, much like in simple alignments. However, a weighted aggregate of pairs is added along with similarity values to punctuate each potential pair that may develop. In addition, other techniques were created to speed upthe computations; three in particular stand out: progressive, hidden Markov models, and iterative techniques.

ISSN PRINT 2319 1775 Online 2320 7876
Research paper Volume 12, Iss 1, 2023

BLAST

A particular local alignment approach called BLAST, which shows the greatest alignment score oftwo sequences, is derived from the Smith-Waterman algorithm. When searching the sequences in the database, BLAST uses a heuristic based on the k-tuple approach in addition to the dynamic programming that results from the algorithm previously discussed. The search is restricted to more significant terms using the k-tuple approach; these terms have a character count of three for amino acids and eleven for nucleotides.

Depending on the kind of sequence of interest and the collection of data to be searched, the BLAST family of programmes is utilised for various reasons (Prosdocimi, 2010). BLAST has a number of applications, some of which are included in Table 2. Megablast and PSI-BLAST (Position Specific Iterative BLAST) are also available, albeit they are less prevalent.

Table 2 An overview of the BLAST family of programmes

Program	Subject	Query
BLASTp	aa	aa
BLASTx	aa	nt*
BLASTn	nt	nt
tBLASTx	nt*	nt*
tBLASTn	nt*	aa

Two factors are used to present the BLAST results: the E-value and the score value (Score bits). Taking into account the alignment score and database size, the E-value is a statistical value that showsthe likelihood that the alignment did not happen at random. Conversely, the algorithm assigns the score based on the sequences that match and those that don't match the database.

Analyses across the genome - from genome to proteome

DNA sequencing is essential to the development of molecular biology since it not only alters the structure of genome designs but also creates new avenues for research and application. NGS technologies have various applications, as previously stated. In order to deal with this infinite, future bioinformatics techniques will centre on transcriptome, proteome, and genome analysis.

ISSN PRINT 2319 1775 Online 2320 7876
Research paper Volume 12, Iss 1, 2023

Genome

Because sequencing has become less expensive, many genomes have been published. The newapproaches do, however, have a constraint with the read quality and size (150–300 bp), which poses aproblem for assembly software. However, they generate a lot more sequences. Assembling the genome requires interpreting the millions of base pairs that have been sequenced. A hierarchical data structure that links the sequencing data to an assumed target reconstruction makes upthe assembly. Assembly is necessary, nevertheless, if a new genome is not previously characterised (de novo). The procedures for constructing the genome are depicted in Figure 3.

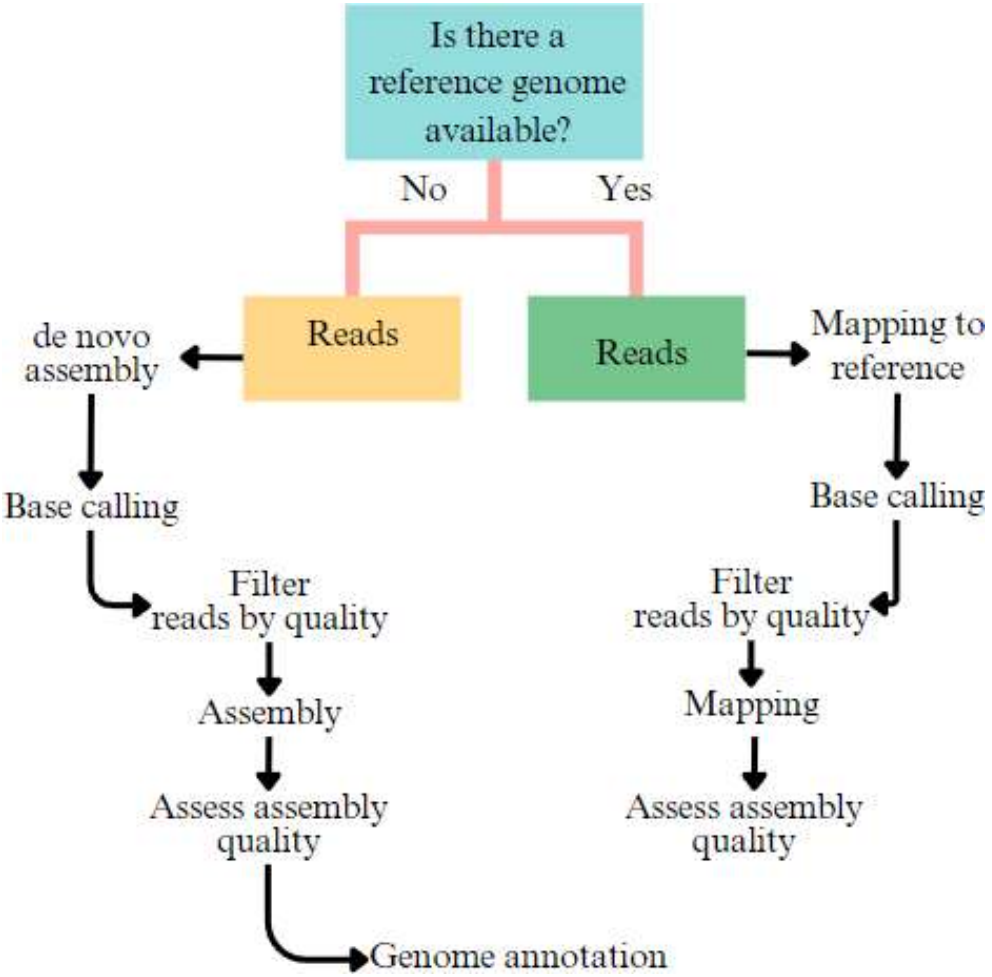

Figure 3. An assembly flowchart for genomes

ISSN PRINT 2319 1775 Online 2320 7876
Research paper Volume 12, Iss 1, 2023

Certain indices are used to assess the quality of the assembly, such as coverage, which measures the number of reads connected to a certain DNA segment. The amount of the genome that is covered by big contigs is shown by the N50. Half of the reads are contained in contigs with a size of n or more when the N50 value is n.

Extracting the biological information from the sequences is the next phase in the genome assembly process, which is correlated with its annotation (Prosdocimi, 2010). The distinctions between prokaryotes and eukaryotes led to the development of different methods for searching genomes for genes. The first phase is using sequence similarity to identify genes. Subsequently, the annotated genefunction is verified through comparison with protein databases, including UniProt and NCBI (Staats et al., 2014). Additionally, functional annotation is carried out, which entails using Gene Ontology (GO) words to link genes to biological processes. According to Prosdocimi (2010), these concepts categorise gene function into three classes: molecular function, biological processes, and cellular components.

Transcriptomics

Technologies such as DNA sequencing and hybridization have been developed to estimate and measure the transcriptome . Methods based on DNA microarray and real-time PCR (qPCR), while greatly advancing, have certain drawbacks. However, NGS platforms have become a viable substitutefor these technologies in the assessment of the global expression.

There are several tools available for data analysis. Three types of analyses are distinguished: i) read mapping; ii) transcript assembly; and iii) gene/transcript quantification. One of the most popular toolshas been the Tuxedo Suite protocol (Tophat/Cufflinks) (Figure 4).

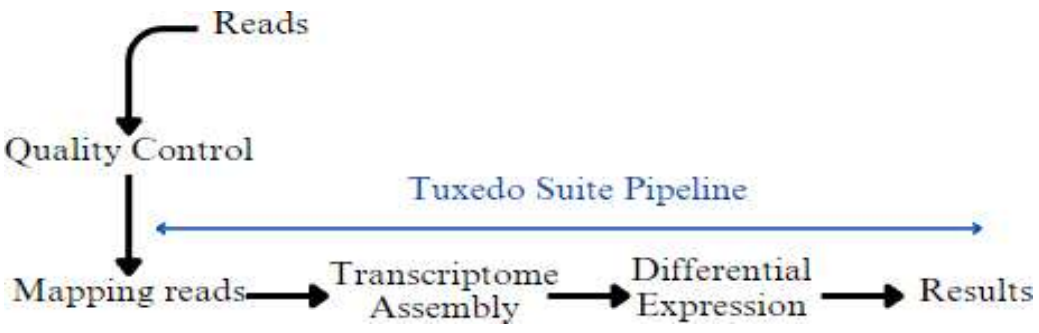

Figure 4. Method for differential expression analysis using Tuxedo Suite..

ISSN PRINT 2319 1775 Online 2320 7876
 Volume 12, Iss 1, 2023

These analyses are often carried out in the following ways: Tophat finds the splicing junctions and carries out the read mapping to the reference genome. The Cufflinks then use these alignments to assemble the transcripts, assess their abundance, and identify the genes and differentially expressed transcripts (Cuffdiff), which CummeRbund may visualise, under the testing conditions. Annotating biological processes and performing functional enrichment analysis are both possible using the list of genes that were produced by differential expression analysis. Additionally, these genes can be found participating in biological processes, for instance in the Reactome and KEGG databases.

Proteomics

Understanding the molecular mechanisms underlying cellular physiology requires the identification, measurement, and characterization of every protein in the cell. In this setting, the field of proteomics seems to have grown quickly in an attempt to organise the study of the connections, framework, motion, and function of proteins over space and time.

The discovered proteins might be connected to GO keywords and used to build biological pathways. A different strategy involves analysing protein-protein interactions through coexpression or by utilising databases like BioGRID and MINT .

Systemsbiology: thewhole isgreaterthanthesum oftheparts

"The whole is greater than the sum of its parts" is the guiding principle of systems biology, which offers a comprehensive method of understanding system complexity. This is a multidisciplinary science that aims to create new technologies, investigate data in new ways, and produce new findings and theories in order to start an innovative cycle (Figure 5).
Compared to a single data analysis, the systems approach at the genomic level enables comprehensive and useful inquiries concerning genotype-phenotype relationships. While the identification of genes and proteins holds significance, it is insufficient in comprehending the intricacy of the system. Understanding the four fundamental aspects of the system—its dynamics, control scheme, method design, and structure—will lead to a deeper understanding of the system.

ISSN PRINT 2319 1775 Online 2320 7876
Research paper Volume 12, Iss 1, 2023

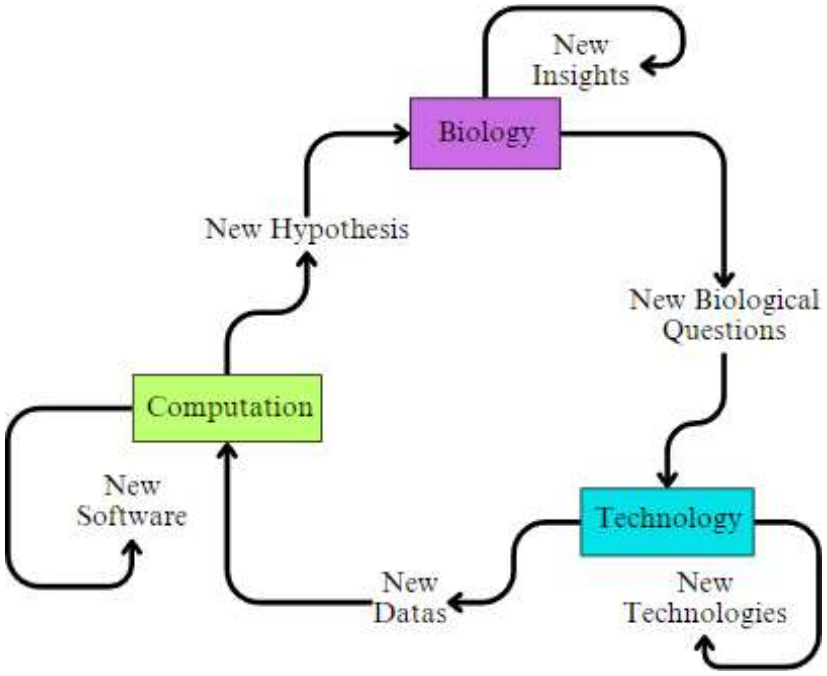

Figure 5. Systems biology is an interdisciplinary field of study.

Conclusion

Technological developments in data collection, analysis, and result interpretation have indicated a bright future. But widespread advancements across the board in science point to the rise of fresh approaches to analysis. The application of molecular-level data should progress to systemic techniques, which have the potential to revolutionise our comprehension of the regulation of intricate biological systems, while also deepening our grasp of how the body functions. However, data integration is not the last step. It's the start of fresh findings and theories that create a feedback loop. Significant advancements in health will also be made, including the application of genomic technologies to personalised medicine and gene therapy.

Abbreviation

KEGG - Kyoto Encyclopaedia of Genes and
GenomesDNA- Deoxyribonucleic acid
RNA - Ribonucleic acid

Competing interests

The authors declare that they have no competing interests.

Consent for publication

Not applicable

Ethics approval and consent to participate

Not applicable

Funding

This research study is sponsored by the institution name. Thank you to this college for supporting thisarticle!

Availability of data and materials

Not applicable

Authors' contribution

Author A supports to find materials and results part in this manuscript. Author B helps to developliterature part.

Acknowledgement

I offer up our fervent prayers to the omnipotent God. I want to express my sincere gratitude to my co- workers for supporting me through all of our challenges and victories to get this task done. I want to express my gratitude for our family's love and support, as well as for their encouragement. Finally, I would like to extend our sincere gratitude to everyone who has assisted us in writing this article.

ISSN PRINT 2319 1775 Online 2320 7876
 Volume 12, Iss 1, 2023

References

1. Bhardwaj, Kartik Krishna, Siddhant Banyal, and Deepak Kumar Sharma. (2019). "Chapter 7 - Artificial Intelligence Based Diagnostics, Therapeutics and Applications in Biomedical Engineering and Bioinformatics." In Internet of Things in Biomedical Engineering, edited by Valentina E. Balas, Le Hoang Son, Sudan Jha, Manju Khari, and Raghvendra Kumar, 161–87. Academic Press.
2. Egger, Maria, Matthias Ley, and Sten Hanke. (2019). "Emotion Recognition from Physiological Signal Analysis: A Review." Electronic Notes in Theoretical ComputerScience 343 (May): 35–55.
3. Evsutin, Oleg, and Kristina Dzhanashia. (2022). "Watermarking Schemes for Digital Images: Robustness Overview." Signal Processing: Image Communication 100 (January): 116523.
4. Gu, Xiaotong, Zehong Cao, Alireza Jolfaei, Peng Xu, Dongrui Wu, Tzyy-Ping Jung, and Chin- Teng Lin. (2021). "EEG-Based Brain-Computer Interfaces (BCIs): A Survey of Recent Studies on Signal Sensing Technologies and Computational Intelligence Approaches and Their Applications." IEEE/ACM Transactions on Computational Biology and Bioinformatics / IEEE, ACM 18 (5): 1645–66.
5. Haldorai, Anandakumar, Suriya Murugan, and Arulmurugan Ramu. (2021). "Evolution, Challenges, and Application of Intelligent ICT Education: An Overview." Computer Applications in Engineering Education 29 (3): 562–71.
6. Hosseini, Mohammad-Parsa, Amin Hosseini, and Kiarash Ahi. (2021). "A Review on Machine Learning for EEG Signal Processing in Bioengineering." IEEE Reviews in Biomedical Engineering 14 (January): 204–18.
7. Karasu, Seçkin, and Zehra Saraç. (2020). "Classification of Power Quality Disturbances by 2D- Riesz Transform, Multi-Objective Grey Wolf Optimizer and Machine Learning Methods." Digital Signal Processing 101 (June): 102711.
8. Lan, Lan, Lei You, Zeyang Zhang, Zhiwei Fan, Weiling Zhao, Nianyin Zeng, Yidong Chen, and Xiaobo Zhou. (2020). "Generative Adversarial Networks and Its Applications in Biomedical Informatics." Frontiers in Public Health 8 (May): 164.
9. Li, Yu, Chao Huang, Lizhong Ding, Zhongxiao Li, Yijie Pan, and Xin Gao. (2019). "Deep Learning in Bioinformatics: Introduction, Application, and Perspective in the Big Data Era." Methods 166 (August): 4–21.
10. Petmezas, Georgios, Kostas Haris, Leandros Stefanopoulos, Vassilis Kilintzis, Andreas Tzavelis, John A. Rogers, Aggelos K. Katsaggelos, and Nicos Maglaveras. (2021). "Automated Atrial Fibrillation Detection Using a Hybrid CNN-LSTM Network on Imbalanced ECG Datasets." Biomedical Signal Processing and Control 63 (January):

102194.

11. Pic, Xavier, Eva Gil San Antonio, Melpomeni Dimopoulou, and Marc Antonini. (2023). "Rotating Labeling of Entropy Coders for Synthetic DNA Data Storage." In 2023 24th International Conference on Digital Signal Processing (DSP), 1–5. IEEE.
12. Slovin, Shaked, Annamaria Carissimo, Francesco Panariello, Antonio Grimaldi, Valentina Bouché, Gennaro Gambardella, and Davide Cacchiarelli. (202)1. "Single-Cell RNA Sequencing Analysis: A Step-by-Step Overview." Methods in Molecular Biology 2284: 343– 65.
13. Varshney, Sandesh, Manisha Bharti, Sonali Sundram, Rishabha Malviya, and Neeraj Kumar Fuloria. (2022). "The Role of Bioinformatics Tools and Technologies in Clinical Trials."In Bioinformatics Tools and Big Data Analytics for Patient Care, 1–16. Boca Raton: Chapman and Hall/CRC.
14. Wang, Yunhao, Yue Zhao, Audrey Bollas, Yuru Wang, and Kin Fai Au. (2021). "Nanopore Sequencing Technology, Bioinformatics and Applications." Nature Biotechnology 39 (11): 1348–65.
15. Wasimuddin, Muhammad, Khaled Elleithy, Abdel-Shakour Abuzneid, Miad Faezipour, and Omar Abuzaghleh. (2020). "Stages-Based ECG Signal Analysis From Traditional Signal Processing to Machine Learning Approaches: A Survey." IEEE Access 8: 177782–803.
16. Wood, Alexander, Kayvan Najarian, and Delaram Kahrobaei. (2020). "Homomorphic Encryption for Machine Learning in Medicine and Bioinformatics." ACM Comput. Surv., 70, 53 (4): 1–35.
17. Zhang, Yongqing, Jianrong Yan, Siyu Chen, Meiqin Gong, Dongrui Gao, Min Zhu, and Wei Gan. (2020). "Review of the Applications of Deep Learning in Bioinformatics." Current Bioinformatics 15 (8): 898–911.