Genomics, High Performance Computing and Machine Learning

Vaidehi Thakre[1], Shreyas Vedpathak[2], and Sejal Sawarkar[3]
[1,2,3] Department of Computer Science and Engineering, MIT World Peace University
Email: [1]vaidehithakre21@gmail.com, [2]shreyasvedpathak@gmail.com and [3]sejalsawarkar2000@gmail.com

***Abstract*—** Genomic data has the potential to improve healthcare strategy in a variety of ways, including illness prevention, improved diagnosis, and better treatment. While Machine Learning may have revolutionized many fields, its implementation in the field of Genomics is new. Currently, Machine Learning is being applied and tested in a lot of genomic processes but all of those have not been clinically validated. Hence, we are far from providing Machine Learning or Deep Learning models for -omics data which can be implemented. This paper aims to explore in a very uncomplicated manner, what exactly is genomics, where does high performance computing and machine learning come into picture, current applications of machine learning in genomics and discuss potential future scope of machine learning in genomics.

***Keywords*—** Deep Learning, Genomics, High-Performance Computing, Machine Learning, Mass Spectrometry, Next-Generation Sequencing.

I. INTRODUCTION

The structure, function, evolution, mapping, and editing of genomes are all studied in genomics, which is an interdisciplinary subject of biology. A genome is a full set of DNA that includes all of an organism's genes.

Genomics is the study of an organism's entire genome, which includes genetic material. To sequence, assemble, and analyse the structure and function of genomes, genomics uses a combination of recombinant DNA, DNA sequencing technologies, and bioinformatics. It varies from traditional genetics in that it considers an organism's entire hereditary material rather than just one gene or gene product at a time.

The ability of a PC to process data and perform complex calculations at rapid speeds is known as high-performance computing (HPC). The supercomputer is one of the most well-known forms of HPC solutions. Thousands of compute nodes work together to execute one or more tasks on a supercomputer. Parallel processing is the term for this.

Machine Learning can be defined as “A branch of research that aims to understand and replicate intelligent behaviour using computational methods.” In simple words “Automate Intelligence in a Machine.”

Machine Learning can be a high-performance computing challenge since it necessitates a lot of computation and data movement (IO and network). Machine learning needs computationally intensive training and lots of computational power help to enable speeding up the training cycles. And HPC is used to analyse these large amounts of datasets.

Machine learning has been used to annotate a wide variety of genomic sequence components and is possibly most beneficial for the interpretation of big genomic data sets.

II. LITERATURE SURVEY

Proteogenomic data sets generate massive volumes of data, necessitating effective storage methods. For the sake of this presentation, we will use NGS data sets that can readily be expanded to MS data sets. Storage solutions are inefficient, thus specific compression methods have been suggested. However, the scalability of these specialised compression methods is limited, necessitating HPC solutions.

Any large data challenge necessitates the use of high-performance analytics. For the analysis of proteogenomic data sets, HPC techniques would be necessary. However, as has been demonstrated, tools designed using traditional parallel computing paradigms are prone to failure due to enormous data volumes. The next generation of HPC methods should be built with the data- and compute-intensive nature of proteogenomic problems in mind. For such analysis, HPC systems with techniques like data and dimensionality reduction, novel sampling, and sketching & streaming must be implemented keeping in mind their efficiency and feasibility. System Biology and Clinical labs usually do not have the resources to host big clusters and in such cases then, ubiquitous architectures like GPUs, TPUs and multi cores can prove extremely advantageous.

The influence of HPC on proteogenomic was investigated in F. Saeed [1]. Proteogenomic research necessitates the generation and integration of data from high-throughput technologies such as Next Generation Sequencing (NGS) and Mass Spectrometers (MS), yet present methods have been shown to be insufficient. [1] highlighted three major areas in which HPC can have a

significant influence in the realm of large data proteogenomic: storage, transmission, and analytics.

Libbrecht1 et al [2] discuss machine learning frameworks for analysing genome sequencing data sets as well as criteria for choosing an ML method. For accurate results, [2] stressed the need of theoretical and practical understanding of the relevant study application sector. As DNA sequencing and high-resolution imaging technologies become more common, new machine learning algorithms and experts will be in more demand. Liu et al [3] addressed how advances in DL and the use of CNN in sequence analysis had transformed genomics. L. Koumakis [4] discusses the growth of deep learning (DL) in genomics and how it surpasses conventional image processing approaches, [4] advocates integrating public and private datasets to improve prediction, but [3] stresses the necessity to analyse, identify, and create algorithms to automate feature extraction from big datasets. Deep Learning, according to Lu Zhang et al [5], is a better technique than Machine Learning. DL's capacity to accomplish difficult tasks on heterogeneous datasets without human input is highlighted. According to [5], finding large volumes of high-quality data is tough.

Genomics data are usually in a logical order and are frequently referred to as biological languages. As a result, recurrent models can be used in a wide range of settings. Cao et al. [6] developed an LSTM-based Neural Machine Translation system that translates protein function prediction into a language translation problem by recognising protein sequences as Gene Ontology terms. DeepNano was proposed by Boza et al. [7] for base calling, DanQ was proposed by Quang and Xie [8] to quantify the function of non-coding DNA, Convolutional LSTM networks were proposed by Snderby et al. [9] to predict protein subcellular localization from protein sequences, and so on. A recently suggested seq-to-seq RNN that can translate a variable-length input sequence to another sequence or anticipate a fixed-size prediction is also intriguing for genomic research.

Deep learning models outperform LASSO in analysing RNA-Seq gene expression profiles data, according to new research. In analysing RNA-Seq gene expression profiles data, Urda et al. [10] used a deep learning approach to surpass LASSO. Enhancer-promoter interactions are always predicted using non-sequence features from functional genomic signals. Singh et al. [11] developed the first deep learning approach for inferring enhancer-promoter interactions across the genome using just sequence-based information, as well as the locations of potential enhancers and promoters in a specific cell type. DeepFinder, a machine learning-based approach, was shown to be less effective than theirs (Whalen et al.,) [12].

The process of turning pre-messenger RNA into mature messenger RNA (mRNA), which may be translated into a protein, is known as splicing. Jha et al. [13] used previously established BNN (Xiong et al.,) [14] and DNN (Leung et al.,) [15] models to construct integrated deep learning models for alternative splicing. Their algorithms can detect splicing regulators and their potential targets, as well as infer regulatory rules from the genomic sequence.

III. METHODOLOGY

Researchers can "read" the genetic code that regulates all of a living organism's behaviours thanks to the ability to analyze DNA. To put it into perspective, the pathway from DNA to RNA to Protein is the basic dogma of biology. Base pairs are made up of four basic units called nucleotides (A, C, G, and T): A couples with T, and C couples with G. Humans have 23 chromosomal pairs that are organised by DNA.

Chromosomes are further divided into genes, which are DNA sequences that produce or encode proteins. The genome is the collection of genes that make up an organism. Humans have around 20,000 genes and 3 billion base pairs. Despite the fact that protein coding is a key focus in genomics academia and practice, only around 2% of the human genome codes for protein.

DNA is a four-letter code that contains all the information required to build a human body. A gene is a segment of DNA that carries the instructions for making a particular protein or group of proteins. On average, each of the human genome's 20,000 to 25,000 genes codes for three proteins.

Genes, which are located on 23 pairs of chromosomes in the nucleus of a human cell, use enzymes and messenger molecules to control the creation of proteins. For example, an enzyme transfers information from a gene's DNA to a molecule known as messenger ribonucleic acid (mRNA).

The mRNA goes from the nucleus to the cytoplasm, which is read by a ribosome, a molecular machine that uses the information to join small molecules known as amino acids in the right order to produce a specific protein.

Proteins are responsible for developing of body structures such as organs and tissue, as well as chemical

process regulation and information transfer between cells. When a cell's DNA is mutated, it produces an abnormal protein, which can impair the body's normal functions and lead to illnesses like cancer.

Almost every human ailment may be traced back to our DNA. Until recently, doctors could only examine the study of genes, or genetics, in cases of birth abnormalities and a small number of other disorders. These were diseases like sickle cell anaemia, which have relatively straightforward, predictable inheritance patterns since they are caused by a single gene mutation.

Thanks to the massive amount of data about human DNA created by the Human Genome Project and other genomic studies, scientists and doctors now have more sophisticated tools to examine the impact that multiple genetic variables working together and with the environment play in far more complicated illnesses. The majority of health problems in the United States include cancer, diabetes, and cardiovascular disease. Genome-based research is resulting in improved diagnoses, more effective therapeutic techniques, evidence-based ways to demonstrate clinical effectiveness, and better decision-making tools for patients and clinicians. In the end, treatments will almost probably be tailored to a patient's unique genetic makeup. As a result, genetics' role in health care is changing drastically, and the age of genomic medicine is quickly coming.

Now that we've established what genomics, HPC, and machine learning are, let's look at where they intersect. Before we accomplish that, we would want to mention or emphasise a few issues in genomics that researchers from Space-Time Insight, Inc. have identified:

- extracting the location and structure of genes
- identifying regulatory elements
- identifying non-coding RNA genes
- gene function prediction
- RNA secondary structure prediction

These are the few challenges identified and the key solution to these problems is to involve ML algorithms. ML has a wide range of applications in genomics right now, and the field's reach is still expanding. ML is now being utilised in applications such as classifying if something is a gene or not, genome sequencing, gene editing, validating DNA sequence strings, and many more. We know that the amount of data generated and required for genomic research is enormous. To be able to process and store that kind of data, we require extremely high-performance algorithm-based software and hardware.

A brief study of which ML algorithms are used in what genomic processes is summarized in Table 1.

Table I: Application of Machine Learning in Genomics [21]

Techniques in ML	Use
PyDNA Library	DNA/RNA/Protein Preprocessing Genomic Data Classification
NumPy Arrays	DNA Sequence String Processing
NLP Bag of Words	DNA Sequence String Processing
Multinomial Naive & Multi-Layer Perceptrons	Classification of DNA data
Convolutional Neural Networks	Classification of DNA data
Long Short-Term Memory	Classification of DNA data
Recurrent Neural Networks	DNA Sequencing
Naive Bayes and Bayseian Neural Networks	Gene Expression Regulation - Splicing
Long Short-Term Memory	Structural Classification of Proteins – Protein Homology
Residual Neural Networks	Protein Contact Prediction from Amino Acid Sequence
Convolutional Neural Networks	Prediction of Functional Activities of DNA
Label Encoding, K-Mer Counting, One Hot Encoding	Encoding DNA Sequences
Decision Trees	Location of Protein Coding Regions
Support Vector Machines	Identification of functional RNA genes

With any DNA sequence length, traditional ML methods (Linear and Logistics Regressions, Decision Trees, Support Vector Machines, Random Forest, Boosting Algorithms, Bayesian Network, and so on) may be utilised. Modern ANN algorithms, such as CNN and RNN, need that the DNA sequence length in each dataset column be consistent.

While all these are currently being used and explored on a more comprehensive manner, there is a lot more scope of expansion in this domain. A few future applications of ML in genomics [17], [18], [19], [20], [21] are:

Precision Medicine – The diagnosis and management of chronic illnesses such as cancer are highly complicated. Doctors and scientists are working on technology that will allow them to not only detect which type of DNA and mutations are present in the genome, but also to cut it off from the affected areas for a more effective therapy. AI-based algorithms can help to make treatment procedures more exact, so that such illnesses can not only be identified, but also treated individually based on a person's DNA. The current method, which is one-size-fits-all, may not be successful enough. As a result, this is one area where machine learning in genomics can be particularly beneficial.

Precision medicine and genomics go hand in hand. Personalized medicine is a patient-centred approach to care that integrates genetics, behaviour, and the environment to create a patient- or population-specific therapeutic intervention rather than a one-size-fits-all strategy, with a market size predicted to reach $87 billion by 2023. A client looking for a new blood transfusion would be paired to a donor who has the same blood type as them rather than a random donor to decrease the chance of problems.

Pharmacogenomics - In case of major outbreaks, like for example, the recent most COVID-19 pandemic, having population specific vaccinations is of utmost importance. Pharmacogenomics, we can say, is a natural progression of precision medicine. The population specific course of treatment ideology of course has a lot of its own hurdles like determining the right genetic set, finding genetic links for complex conditions and many more.

The forever increasing genomic data because of increasing population means that analysing all that data can be a daunting task. Supervised Machine Learning methods may offer a new approach for analysis, one that is particularly suitable for inferring from high-dimensional data produced by an unknown or imprecisely specified model.

Next Generation Breeding - Food production is majorly affected by plant diseases and unfavourable environmental conditions. For developing nations, this can be a very serious issue to deal with as they lead to losses. ML plays an integral part in the process of analysing phenotypes and providing relevant information or patterns. ML algorithms, on the other hand, aren't just for detecting variations from long-read methods; they may also be utilised by population genetics researchers. Indeed, supervised machine learning has been used to investigate recombination rates in a target genome.

ML applications aimed at soil health using proprietary ML models to identify the factors responsible for driving crop outcomes. The future of ML will be focused on how to cope with numerous species at the same time. Deep-learning methods may be able to handle comparative genomics investigations or information transfer from a model plant to a crop of interest.

Big Data Management - Handling NGS and MS data (Two major sources of genomic data production) requires high storage spaces and innovative computational capacities for the management of big data. Several repositories have been established in recent years to solve this problem. However, the amount of data created by NGS is quickly increasing (from hundreds of terabytes to petabytes in recent years), making storage a key problem in data computing.

The majority of NGS analysis software is command-line based, posing accessibility issues for many biologists, and making it difficult to choose the best performing/most suitable tools. In this context, ML methods, defined as a computer's ability to learn and interpret data without being pre-programmed, have the potential to improve accessibility.

Indeed, in addition to predictive analysis, Data classification and cataloguing algorithms can be used to incorporate ML in systems that can execute and handle data automatically.

Different types of data, ranging from huge datasets to individual tables, may be categorised in a variety of ways to meet the needs of users, allowing for stronger cross-category analytics correlations through identifier-driven searches and queries.

The potentialities of ML can be applied to give a better integration of data retrieved from genomics and phenomics, which will accelerate the process of developing prediction models.

A few barriers to applying ML approaches in -omics domain:
Dimensionality – The genomic data produced is observed to have a greater number of variables and small number of samples. Hence, to have the perfect standardized template for reference, a lot of preprocessing and harmonization is required.

The Classification Problem - The majority of DL and ML models for genomics are used to solve classification problems, such as distinguishing between disease and healthy samples. It is common knowledge that genomics trials and data gathered from different sources are fundamentally class imbalanced, and ML/DL models cannot be accurate unless enough instances per class have been fitted.

Heterogeneous Data issues - Since we deal with subgroups of the population, the data in most genomic applications is heterogeneous. In the system's biology level, genomic data includes I gene or non-coding transcript sequencing, (ii) quantitative gene expression profiles, (iii) gene variations, (iv) genome alternations, and (v) gene interactions. The covariates between the underling interdependencies among these heterogeneous data are one of the obstacles to integrating different data. There are numerous data tools available, but none are well structured, making model training extremely difficult.

WHY ARE THERE ONLY A FEW PEOPLE WORKING ON APPLYING ML/DL IN GENOMICS?

How do you define few? How many are now employed, and how many would suffice for you?

Most individuals in genomics, I believe, use some form of machine learning. PCA, for example, is a type of machine learning. Many genomics projects include a PCA step as a QC measure.

On the other hand, as shown in Fig 1 have you seen the scikit-learn Choosing the right estimator chart? [16]

See where it says ">50 samples" in the first step? If you work in genomics, it is almost impossible to find a project with 50 samples. By that notion, machine learning cannot be used in most genomics research.

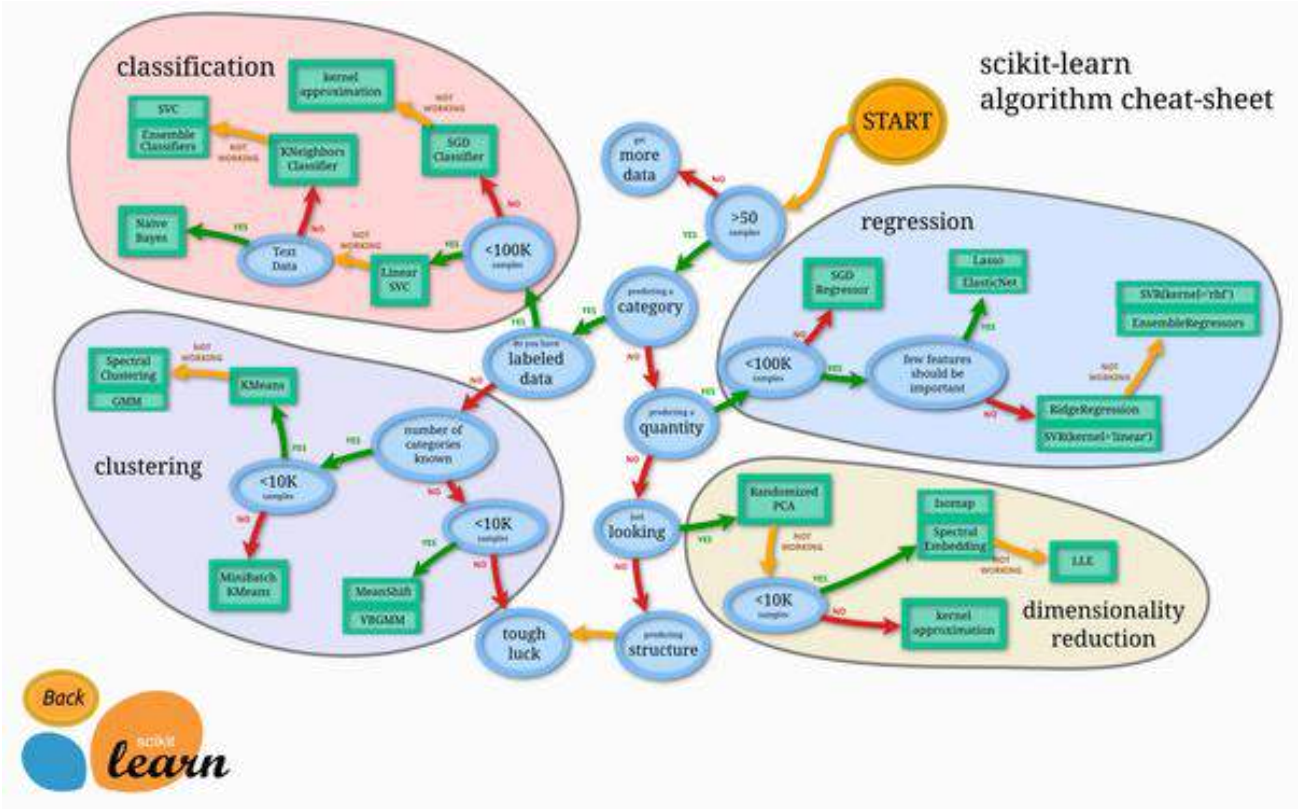

Fig. 1 scikit-learn algorithm cheat sheet

IV. CONCLUSION

To summarize, machine learning is a huge and difficult subject. Algorithms can be developed that allow for considerably more precise data analysis than many other approaches now available.

The type of machine learning approach chosen will be determined by the nature of the data provided and the goal of the researchers.

In the future, further research into machine learning and artificial intelligence will lead to more precise techniques to evaluate genetic data, resulting in additional discoveries. ML is being used and evaluated in a variety of genomic processes right now, but none of them have been clinically verified. As a result, we are still a long way from having DL models for -omics data that can be employed in precision medicine.

To improve the function of ML/DL genomics in prediction and prognosis, more efforts should be made to evaluate and integrate datasets (private and public). To cope with -omics-like data, more effort must be done in the HPC area, where methods must be both computational and data heavy.

In the future, machine learning models are projected to be widely used across the many -omics disciplines, increasing their integration and allowing for the resolution of important biological issues. This process will necessitate not just computing infrastructures and data analysis abilities, but also a higher level of sensitivity and an open mind when it comes to innovative models that may be used in many scientific fields. This will be aided through information exchange and multidisciplinary projects.

While there is a lot of potential, making the case for precision medicine is still a long way off, with many physicians wanting more clarity on clinical utility and insurance companies not seeing it as a need.

As a result, machine learning's data interpretation capabilities will need to be supplemented by education and clear explanations of the technology's utility and worth.

Pharmacogenomics is one of the most prominent developing uses of machine learning in genomics, although it is only one example of many possible future applications. However, with insufficient data on results, only time will tell which areas will profit the most from AI investments.

REFERENCES

[1] F. Saeed, "Big Data Proteogenomics and High-Performance Computing: Challenges and Opportunities", IEEE GlobalSIP 2015 -- Symposium on Signal and Information Processing for Software-Defined Ecosystems, and Green Computing, 2015.

[2] Maxwell W. Libbrecht1 and William Stafford Noble1, "Machine learning applications in genetics and genomics", Nature Reviews Genetics, 2015.

[3] Liu, J., Li, J., Wang, H., & Yan, J. “Application of deep learning in genomics”, Science China Life Sciences, 2020.

[4] L. Koumakis, "Deep learning models in genomics; are we there yet?", Computational and Structural Biotechnology Journal, 2020.

[5] Lu Zhang, Jianjun Tan, Dan Han and Hao Zhu, ” From machine learning to deep learning: progress in machine intelligence for rational drug discovery”, Elsevier, 2017

[6] R. Cao, C. Freitas, L. Chan, M. Sun, H. Jiang and Z. Chen, "ProLanGO: Protein Function Prediction Using Neural Machine Translation Based on a Recurrent Neural Network", Molecules, vol. 22, no. 10, p. 1732, 2017. Available: https://www.mdpi.com/1420-3049/22/10/1732.

[7] V. Boža, B. Brejová and T. Vinař, "DeepNano: Deep recurrent neural networks for base calling in MinION nanopore reads", PLOS ONE, vol. 12, no. 6, p. e0178751, 2017. Available: https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0178751.

[8] D. Quang and X. Xie, "DanQ: a hybrid convolutional and recurrent deep neural network for quantifying the function of DNA sequences", Nucleic Acids Research, vol. 44, no. 11, pp. e107-e107, 2016. Available: 10.1093/nar/gkw226.

[9] S. Sønderby, C. Sønderby, H. Nielsen and O. Winther, "Convolutional LSTM Networks for Subcellular Localization of Proteins", Algorithms for Computational Biology, pp. 68-80, 2015. Available: 10.1007/978-3-319-21233-3_6

[10] D. Urda, J. Montes-Torres, F. Moreno, L. Franco and J. Jerez, "Deep Learning to Analyze RNA-Seq Gene Expression Data", Advances in Computational Intelligence, pp. 50-59, 2017. Available: 10.1007/978-3-319-59147-6_5

[11] S. Singh, Y. Yang, B. Póczos and J. Ma, "Predicting enhancer-promoter interaction from genomic sequence with deep neural networks", Quantitative Biology, vol. 7, no. 2, pp. 122-137, 2019. Available: 10.1007/s40484-019-0154-0

[12] S. Whalen, R. Truty and K. Pollard, "Enhancer–promoter interactions are encoded by complex genomic signatures on looping chromatin", Nature Genetics, vol. 48, no. 5, pp. 488-496, 2016. Available: 10.1038/ng.3539.

[13] A. Jha, M. Gazzara and Y. Barash, "Integrative deep models for alternative splicing", Bioinformatics, vol. 33, no. 14, pp. i274-i282, 2017. Available: 10.1093/bioinformatics/btx268.

[14] H. Xiong, Y. Barash and B. Frey, "Bayesian prediction of tissue-regulated splicing using RNA sequence and cellular context", Bioinformatics, vol. 27, no. 18, pp. 2554-2562, 2011. Available: 10.1093/bioinformatics/btr444.

[15] M. Leung, H. Xiong, L. Lee and B. Frey, "Deep learning of the tissue-regulated splicing code", Bioinformatics, vol. 30, no. 12, pp. i121-i129, 2014. Available:10.1093/bioinformatics/btu277.

[16] scikit-learn, Choosing the right estimator

[17] K. Sennaar, "Machine Learning in Genomics - Current Efforts and Future Applications | Emerj", Emerj, 2021. [Online]. Available: https://emerj.com/ai-sector-overviews/machine-learning-in-genomics-applications/.

[18] A. Wickramarachchi, "Machine Learning For Genomics", Medium, 2021. [Online]. Available: https://towardsdatascience.com/machine-learning-for-genomics-c02270a51795.

[19] E. Bonat, "Apply Machine Learning Algorithms for Genomics Data Classification", Medium, 2021. [Online].Available: https://medium.com/mlearning-ai/apply-machine-learning-algorithms-for-genomics-data-classification-132972933723.

[20] R. Dias and A. Torkamani, "Artificial intelligence in clinical and genomic diagnostics", GenomeMedicine, 2021. [Online]. Available: https://genomemedicine.biomedcentral.com/articles/10.1186/s13073-019-0689-8.

[21] E. Bonat and B. Rayamajhi, "Apply Machine Learning Algorithms for Genomics Data Classification", Medium, 2021. [Online]. Available: https://medium.com/mlearning-ai/apply-machine-learning-algorithms-for-genomics-data-classification-132972933723#0f4e.