2022 International Conference on Innovative Computing, Intelligent Communication and Smart Electrical Systems (ICSES) | 978-1-6654-7413-9/22/$31.00 ©2022 IEEE | DOI: 10.1109/ICSES55317.2022.9914340

A Novel Machine Learning Approach of Multi-omics Data Prediction

Dr. B. Srinivasa Rao[1], S.Lavanya[2], K.Kajendran[3], Dr.Prince Prashant Sharma[4], Dr. Devvret Verma[5], Radhakrishnan P[6] and Dr.G.Manikandan[7]

[1]Professor, Department of Computer Science and Engineering, Gokaraju Rangaraju Institute of Engineering and Technology, Hyderabad Telangana, India
[2]Assistant Professor, Department of Information Technology, Karpagam College of Engineering, Coimbatore, Tamil Nadu, India
[3]Associate Professor, Department of Computer Science and Engineering, panimalar Engineering College, Chennai, TamilNadu, India.
[4]Assistant Professor, Department of Pharmaceutical Sciences, GurukulaKangri Deemed to be University, Haridwar, Uttarakhand
[5]Assistant Professor Department of Biotechnology, Graphic Era Deemed to be University, Dehradun, Uttarakhand, India, 248002
[6]Assistant Professor, Department of Computer Science and Engineering, Panimalar Engineering College, Chennai, TamilNadu, India.
[7]Assistant Professor, Department of Electronics and Communication Engineering, Saveetha School of Engineering, Saveetha Institute of Medical and Technical Sciences (Deemed to be University), Thandalam, Chennai-602105

E-mail : bsrgriet2015@gmail.com lavanya.skar29@gmail.com kajendran@yahoo.com devvret@geu.com rksiva13@gmail.com mrg.manikandan@gmail.com

Abstract- An integrated (combined) approach is needed in today's biomedical research in order to effectively use this data and obtain insights into natural systems. Genome, proteome, and metaproteomic data may be used to comprehend the complexities of molecular genetics utilising "machine learning" performance tracking methods derived from diverse omics sources. New biomarkers may be discovered by merging and analysing omics data using machine learning techniques. These biomarkers may aid in the proper identification of illness, the separation of patients, and the provision of tailored therapy. This study looks at a variety of "integrative machine learning or ML techniques" that are being used to obtain a better understanding of biological systems during natural bodily performance as well as when systems are diseased.Secondary data collection method has been used for this paper to gather relevant and factual data related to ML techniques of multi omics data prediction.

Keywords: "Machine Learning or ML", Omics, Biomarkers, Technology, omics evaluation.

I.INTRODUCTION

Medical technology is focusing on the collection and interpretation of high-throughput molecular tests in order to better understand patient and illness-specific variations. Health-related big data, such as connected patient clinical evidence (for instance, sex, age, clinical as well as physiologic background) and omics data (for instance, genotypes, proteomics, as well as metabolomics), is becoming more commonly available. A growing number of healthcare providers are turning to precision treatment, also known as individualised or stratified care, to provide patients with more personalised care [1]. ML (machine learning) methodologies and data mining technology have aided tremendously in the advancement of precision medicine. Researchers have utilised these techniques to find novel omics biological markers that may be used to pinpoint the cause of illness at its molecular level.

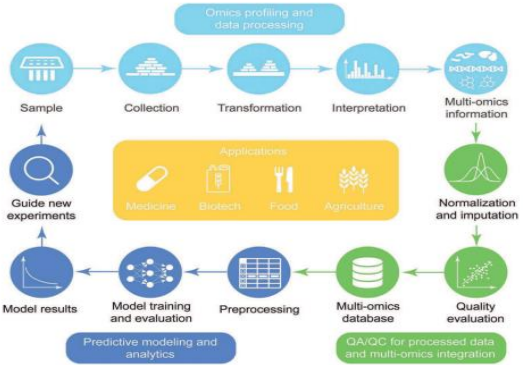

Fig 1: Structural pipeline of multi omics (Source: [2])

II. LITERATURE REVIEW

In the human body or in its by-products, a biomarker may be measured and utilised to infer the presence of disease or sickness. The identification of particular biomarkers is based on a series of omics data. Testing for C-reactive proteins with great sensitivity, for example, may identify reliably and quantitatively the risk of cardiovascular disease.

Treatments and options for patients are greatly influenced by biomarkers, which may be classified as diagnosing, prognosticating, or anticipating [3]. Patients may be diagnosed with the condition using diagnostic biomarkers, while prognostic biomarkers provide insight into how well they will do with or without conventional therapy. In order to identify those that pose a risk, predictive biomarkers are utilised in the process.

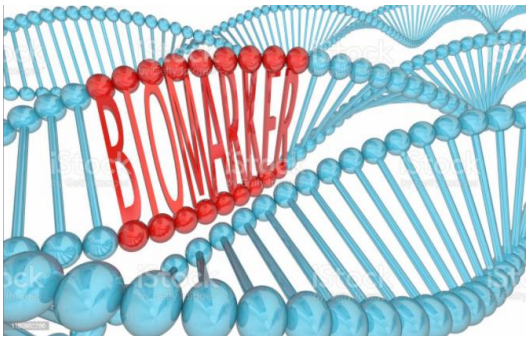

Fig 2: Biomarker Structure (Source: [3])

Using any or all of these signals, a patient's best course of treatment may be decided. According to the "ADNI (Alzheimer's Disease Neuroimaging Project) study," a combination of neurological, pharmacological, and genetic indications may reliably distinguish early-stage Alzheimer's patients from healthy controls.A computerised system that combines biomarkers from multi-site diffusion-weighted MRI imaging with a disease assessment score has also been utilized to analyze different forms of Parkinson's disorders (MDS-UPDRS III) (MDS-UPDRS III). High-risk people may be detected using biomarkers before they display any physical signs. They are also beneficial in monitoring the course of disease [4]. As a form of personalised medicine, ML has been used to create diagnostic, pathophysiologic, and parametric algorithms from single omics data. Also, ML's performance for some omics, like gene data, may have gotten worse because of things that are built into them.

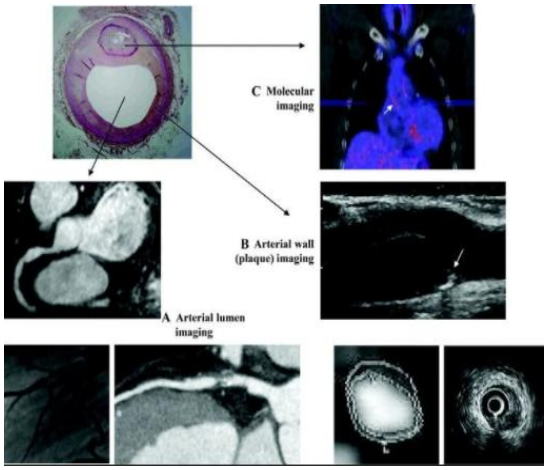

Fig 3: Imaging biomarker (Source: [5])

ML approaches are increasingly being used to examine and analyse the links between information as well as phenotypes in multi-omics datasets. While multi-omics ML evaluation is still in its infancy, this has already been investigated for a multitude of scenarios, as documented in comprehensive studies on mental illnesses, diabetes, leukaemia, coronary heart disease, medical image processing, single-cell evaluation in people, as well as plant science research. Many multi-omics analyses are now concentrated on certain sub-topics. Conducting research, establishing procedures, selecting software applications, and assessing effective feature outcomes are just a few examples [6]. ML techniques are computerized strategies that can assist in understanding complex features from experimental observations in big data research. A pattern recognition technique's purpose is to allow an algorithm to understand historical or relevant information and use that information to make recommendations or choices about uncertain potential events [7]. In a broad sense, an ML solution's workflow involves three stages: Classifying data using a network built from sample inputs, analysing and adjusting the method, and finally using the prototype. “Naive Bayes”, “C4.5 decision trees”, “Artificial neural networks (ANNs)”, “support vector machine (SVM)”, “k-Means”, “k-nearest neighbours (KNN)”, and “regression” are some of the most well-known ML approaches.

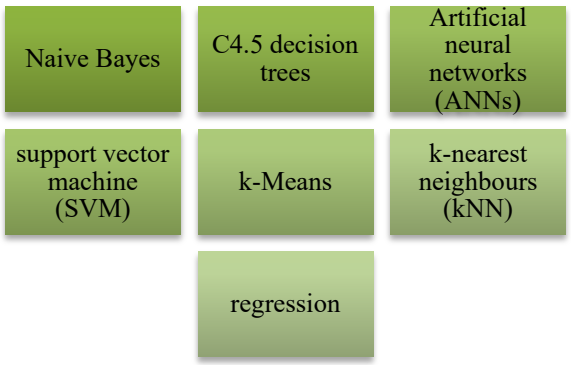

Fig 4: Different types of ML approaches
(Source: Created by researcher)

This analysis, on the other hand, intends for a larger audience, offering a grounding on multi-omics and machine learning to newcomers to this field. It advances expert-invented fusion terminology and presents contemporary integrated state-of-the-art methodologies.

DNA (deoxyribonucleic acid) and mRNA (messenger ribonucleic acid) are thought to be the primary pathways by which genetic mutations are transmitted from one cell to another. The analogy of a computer network to this data flow has made it easier to understand how biological information moves. Genomic research, mRNA research, and protein research are all terms for the study of DNA, mRNA, and proteins. Genetics is used to study the structure of the genome. It looks at certain parts of DNA to see if certain genes are there or not [8]. [8][9] In order to understand what is going on at the microscopic level, transcriptomics studies the genes that are efficiently transmitted. To better understand how cells and animals communicate information, proteomics may help identify protein channels and networks.Despite the fact that metabolomics, molecular techniques, and Glycosmis are not included in the core dogma analysis, they often provide a great deal of information about metabolites, fatty acids, and glycoprotein (Biotransformation processes are used by the proteomics to generate proteins). Since these compounds are intermediary by-products of a cell's communication process, they are great markers of the cell's functioning [9]. Metagenomics is a technique for sequencing genetic data from ambient materials without isolated specific species, comparable to single-genome investigations.

Every piece of omics information can be used as a biomarker to help understand and study the basic traits and complexity of living things.They're all elements of the same biological information chain, with varied sources and controls affecting the outcome. For genomes and metabolomics, specialised high-throughput methods and spectrophotometry may be used to analyse each one of these omics. The goal of ML techniques is to gain information from historical or current data and use that expertise to produce projections or decisions for unspecified future data measurements. Many books and articles have been written about the basics of machine learning and how it can be used.In circumstances where inventing and implementing explicit strategies with optimal outcomes is difficult, such as email classification, hand-written text categorization, and object recognition, machine learning is used [10]. It has also been used in self-driving cars, cyber-security, computerised assistants like 'Siri,' portals that suggest things based on other person's shopping habits, and innovative approaches to some of the world's most difficult challenges. Deep learning has risen to prominence as the most popular type of machine learning algorithm in recent years. It finds sophisticated representations of the data using neural networks, which are made up of hidden layers that execute various functions. It has advanced classification performance exceeding typical machine learning techniques, particularly in circumstances involving huge databases with high dimensionality. As a result, it's computationally taxing, necessitating powerful hardware, and it lacks interpretability (accountability) in feature extraction (a black-box method). This is because the system's connectedness makes it difficult to obtain features like classification from the system.Deep learning, on the other hand, presents an interesting potential in the framework of multi-omics unification. iOmicsPASS uses a network-based approach to integrate multi-omics profiles across genome-scale intracellular pathways [11]. The method includes analysis elements that convert descriptive multi-omics information into numbers for biological engagement, then uses the eventually results in numbers as information to select model-based subnets; ultimately, it uses a unique nearby shrunken centroid method to obtain forecasting corners for phenotypic clusters. The researchers measured iOmicsPASS on data from "Breast Invasive Ductal Carcinoma", combining mRNA activity and polypeptide accumulation with and without normalising the mRNA information by DNA "Copy Number Variation (CNV)".

iOmicsPASS outperforms the step in order when compared to the earlier closest "shrunken centroid classification" technique, emphasizing the significance of choosing predictive signatures from fully connected subnets, thus restricting the search area of input to the model to recognised connections. By using a specific phrase as well as scanning throughout all resources, this data was gathered from the Science Citation Index. Even though the application of machine learning in medical technology goes all the way back to the 1970s, it has grown at a far faster rate in the last ten years. Also, in the last five years, papers on multi-omics fusion and multi-omics integrated ML have become more popular in the area of accuracy and analytical pharmacy. While supervised learning is frequently used in other fields (such as diagnostic imaging and medicinal computational linguistics), demand in multi-omics research has been restricted. This is due to the fact that multi-omics investigations are difficult to implement since they need specialised high-throughput omics facilities. This is supported by the finding that much of the ongoing project for cancer prediction and anti-cancer therapeutic response uses machine learning on large-scale multi-omics information from various sources such as “The Cancer Genome Atlas (TCGA)”, “Cancer Cell Line Encyclopaedia (CCLE)”, and “Genomics of Drug Sensitivity in Cancer (GDSC)”. The application of ML to the analysis of high-throughput produced multi-omics data presents a number of distinct obstacles.

III. Methodology

Secondary research method has been considered for this research to manage data related to ML approaches of multi omics data prediction. In order to perform this, researchers have used different databases such as PubMed, Google scholar and ProQuest to manage data. Keywords are used to select proper articles.

Research questions

Why ML approaches are used in multi omics data prediction?

What are the benefits of implementing ML approaches?

IV. Analysis And Discussion

From this research paper, the following is a summary of what they include. Regarding omics evaluation, transcriptomics as well as proteomics, for instance, utilise various regularisation and scalability procedures. As a result, the variable bounds and distributed processing fluctuate. Furthermore, certain omics are more susceptible to producing discrete data than others (for example, in metabolomics, certain readings may be underneath the detection limits and thus given a null value). As a result, before designing their integration, each omics rectification and anomaly analysis should be evaluated independently. Several sickness categories are more uncommon than others in illness categorization, resulting in imbalanced data in the multi-omics database [12]. High blood pressure, for instance, is by far the most prevalent kind of hypertension, accounting for 95% of cases, whereas neuroendocrine hypertension affects just 5% of people. Overfitting occurs when a machine learning classifier is developed on an unbalanced database, resulting in high reliability for the testing phase but poor performance for an unknown testing dataset. As a result, one of the accompanying methodologies can be used to categorize these two different forms of hypertension:

1. if appropriate, gather more information,
2. explore using balanced or standardized measures to quantify ML effectiveness (such as F1-Score or Kappa), or
3. investigate over or principal component analysis the under or over-represented category, accordingly and
4. for the under-represented category, attempt synthesized collection production (such as SMOTE or ADASYN).

To address the bias-variance trade-off, approaches like normalisation, compressing, hyperparameter adjustment, as well as cross-validation, can be applied. To tackle the imbalance problem and generalization error difficulties, any of the abovementioned ways can be utilised, based on the information and situation. Several novel data integrating techniques have been investigated in recent years as a result of recent advances in theoretical, analytical, and computer sciences. This article contains an overview of a few evaluations that span the scope of multi-omics fusion for broad as well as speciality fields such as cancer and toxicity for the convenience of the users. The majority of these evaluations have attempted to propose several category terminologies (for instance, "initial," "later," and "middle" in or "bottom-up" as well as "top-down" in) that allow them to categorise

integrating techniques depending on various factors/parameters. As previously stated, this section uses Ritchie's category terms and expands on them to cover the whole range of contemporary integration approaches. It succinctly covers them, offering a novice multidisciplinary user a clear viewpoint. The different integration approaches are classified as "concatenation-based," "model-based," or "transformation-based." Concatenation-based integrating approaches use a joint set of data created by merging several omics datasets to build a model. Stage 1 contains the raw data from three different omics (for example, genomes, proteomics, as well as metabolomics) as well as the phenotypic characteristics. Concatenation-based integrating often does not need any pre-processing, and therefore there is no Stage 2. The information from each omics is synthesized in Stage 3 to generate a single big matrix of multi-omics information. Ultimately, the connection matrix is employed for either supervised or unsupervised evaluation in Stage 4. When all separate omics have been concatenated, the key benefit of adopting concatenation-based approaches is the ease with which ML may be used to analyse continuum or categorized information. These approaches employ all of the concatenated characteristics, in the same way, allowing them to choose the most discriminative information for a particular phenotype. Researchers have presented iClusterBayes, which is a completely Bayesian latent factor framework. In terms of numerical interpretation and processing performance, it addresses the constraints of iCluster+. iClusterBayes has a binary indicator prior to selecting features that extrapolate to binary as well as count data. MOFA (Multi-Omics Factor Analysis) was also created by researchers, which disentangles the heterogeneity maintained across distinct omics to identify the primary cause of unpredictability. It can combine datasets that are partially redundant.

V. Conclusion

A wide range of multi-omics integration methods are now available for both supervised and unsupervised learning, as shown in this work. It is possible that interdisciplinary scientists may be overwhelmed by this information, which would need a lengthy effort to master the rigorous mathematics and computing concepts that support it. This means that multi-omics teams should include ML experts to aid with approach selection, solution building, and result interpretation. They should also discuss the value and limits of ML specialists in their work.Such really multidisciplinary teams provide an actual opportunity for cooperation and collaboration of the many areas, practises, and skills required, eventually leading to more solid findings. It depicts the many decision phases involved in selecting the best technique (or range of techniques) for a particular situation. A summary of existing multi-omics investigations has also been provided. Furthermore, a multidisciplinary expert can use a suggestion process to select an acceptable approach for a piece of multi-omics information. Ultimately, this research emphasises current discoveries in the multi-omics sector and underscores the crucial relevance of machine learning in the creation of customised therapy.

References

[1] Reel, P.S., Reel, S., Pearson, E., Trucco, E. and Jefferson, E., 2021. Using machine learning approaches for multi-omics data analysis: A review. Biotechnology Advances, 49, p.107739.

[2] Nicora, G., Vitali, F., Dagliati, A., Geifman, N. and Bellazzi, R., 2020. Integrated multi-omics analyses in oncology: a review of machine learning methods and tools. Frontiers in oncology, 10, p.1030.

[3] Vijayakumar, K., Pradeep Mohan Kumar, K. &Jesline, D. Implementation of Software Agents and Advanced AoA for Disease Data Analysis. J Med Syst 43, 274 (2019). https://doi.org/10.1007/s10916-019-1411-5.

[4] Ma, B., Meng, F., Yan, G., Yan, H., Chai, B. and Song, F., 2020. Diagnostic classification of cancers using extreme gradient boosting algorithm and multi-omics data. Computers in biology and medicine, 121, p.103761.

[5] V. Ganesan, K. Vijayakumar, V. A. Devi and P. Ramadoss, "Hybrid Intelligence for Multimedia Data in Intra IoT (IIoT) Cloud by Persistent homology," 2022 International Conference on Advances in Computing, Communication and Applied Informatics (ACCAI), 2022, pp. 1-5, doi: 10.1109/ACCAI53970.2022.9752558.

[6] M. Sonika and S. B. G. T. Babu, “Analysis of Channel coding performance for wireless communications,” J. Study Res., vol. XII, no. 29, pp. 29–49, 2020.

[7] Zhang, L., Lv, C., Jin, Y., Cheng, G., Fu, Y., Yuan, D., Tao, Y., Guo, Y., Ni, X., Shi, T., 2018. Deep learning-based multi-omics data integration reveals two prognostic subtypes in high-risk neuroblastoma. Front. Genet.

[8] S. B. G. T. Babu and C. S. Rao, “Statistical Features based Optimized Technique for Copy Move Forgery Detection,” 2020 11th Int. Conf. Comput. Commun. Netw. Technol. ICCCNT 2020, 2020.

[9] C. Srinivasa Rao and S. B. G. TilakBabu, “Image Authentication Using Local Binary Pattern on the Low Frequency Components,” in Lecture Notes in Electrical Engineering, vol. 372, Springer Verlag, 2016, pp. 529–537.