

Machine Learning-based Multi-Omics Data Analysis for Cancer Diagnosis and Prediction

Shweta A. Gode¹ and Roshan Umate²

¹Department of Computer Technology/ Yeshwantrao Chavan College of Engineering, Nagpur, India

²Research consultant, Dept. of Research and Development/ Jawaharlal Nehru Medical College, Datta Meghe Institute of Higher Education and Research, Wardha, India

Email: shwetagode16@gmail.com

Abstract—Recent progress in bioinformatics has made it possible to extract useful information from massive datasets of gene expression. Researchers are increasingly concerned with how to identify important genes among these massive databases. MicroRNAs were identified as short non-coding RNAs that regulate gene expression by binding to the 3' UTR of target messenger RNAs and so inhibiting protein synthesis in accordance with the findings of the in-depth study. A variety of diseases, including cancer, have been linked to deregulation of microRNAs since their discovery. Using multi-omics data analysis, this study intends to delve into the characteristics and roles of these microRNAs and genes. MicroRNAs and genes are singled out for their functional and biological relevance across a wide range of cancer types using state-of-the-art machine learning algorithms. Furthermore, ligand binding activity prediction of Cyclin-Dependent Kinases is predicted using machine learning approaches. Eukaryotic cell division is coordinated by a group of proteins called cyclin-dependent kinases. Cyclin-dependent kinase crystal structures often reveal potential molecular processes of ligand binding.

Index Terms— microRNAs, Cyclin-Dependent Kinases, multi-omics, machine learning, ligand binding.

I. INTRODUCTION

According to research cancer is considered a leading global killer. International Agency for Research on Cancer just published data The International Agency for Research on Cancer (IARC) projects that 19.3 million people will be diagnosed with cancer by 2020, with 10.0 million losing their lives to the disease. The fact that 28.4 million individuals will be diagnosed with cancer during the next 20 years highlights the critical nature of this problem and the need for preventative actions. Therefore, it is clear that cancer is a major issue that people face and that it may have negative impacts on both individuals and communities. Therefore, learning the ins and outs of cancer's cellular processes might be a huge step towards better cancer prediction, diagnosis, and therapy, and would have far-reaching consequences in the long run Cancer is a general term used to describe a group of diseases in cellular and molecular biology that include conditions where cells can multiply and spread unchecked, a phenomenon known as metastasis. Cancer does not start and then develop all at once, according to the available evidence. On the contrary, the hallmarks of cancer cells emerge in a sequential fashion. These processes reflect the gradual buildup of molecular alterations in many biological components such as DNA, RNA, proteins, metabolites, and so on. The discipline of omics research and associated technologies have made

significant strides in understanding the processes behind cancer's start and progression in recent years. This is because molecular abnormalities may impact several levels of cell biology. Although researchers have gained a better understanding of cancer through single-layer-omics studies, a multi-omics approach to cancer biology is necessary to unravel the complex web of genotype-phenotype environmental interactions.[1][2] The use of multi-omics research has the potential to open up new avenues for cancer study, including studying how cells react to therapies and finding new ways to diagnose and treat the disease. Data explosions are so commonplace in cancer research of this magnitude.[3] A wealth of cancer research and multi-omics databases are at your fingertips, allowing you to piece together the molecular patterns at work in cancer across the genome, transcriptome, proteome, metabolome, and beyond. Complex computational analysis is necessary due to the presence of huge data in the domain of multi-omics investigations. Thus, thanks to the lightning-fast progress in technology, AI and related subfields, especially ML, have made great strides in aiding the analysis of big data resulting from multi-omics studies in the fight against various diseases, including cancer. Collaborations between ML and multi-omics data analysis of cancer have made some progress in recent years.[4] The main goal is to provide a comprehensive picture of the intricate patterns involved in the cancer process. The purpose of this study is to encourage future work in the area of machine learning (ML) applications in cancer multi-omics data analysis by outlining the many facets of cancer biology and the approaches currently used to treat the disease. Next, we will focus on the multi-omics arena and how it relates to cancer research.[5] Afterwards, the most up-to-date developments in combining ML algorithms with cancer omics data evaluations are detailed. In the end, we go over some of the problems with using ML for omics data analysis as well as potential viable solutions.

II. CYCLIN-DEPENDENT KINASES

Members of the family of serine-threonine protein kinases are referred to as cyclin-dependent kinases (CDKs). They contribute to the regulation and maintenance of the cellular cycle. Cyclin-dependent protein kinases and their associated cyclins control the cell cycle at discrete checkpoints. Either cell death or an out-of-control cell cycle may result from cyclin activity being disrupted. Cell cycle progression is controlled by cyclin concentration, T-loop phosphorylation [68], and the presence of endogenous cyclin-dependent kinase (CDK) inhibitors (CDKIs). Some CDKs must be activated before the cell cycle may advance to the next checkpoint, whereas others must be inhibited if the cycle is to be stopped. Beginning S-phase and entering mitosis are the primary cell-cycle transitions that call for CDK activity. Most of the target proteins that were phosphorylated when the cells entered mitosis are likely dephosphorylated together with the loss of CDK activity that permits the return to interphase. In addition, there are several research initiatives aimed at identifying novel CDKIs that have the potential to serve as therapeutics for disorders involving cell-cycle progression, such as malignancies. CDKs function in both cell cycle regulation and mRNA processing.[6] Only 8 of the 11 CDKs discovered so far are thought to be involved in cell cycle control. The eight cyclin-dependent kinases (CDKs) CDK1, CDK2, CDK3, CDK4, CDK5, CDK6, CDK7, and CDK9 are being studied for their roles in cell cycle control.[7]

Machine learning is a kind of data analysis that gives computers the ability to "learn" from their own mistakes and successes. In bioinformatics, it may be used in a wide variety of contexts. Most biological data is too complex for the efficiency of traditional approaches based on explicit programming to handle. Here, computers can sift through mountains of biological data using a technique called "machine learning." Furthermore, machine learning enables fully and partially automated data-driven learning. Data from the fields of genomics, proteomics, and system biology have been used extensively in studies using machine learning. Genomics is the study of the structure and function of complete genomes, such as that of humans. Scientists in the area of genomics have been compelled to utilize machine learning due to the exponential growth of raw genomics data and the need to physiologically interpret these data.[8] Machine learning algorithms were originally put to use in genomics for the purpose of gene prediction. Since then, machine learning methods have been used in many areas of genomics. Finding common ancestry among closely related sequences was a significant field of study in this context. Proteomics refers to the systematic study of proteins and their roles. Protein sequence prediction was formerly performed by the laborious process of manually analyzing amino acid sequence. Given the ever-increasing volume of proteomics data, the event was not only prohibitively costly and time-consuming, but also impractical.[9] These issues are addressed by the automated feature learning and prediction introduced by machine learning methods. In System Biology, researchers are looking for evidence of communication between biologically basic building blocks including DNA, RNA, and proteins. Specifically, machine learning has been used to locate binding sites for transcription factors. In addition, machine learning algorithms have made substantial use of genetic and regulatory networks. The use of machine learning to the study of gene expression data opens up new avenues for categorizing cancers. In addition to the above, clustering on the gene level is a

powerful tool for analyzing gene expression data. These techniques may identify a set of functionally related genes that are co-expressed. Therefore, a clustering approach may also be used to understand the functioning of genes with uncertain function.[10] Class prediction and regression are two common applications of supervised machine learning algorithms. They deviate from the rule that "one size fits all." When the number of genes to be analyzed exceeds the size of the available data sample, supervised approaches excel. The interpretability of classifiers may be enhanced by using biological information in conjunction with other approaches. Support Vector Machines (SVM) is a widely used learning classifier in the field of gene expression research, and it is used here. SVM can compare gene expression data with relative ease by using distance functions. Correlations between gene expression measures may be used by the SVMs since distance functions work in very high-dimensional feature spaces.[11] Furthermore, SVM is a supervised approach, meaning it relies on expert knowledge to make classifications between gene subtypes. This study aims to create computer models that can accurately identify the miRNAs and genes that cause various forms of cancer. Additionally, a model is constructed to predict CDK ligand binding activity. Machine learning is used here to both analyze and actualize the learning mechanism. High-throughput sequencing's rapid expansion and the resulting flood of multi-omics data made it difficult to zero in on genes and miRNAs of interest. The goal is to research the feasibility of creating new strategies for predicting miRNAs and genes relevant to cancer. Understanding how miRNAs and mRNAs interact together as a regulatory system is a major focus of this study. In most cases, there isn't a panacea that can solve all of these issues, so it's up to us to figure out what works best for each one individually. This study creates new computational models tailored to the situation at hand and evaluates their performance in comparison to state-of-the-art models.[12]

Panels of miRNAs that are useful indicators for various cancers are extracted using an in-silico analysis. The miRNAomes of a set of healthy and malignant samples are used to simulate the challenge of biomarker discovery as a two-class classification job. The goal is to identify the miRNA combination that yields the most precise categorization. Gene expression analysis provides insight into gene function and regulatory processes. However, researchers have recently shown that it is difficult to choose meaningful genes from big datasets.[4] A computational approach is provided here to distinguish between breast cancer subtypes like Basal, Human epidermal growth factor receptor 2 (Her2), luminal A (LumA), and luminal B (LumB), and to discover important genes for each. To determine the miRNAs that are dysregulated and their mRNA targets using next-generation sequencing data, a unified strategy is presented. The cyclin-dependent kinases (CDKs) are the key regulators of the cell cycle in eukaryotes. In most cases, we can learn more about potential molecular processes of ligand binding from the crystal structure of CDKs. Here, we focus on the development of a computer model for accurate and precise prediction of ligand binding activity.[2]

III. METHODOLOGY

The suggested strategy is predicated on the use of optimization algorithms to identify candidate gene biomarkers. Here, we evaluate the accuracy of predictions made using several methods, including Ant colony optimization, Differential evolution, Particle swarm optimization, and Artificial bee colonies. These algorithms iteratively enhance their potential solutions as they hunt for the best feature subsets or genes. Different optimization methods may use different sets of genes because of the inherent randomness of the algorithms. The likelihood of the algorithms giving poor answers with false positives and/or false negatives is mitigated by taking into account the randomization impact of the applied methods. False positives and negatives are still possible with this method, although they can be mitigated to some extent.

A. Fitness Computation using Support Vector Machine

The purpose of a fitness function is to provide a numeric measure of the performance of optimisation techniques. The goal here is to choose a function that improves the likelihood of discovering genes with differential expression levels between normal and tumour samples. This, in turn, may be seen as a problem in which two classes need to be distinguished. Regarding classification, an SVM classifier with an RBF kernel is used here. The SVM is given individual datasets ranging in size from P to L, and uses their classification accuracy to determine a fitness function. Classification accuracy is enhanced by using 10-fold cross-validation. In Fig. 1, we see a block schematic of the suggested procedure.

III. CONCLUSIONS

Research on analyzing multi-omics data is on the rise thanks to the widespread use of Next-Generation Sequencing and the fast development of bioinformatics analysis tools. Various machine learning technologies

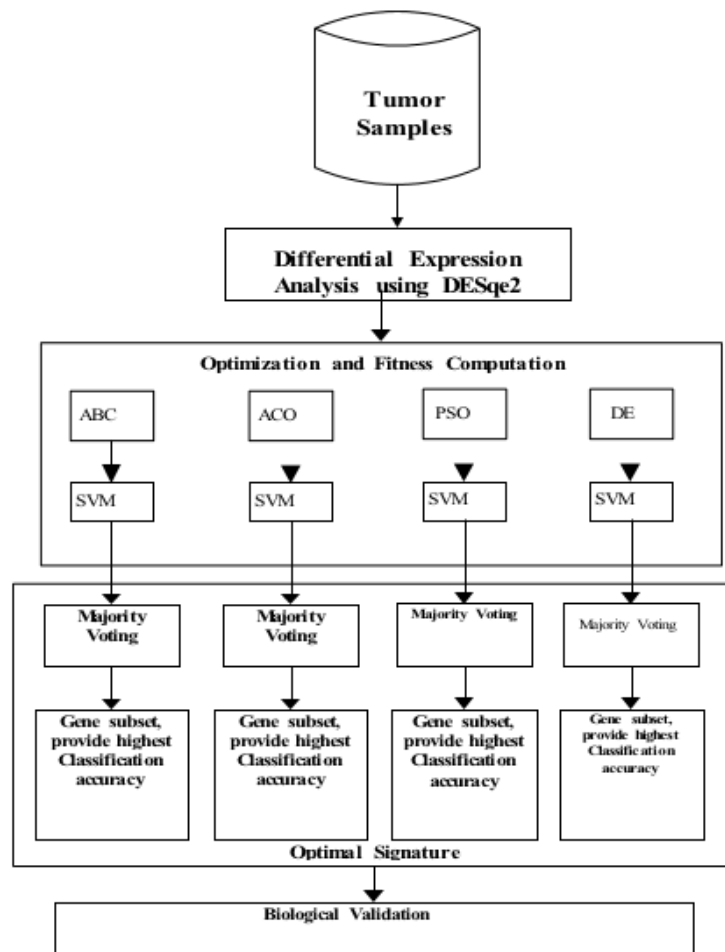


Figure 1: Block diagram of the proposed work

are seeing rapid growth in this area. The study of multi-omics datasets, including microRNA and messenger RNAs, is now dominating efforts in bioinformatics. Several features of microRNA and messenger RNA expression data that have long been linked to cancer and other disorders are investigated in this study. Extraction of reliable and predictive biomarkers responsible for various malignancies and cancer subtypes, as well as the development of novel methodologies for the analysis and prediction of these biomarkers, is the primary focus of current research. The messenger RNA targets of misregulated microRNAs and the nature of the regulatory interaction between microRNAs and mRNAs are both recommended to be recognised using a unified strategy. When it comes to eukaryotic cells, however, the Cyclin-dependent kinases are the primary players in directing the division process.

REFERENCES

- [1] G. Sruthi, C. L. Ram, M. K. Sai, B. P. Singh, N. Majhotra and N. Sharma, "Cancer Prediction using Machine Learning," 2022 2nd International Conference on Innovative Practices in Technology and Management (ICIPTM), Gautam Buddha Nagar, India, 2022, pp. 217-221, doi: 10.1109/ICIPTM54933.2022.9754059.
- [2] R. R. Sekar, M. R. Kumar, M. V. Vardhan, K. V. Gopal and C. Jayakkam Allan Tilak, "Approach to Predict Early Stage of Breast Cancer using Machine Learning," 2023 7th International Conference on Intelligent Computing and Control Systems (ICICCS), Madurai, India, 2023, pp. 122-125, doi: 10.1109/ICICCS56967.2023.10142851.
- [3] P. Rawat, M. Bajaj, S. Mehta, V. Sharma, A. Jain and M. Manjul, "Cancer Malignancy Prediction Using Machine Learning: A Cross-Dataset Comparative Study," 2023 International Conference on Computational Intelligence, Communication Technology and Networking (CICTN), Ghaziabad, India, 2023, pp. 699-704, doi: 10.1109/CICTN57981.2023.10140340.
- [4] R. Gurusamy, V. Rajmohan, N. Sengottaiyan, P. Kalyanasundaram and S. M. Ramesh, "Comparative Analysis on Medical Image Prediction of Breast Cancer Disease using Various Machine Learning Algorithms," 2023 4th

- International Conference on Electronics and Sustainable Communication Systems (ICESC), Coimbatore, India, 2023, pp. 1522-1526, doi: 10.1109/ICESC57686.2023.10193635.
- [5] D. Rawat, "Validating and Strengthen the Prediction Performance Using Machine Learning Models and Operational Research for Lung Cancer," 2022 IEEE International Conference on Data Science and Information System (ICDSIS), Hassan, India, 2022, pp. 1-5, doi: 10.1109/ICDSIS55133.2022.9915898.
 - [6] A. Verma, C. K. Shah, V. Kaur, S. Shah and P. Kumar, "Cancer Detection and Analysis Using Machine Learning," 2022 Second International Conference on Computer Science, Engineering and Applications (ICCSEA), Gunupur, India, 2022, pp. 1-5, doi: 10.1109/ICCSEA54677.2022.9936457.
 - [7] S. Anklesaria, U. Maheshwari, R. Lele and P. Verma, "Breast Cancer Prediction using Optimized Machine Learning Classifiers and Data Balancing Techniques," 2022 6th International Conference On Computing, Communication, Control And Automation (ICCUBEA, Pune, India, 2022, pp. 1-7, doi: 10.1109/ICCUBEA54992.2022.10010783.
 - [8] L. Akter and N. Akhter, "Ovarian Cancer Classification from Pathophysiological Complications using Machine Learning Techniques," 2021 12th International Conference on Computing Communication and Networking Technologies (ICCCNT), Kharagpur, India, 2021, pp. 1-6, doi: 10.1109/ICCCNT51525.2021.9580067.
 - [9] A. Rehman, M. Kashif, I. Abunadi and N. Ayesha, "Lung Cancer Detection and Classification from Chest CT Scans Using Machine Learning Techniques," 2021 1st International Conference on Artificial Intelligence and Data Analytics (CAIDA), Riyadh, Saudi Arabia, 2021, pp. 101-104, doi: 10.1109/CAIDA51941.2021.9425269.
 - [10] M. del Valle, M. O. dos Santos, S. N. dos Santos, E. S. Bernardes and D. M. Zezell, "Evaluation of machine learning models for the classification of breast cancer hormone receptors using micro-FTIR images," 2021 SBFoton International Optics and Photonics Conference (SBFoton IOPC), Sao Carlos, Brazil, 2021, pp. 1-3, doi: 10.1109/SBFotonIOPC50774.2021.9461946.
 - [11] D. Moldovan, "Cervical Cancer Diagnosis Using a Chicken Swarm Optimization Based Machine Learning Method," 2020 International Conference on e-Health and Bioengineering (EHB), Iasi, Romania, 2020, pp. 1-4, doi: 10.1109/EHB50910.2020.9280215.
 - [12] Puneet and A. Chauhan, "Detection of Lung Cancer using Machine Learning Techniques Based on Routine Blood Indices," 2020 IEEE International Conference for Innovation in Technology (INOCON), Bangluru, India, 2020, pp. 1-6, doi: 10.1109/INOCON50539.2020.9298407.