Machine Learning Based Framework for Drug Prediction of Cancerous Genomic Profiles

Amit Maurya, Department of Computer Science & Engineering, RDEC, Ghaziabad
maurya.amit@gmail.com

Abstract

One of the growing approaches to clinical research and patient treatment is known as precision medicine. This method focuses on understanding and treating illness by combining multi-modal or multi-omics data from an individual in order to make choices that are suited to the specific needs of the patient. The huge and complicated datasets that were produced by the diagnostic methods used in precision medicine necessitated the development of innovative methodological approaches that could handle and comprehend this intricate data. While this is going on, the field of computer science has made tremendous strides in the development of methods that allow the storage, processing, and analysis of these complicated datasets. This is a feat that conventional statistics and early computing technologies were not able to achieve. In the field of computer science, machine learning is a process that seeks to uncover complicated patterns in data. These patterns may be used to generate predictions or classifications on fresh data that has not been seen before, or they can be used for sophisticated exploratory data analysis. Machine learning is a subfield of artificial intelligence. The use of machine learning to the study of multi-modal data in precision medicine enables the comprehensive examination of big datasets, which eventually leads to a deeper comprehension of human health and its associated diseases. The purpose of this study is to examine the use of machine learning to the "big data" of precision medicine, namely in the context of genetics, genomics, and other related fields.

Keywords:precision medicine,computer science, machine learning,big data

Introduction

Precision medicine, which is also frequently referred to as precision health, is an innovative method of understanding health and disease that is based on patient-individual data. This data includes medical diagnoses, clinical phenotype (such as the severity of the disease or the amount of functional impairment), biologic investigations (such as laboratory studies and imaging), as well as environmental, demographic, and lifestyle factors. When viewed as a whole, these data are referred to as multi-modal since they include information from a variety of different domains. The exponential growth in the amount of biologic data that can now be collected for each individual patient has had a significant impact on the development of precision medicine. This occurrence is largely attributable to the introduction of new technologies in the fields of medicine, genetics, metabolics, and imaging, amongst others. Because of the vast number and diversity of diagnostic tests that may be carried out, an enormous quantity of data is generated. This data is difficult to comprehend and analyse for a single patient, and it is considerably more complicated to do so for a dataset that contains information from numerous patients. Thankfully, while more advanced diagnostic tests were being created, the area of computer science also saw an evolution. This development made it possible to store and analyse these vast amounts of data in a more effective manner than have ever been possible before. The progress of precision medicine diagnoses and therapies is made possible by the use of computer science approaches that make use of the vast amounts of deep data generated by the health care system. These two advances go hand in hand with one another.

There is a lack of clarity on the beginnings of precision medicine, which may be attributed, in part, to the fact that the phrase has developed from time to time (Phillips 2020). However, one of the first fields to use a precision medicine approach to treat human disease was transfusion medicine, where the discovery of blood types in the early 1900s revolutionised blood transfusions, allowing for matching of donor and recipient blood types, and avoiding complications associated with mismatched donor and recipient blood (Dance 2016; Giangrande 2000; Hodson 2016). Since that time, precision medicine has seen significant development, which has resulted in the incorporation of innovative techniques to therapy, intervention, diagnosis, and prevention. These developments are all contributing to a shift in

the landscape of medicine. Many terminal diseases now have precision medicine treatments that are extending life and enhancing quality of life for patients. One example of this is gene therapy for newborns with spinal muscular atrophy (SMA) type I, which had a condition that was formerly fatal before the age of two. According to Singh et al. (2017), children with spinal muscular atrophy type I who have been treated with gene therapy are now surviving longer and experiencing far fewer major respiratory issues that need invasive breathing assistance. This information has had a profound impact on the lives of these patients and their families. Precision medicine is today a discipline that enjoys enormous support from research and clinical funding organisations, government administrations, and among the general public, including private contributors and politicians. This support is a result of the field's success and the promise it has for the future. The objective of this brief study is to explain how machine learning has the potential to become an indispensable instrument for the development of precision medicine in the future. In addition to discussing ethical and legal issues, the primary emphasis of this study will be on genetics and genomics, "big data," and the most cutting-edge applications of machine learning.The building blocks The building blocks of every known living entity are cells.

In comparison to humans, they are capable of a broad range of complex activities due to a distinct collection of genes. In every cell, there is a piece of genetic material called a gene that serves as both a structural and functional unit of inheritance. Dissimilarities in genes may account for the observed diversity in animal phenotype and genotype. An organism's phenotype is constructed according to the instructions contained in its genes. Since its inception, genetics has developed into its own branch of science. Patients with numerous hereditary disorders now have better treatment options and a longer life expectancy because to bioinformatics. The diagnosis of major diseases such as diabetes, cancer, and heart attacks has become much simpler in recent times. The healthcare industry is looking to chip technology for the future, as it has already provided lab-on-a-chip equipment. Using these chips, it may be possible to precisely evaluate patients' genetic profiles. New medical technologies are making it easier to detect and assess the prognosis of cancer and other potentially fatal diseases at an earlier stage [3]. Researchers in the field of genetics have uncovered the mechanisms by which certain characteristics are transmitted from parents to offspring. They are also investigating gene expression in an effort to determine the internal and external factors that affect the expression of certain genes. This gene expression data could be subject to a plethora of studies using statistical and computational approaches. Gene expression data may be utilised in combination with a multitude of other omic data, including copy number variations, gene mutations, and proteome, transcriptome, and genome. Gene expression data is crucial for many medical tasks, including diagnosis, therapeutic target discovery, drug pathway analysis, and disease classification. An accurate diagnosis and treatment of diseases like cancer may be possible in the future thanks to the persistent efforts of scientists and researchers who are trying to find the underlying processes [4, 5, 6, 7]. Data mining and machine learning methods are assisting data-driven analyses in this way.

Genetics, genomics, and precision medicine

Genetics and genomics have come a long way since 1869, when Friedrich Miescher found DNA, and since 1953, when Watson, Crick, and Franklin first described DNA. More recently, in the late 80s, researchers found specific mutations for colour vision and cystic fibrosis. Our current understanding of the genetic basis of health and disease is mind-boggling. Genomic research focuses on a person's complete genetic makeup and how it interacts with their environment, while genetics studies genes and their functions in inheritance; both fields were instrumental in bringing about the precision medicine revolution, and thus, genomics and genetics have occupied most of the early attention in the field. Grainger (2016) notes that genetics and genomics provide the bulk of the data used in precision medicine. This is also made possible by the dramatic reductions in both the time and money needed to perform genetic testing; in 2001, it took more than a decade and about $3 billion to sequence a genome, but today, the same task can be done in about 24 hours for

about $1000 (Hodson 2016). Without a doubt, genomics and genetics have been the driving forces behind precision medicine and will remain so going forward.

An individual gene's expression value may be calculated by comparing its levels in two different settings using DNA microarray hybridization. Gene expression, which involves reading instructions from the genome, aids protein production. An indicator of gene expression is the amount of messenger ribonucleic acid (mRNA) that a gene generates during a certain time period. Internal and external stimuli, together with the presence or absence of biological regulators and pathways, may alter the values given to genes during expression. Messenger RNA (mRNA) is a molecule that helps transmit the genetic instructions needed to make proteins. Transcription and translation are the two subprocesses involved here. Thetranscription happens when RNA is converted into a transcript. A messenger RNA (mRNA) molecule is a copy of genetic material that must travel from the nucleus to the cytoplasm before it can direct protein synthesis. Protein synthesis begins with the translation of a molecule of messenger RNA (mRNA) containing the sequence of amino acids. The genetic code explains how the sequence of base pairs in a gene becomes a particular sequence of amino acids. Within the cell's cytoplasm, the ribosome reads the messenger RNA sequence in three-base blocks at a time, allowing it to create the protein. Data on gene expression is useful for a wide variety of scientific research. By making the connection between genotype and phenotypic traits more clear, it aids in distinguishing between different phenotypic articulations. We use it to classify disease characteristics and find potential biomarkers for diseases. Machine learning models get these data from genomic studies like m-RNA, DNase-seq, and MNase-seq. By tapping into this potential, scientists have discovered a plethora of new information on cancer and other long-term diseases. There are several varieties of cancer, which makes it a complicated sickness. There is an immediate need for systems or techniques that may help with cancer type prognosis and early identification. Over the last decade, a number of new approaches to cancer research and treatment have surfaced [13]. The literature[14,15] proposes a number of computational and biological approaches to early cancer diagnosis. In addition to searching for novel biomarkers, researchers are attempting to determine how to predict the effects of various medications on various illnesses and targets using computational (in-silico) models and algorithms. More research is being poured into this area because to the proliferation of large cancer data banks. We can now use machine learning to forecast the tumor's malignant potential.

Cancer Classification

For decades, gene expression patterns have been used for the purpose of interpreting biological significance and associating genes with disorders. These profiles are derived from a diverse group of patients and are gathered from several different biological contexts. To better understand the disease, it may be helpful to compare expression patterns in normal and malignant tissues. The quantity of messenger RNA (mRNA) produced by a gene at any one moment is a measure of its expression, or its state (active or inactive). The advancement of biological computational tools has spurred further research into cancer classification and analysis of microarray data. The correct subtyping and categorization of tumour samples is crucial for cancer diagnosis and prognosis. The precise classification of cancer types and the subsequent discovery of subtype-specific treatment modalities are both aided by this. Several writers have created classification algorithms that use gene expression data as their basis [11]. Statistical methods and machine learning systems are two examples of the many cancer categorization options available. Classification is a difficult problem due to the large complexity of gene expression data, and the majority of classifiers start with a genes selection phase [12]. Through the removal of unnecessary attributes, it helps to simplify the classification process in terms of both accuracy and time. A classifier built with only one feature selection method on one dataset may not work so well when used on other datasets since current "feature selection algorithms" can't scale or generalise well enough. Deep Neural Networks (DNN) [13] might be useful for automatic feature extraction and building general, scalable classifiers. Innovations in DNA microarray technology have revolutionised

research in the life sciences. At the same time, scientists may study hundreds of genes to determine their function and significance in the body. Furthermore, microarray technology enables the simultaneous examination of genomic profiles, which provides important information on a wide variety of genetic variations and alterations. Cancer and other life-threatening diseases may be detected earlier with its help. Over the last 20 years, several researchers have contributed standard microarray datasets for various tumour types, which has greatly aided cancer research. Databases for cancer offer thousands of genes extracted from many types of samples. Several approaches to cancer classification using genetic profiles make use of these data sets as benchmarks. Various computational approaches have been developed for the purpose of cancer classification.

The "omics" revolution

But precision medicine is quickly expanding beyond genomics and genetic data to include data from other domains (Peck 2018). Data from several other omics sources, such as microbiome studies, epigenetics, proteins, metabolism, radiology, pharmacology, and environmental omics, have been incorporated into the field as a whole as a result of methodological advancements. This is why it's not uncommon to hear the term "multi-omics" used to describe the incorporation of data from several fields. These deep phenotyping datasets are notoriously difficult, if not impossible, to analyse without the assistance of data science and ever-increasing computing power and technology. Big data sets containing information from a variety of sources necessitate new approaches to data processing, understanding, and utilisation. As a result, machine learning became a prominent tool in the industry. A subfield of AI known as "machine learning" entails teaching computers to analyse and interpret data in order to draw conclusions about new instances based on previously discovered complicated patterns. Already, machine learning has changed the way we live a lot, and it will continue to play a crucial role in precision medicine in the years to come.

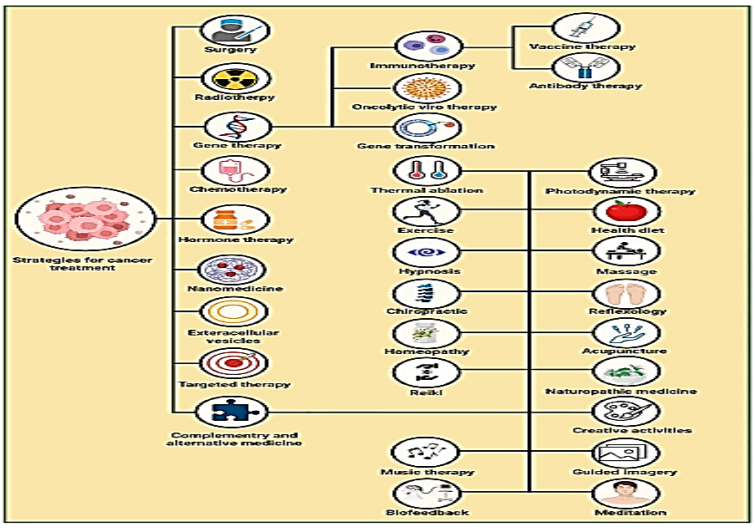

FIGURE 1. Cancer treatment approaches

Drug Response Prediction

Cancer is a genetic disease because it develops from changes and mutations in the genes that code for tumour cells. Because they change the genes that are responsible for many different cellular activities, mutations in genes directly affect how cells work. Contact with a harmful environment that encourages cancer growth is the most common cause of mutations. Because the tumour microenvironment is so intricate, cancer is a difficult disease to cure. Even while a targeted treatment is effective against all types of cancer, various patients react differently to it. The different reactions seen in individuals may be explained by their individual genetic compositions. The tumor's precise location cannot be used to make optimal treatment decisions for cancer. Precision medicine is an approach to cancer treatment that aims to provide personalised therapy by considering each patient's unique genetic profile in an effort to reduce or halt the disease's progression [14]. Finding the optimal treatment for each kind of cancer is challenging, but researchers are making progress. Numerous large-scale high-throughput pharmacological tests have shown a link between patient genes and therapy

efficacy. Data on the reactions of several human cancer cell lines to different medications is part of the pharmacogenomics databases that are created from these screens. Some examples of such extensive databases that aim to advance cancer research include the Cancer Cell Line Encyclopaedia (CCLE) [16] and the Genomics of Drug Sensitivity in Cancer (GDSC) [15]. For the purpose of drug (responses/combinations/repositioning) forecasting, these data sets are vital to modern drug development. In order to build reliable prediction models from this large screening datasets, computational methods must be refined. One of the most important issues is predicting which medications will work against a certain cell line by analysing the association between preexisting cancer genetic profiles and treatment responses.

Conclusion

Using information gathered from a variety of sources, precision medicine is reshaping the way we study and treat human illness. The analysis and processing of multi-omics data, which enables the categorization and prediction of outcomes for individuals and groups, relies heavily on machine learning approaches, such as the developing deep learning models. Precision medicine has relied heavily on genetics and genomics, which lend themselves well to machine learning; this data, along with other evaluations, will be crucial to the ongoing development of these areas. We have come a long way in our knowledge of human health and illness thanks to precision medicine and machine learning, and both fields offer enormous promise for the future of mankind.

References

[1]H. Chen, Y. Zhang, and I. Gutman, "A kernel-based clustering method for gene selection with gene expressiondata," Journal of Biomedical Informatics, vol. 62, pp. 12–20, 2016

[2]L.-J. Zhang, Z.-J.Li, and H.-W. Chen, "An effective gene selection method based on relevance analysis and discernibility matrix," in Proceedings of Pacific-Asia Conference on Knowledge Discovery and Data Mining, pp. 1088–1095, Springer, 2007

[3]G. Ji, Z. Yang, and W. You, "Pls-based gene selection and identification of tumor-specific genes," IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews), vol. 41, no. 6,pp. 830–841, 2011.

[4]D. K. Slonim, P. Tamayo, J. P. Mesirov, T. R. Golub, and E. S. Lander, "Class prediction and discovery using gene expression data," in Proceedings of the fourth annual international conference on Computational molecular biology, pp. 263–272, ACM, 2000.

[5]S. L. Pomeroy, P. Tamayo, M. Gaasenbeek, L. M. Sturla, M. Angelo, M. E. McLaughlin, J. Y. Kim, L. C. Goumnerova, P. M. Black, and C. Lau, "Prediction of central nervous system embryonal tumour outcome based on gene expression," Nature, vol. 415, no. 6870, pp. 436–442, 2002.

[6]D. Beer, S. Kardia, C. Huang, A. Gautam, Z. Li, and G. Bepler, "Ten best readings," Group, vol. 343, pp. 1217–1222, 2000.

[7] J. Khan, J. S. Wei, M. Ringner, L. H. Saal, M. Ladanyi, F. Westermann, F. Berthold, M. Schwab, C. R. Antonescu, and C. Peterson, "Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks," Nature Medicine, vol. 7, no. 6, pp. 673–679, 2001.

[8]S. A. Armstrong, J. E. Staunton, L. B. Silverman, R. Pieters, M. L. den Boer, M. D. Minden, S. E. Sallan, E. S. Lander, T. R. Golub, and S. J. Korsmeyer, "Mll translocations specify a distinct gene expression profile that distinguishes a unique leukemia," Nature Genetics, vol. 30, no. 1, pp. 41–47, 2001.

[9]K. E. Lee, N.Sha, E. R. Dougherty, M. Vannucci, and B. K. Mallick, "Gene selection: a bayesian variable selection approach," Bioinformatics, vol. 19, no. 1, pp. 90–97, 2003

[10]K. L. Tang, W. j. Yao, T. H. Li, Y. x. Li, and Z. W. Cao, "Cancer classification from the gene expression profiles by discriminant kernel-pls," Journal of Bioinformatics and Computational Biology, vol. 8, pp. 147–160, 2010.

[11]D. Wang, A. Khosla, R. Gargeya,H. Irshad, and A. H. Beck, "Deep learning for identifying metastatic breast cancer," arXiv preprint arXiv:1606.05718, pp. 1–6, 2016.

[12]H. Liao, "A deep learning approach to universal skin disease classification," University of Rochester Department of Computer Science, CSC, 2016.
[13]R. Fakoor, F. Ladhak, A. Nazi, and M. Huber, "Using deep learning to enhance cancer diagnosis and classification," in Proceedings of the International Conference on Machine Learning, vol. 28, pp. 1–7, ACM New York, USA, 2013.
[14]A. Chinnaswamy and R. Srinivasan, "Hybrid feature selection using correlation coefficient and particle swarm optimization on microarray gene expression data," in Innovations in Bio-Inspired Computing and Applications, vol. 424, pp. 229–239, Springer, 2016.
[15]S. S. Shreem, S. Abdullah, and M. Z. A. Nazri, "Hybridising harmony search with a markov blanket for gene selection problems," Information Sciences, vol. 258, pp. 108–121, 2014
[16]L.-Y. Chuang, C.-H.Yang, J.-C.Li, and C.-H. Yang, "A hybrid bpso-cga approach for gene selection and classification of microarray data," Journal of Computational Biology, vol. 19, no. 1, pp. 68–82, 2012.
[17]F. V. Sharbaf, S. Mosafer, and M. H. Moattar, "A hybrid gene selection approach for microarray data classification using cellular learning automata and ant colony optimization," Genomics, vol. 107, no. 6, pp. 231–238, 2016.
[18]G. R. Zimmermann, J. Lehar, and C. T. Keith, "Multi-target therapeutics: when the whole is greater than the sum of the parts," Drug Discovery Today, vol. 12, no. 1-2, pp. 34–42, 2007
[19]L. Huang, F. Li, J. Sheng, X. Xia, J. Ma, M. Zhan, and S. T. Wong, "Drugcomboranker: drug combination discovery based on target network analysis," Bioinformatics, vol. 30, no. 12, pp. 228–236, 2014.
[20]A. Polynikis, S. Hogan, and M. di Bernardo, "Comparing different ode modelling approaches for gene regulatory networks," Journal of Theoretical Biology, vol. 261, no. 4, pp. 511–530, 2009.