Acta Scientific DENTAL SCIENCES (ISSN: 2581-4893)

Volume 4 Issue 11 November 2020

Research Article

Machine Learning in Multidimensional Biomarker Design: A Milestone in Precision Medicine - A Systematic Review

Raajasiri Iyengar[1] and Gaurav[2]*

[1]*Final Year Undergraduate Student, NSVK Sri Venkateshwara Dental College and Hospital, Bangalore, Karnataka, India*

[2]*Consultant Oral Physician and Maxillofacial Radiologist, Assistant Professor, Department of Oral Medicine and Maxillofacial Radiology, NSVK Sri Venkateshwara Dental College and Hospital, Bangalore, Karnataka, India*

***Corresponding Author:** Raajasiri Iyengar, Final Year Undergraduate Student, NSVK Sri Venkateshwara Dental College and Hospital, Bangalore, Karnataka, India.

Received: September 08, 2020

Published: October 14, 2020

Abstract

Background: With the emerging era of precision medicine and high-throughput sequencing technologies, huge amount of 'omics' data has been gathered. Data interpretation remains a challenge due to the large and evolving magnitude of the dataset [3]. Precision medicine not only includes targeted therapeutics, but also necessitates 'precision diagnostics'. Rational design of multidimensional biomarkers is the keystone of precision diagnostics, wherein multiple biomarkers are cumulatively assessed using computational techniques to yield specific patterns [1]. Patterns thus obtained are analyzed with Machine learning algorithms to arrive at a diagnosis that is both reliable and accurate.

Aim of the Study: To determine the role of Machine Learning in the rational design of multidimensional biomarkers.

Research Question: Can Machine Learning enable rational design of multidimensional biomarkers which serve as molecular signatures for specific disease states?

Materials and Methods: With the Medline database and Cochrane Collaboration taken as a source for authenticated scientific research data, 45 articles were selected having undergone randomized control trial. Out of these, 14 articles (studies) were chosen which met the criteria for systematic review.

Results and Conclusion: Machine learning enables identification of molecular signatures of specific disease states, by cumulative interpretation of multiple biomarkers simultaneously.

ML Algorithms can discover higher-order interactions among biomarkers, greatly improving the diagnostic performance.

Keywords: Multidimensional Biomarkers; Machine Learning; Rational Design; Precision Medicine

Introduction

Beginning with the completion of a milestone in our genome sequencing endeavor, the Human Genome Project (HGP), we have now evolved to complex sequencing technologies known as high-throughput sequencing or next-generation sequencing (NGS). As a result of this dynamic evolution, we now have enormous genomic datasets have been mined to identify both drug targets and biomarkers [3]. Acquisition of this data has become relatively easier, but interpreting this data efficiently still remains far-fetched. Data comprehensibility is the need of the hour in this era of 'precision medicine'.

The twenty first century healthcare is dominated by the concept of tailor-made therapeutic approaches that cater to individual healthcare needs. And, biomarkers are serving as essential diagnostic aids in this era of precision medicine. What are biomarkers? The FDI-NIH Working group on biomarkers has defined biomarkers as "substances used to detect or confirm the presence of a disease or condition of interest, or to identify individuals with a subtype of the disease. Individual biomarkers have since long, been used for the diagnosis of particular disease states. But, human diseases are phenotypically heterogeneous. This implies that each disease has multiple underlying genomic and proteomic changes, which cannot be pin-pointed using a single biomarker. Diagnoses obtained using a single biomarker remains unreliable [1]. This necessitates the use of advanced computational methods and Machine Learning algorithms to identify 'multidimensional biomarkers' [2]. Multidimensional biomarkers unlike single-analyte biomarkers are a group of biomarkers associated with a specific disease state that cumulatively form patterns unique to that particular disease state, in other words forming pathognomonic biomarker patterns. Machine Learning and advanced computational techniques enable identification and rational design of such multidimensional biomarkers. This greatly improves the accuracy and diagnostic performance of microchip-based diagnostic devices, point-of-care devices etc. [4].

Single-analyte vs. multidimensional biomarkers

Single-analyte biomarkers (for example, gene fusions or mutations) limit biomarker analysis, and set out to predict patient response based on merely a single facet of biology, attempting to oversimplify diagnostic techniques. As combination therapies in precision medicine, have evolved to target multiple diseases using a single therapeutic modality, biomarkers must as well seek to capture maximum possible information from data obtained through molecular profiling. By harnessing the potential of machine-learning tools, it is possible to filter out large amounts of noise within datasets and identify only the most useful data. Using machine-learning algorithms and models, high-accuracy predictive biomarkers called 'multidimensional biomarkers' can be identified and built by employing rational, data-driven methods. Hence, multidimensional biomarkers are superior to single-analyte biomarkers in all aspects [12] (Figure 1).

What is machine learning?

Machine Learning encompasses a set of computational techniques employed to reduce a large number of measurements into smaller dimensional outputs. With the dawn of the 'Omics' era and Next generation Sequencing (NGS), there is an ever expanding biological database. Machine learning enables clustering and stratification of this data, in order to reduce the data to manageable sizes and permit easy analysis.

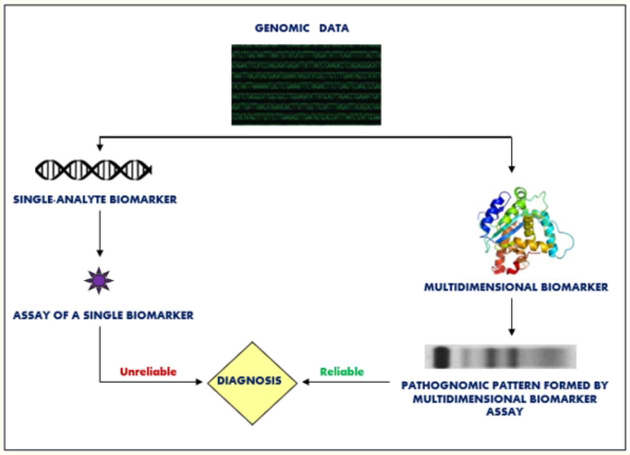
GENOMIC DATA
SINGLE-ANALYTE BIOMARKER
ASSAY OF A SINGLE BIOMARKER
Unreliable
DIAGNOSIS
Reliable
MULTIDIMENSIONAL BIOMARKER
PATHOGNOMIC PATTERN FORMED BY MULTIDIMENSIONAL BIOMARKER ASSAY

Figure 1: Single-analyte biomarker vs multidimensional biomarker - schematic representation.

Based on the data set used machine learning has been categorized into three major types:

- Supervised learning
- Unsupervised learning
- Reinforcement learning.

Supervised learning

Supervised learning applications are most relevant to diagnostic devices. How does supervised learning work? In supervised learning algorithms, the machine learning algorithm is first provided a set of labeled data called as - 'training data'. It is subsequently tested for its Diagnostic performance by providing a set of unlabelled data known as - 'test data'. The train data is the input given to the machine learning algorithm and it is categorized as either healthy or diseased. Based on the training data provided generate a deep model. Based on this deep model generated, the algorithm interprets the test data provided stratified i.e. labeled the test data as either healthy or diseased. It is to be noted that larger the size of the training data set provided to the algorithm, better is its diagnostic performance.

Supervised ML algorithms best address 'stratification/categorization problems'. Biomarker patterns aid in this stratification i.e.

individual sample units with a specific pattern of a given set of biomarkers are categorized as 'diseased', while those units without the defined biomarker pattern are categorized as 'healthy'.

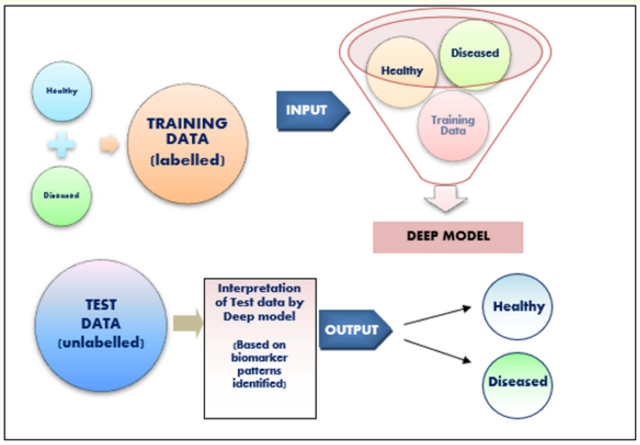

Figure 2: Schematic representation of how supervised learning algorithms work.

Unsupervised learning

In case of unsupervised machine learning algorithms, a set of input data is given, which is heterogeneous in its nature and structure. The algorithm aids in homogenizing such data by a process of clustering and reducing data dimensionality. Hence, such algorithms are best suited for 'clustering problems'.

Reinforcement learning

Reinforcement based ML algorithms on the other hand function on a trial and error basis with the predetermined reward to come up with a solution to a given problem. The algorithm maximizes the reward by starting with random trials and multiplying the number of trials sufficiently to obtain the best outcome.

Materials and Methods

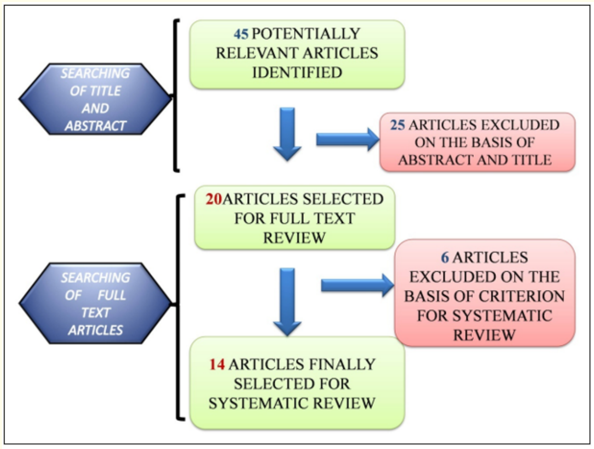
SEARCHING OF TITLE AND ABSTRACT
45 POTENTIALLY RELEVANT ARTICLES IDENTIFIED
25 ARTICLES EXCLUDED ON THE BASIS OF ABSTRACT AND TITLE
SEARCHING OF FULL TEXT ARTICLES
20ARTICLES SELECTED FOR FULL TEXT REVIEW
6 ARTICLES EXCLUDED ON THE BASIS OF CRITERION FOR SYSTEMATIC REVIEW
14 ARTICLES FINALLY SELECTED FOR SYSTEMATIC REVIEW

Figure a

A literature-based systematic review was carried out to fulfill the aim of this study. With the Medline database and Cochrane Collaboration taken as a source for authenticated scientific research data, 45 articles were selected having undergone randomized control trial. Out of these, 14 articles (studies) were chosen which met the criteria for systematic review.

Selection criteria (Inclusion and Exclusion Criteria)

Exclusion criteria

Articles which were not published in English and where the full text could not be obtained were excluded. Non-case control articles were also not included in the study.

Inclusion criteria

45 Articles published within a period of 5 years (2015 - 2020) were checked for validity, eligibility and design of the study, 14 of which were included in this systematic review based on their statistical significance.

Results

The need for this systematic review was laid on the foundation of following major conclusions which were drawn from the final 14 studies which were selected:

- Machine learning enables identification of molecular signatures of specific disease states, by cumulative interpretation of multiple biomarkers simultaneously.
- ML Algorithms can discover higher-order interactions among biomarkers, greatly improving the diagnostic performance.

Discussion

Applications of machine learning in healthcare

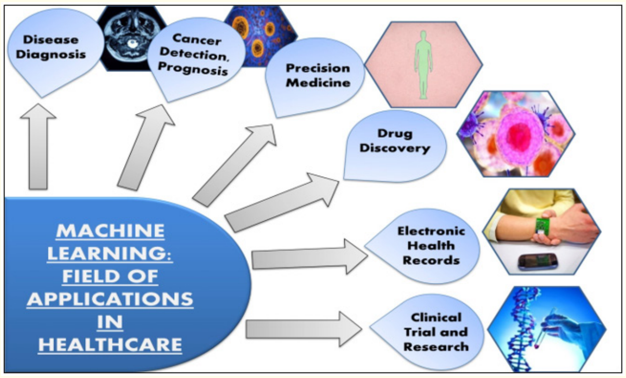
Disease Diagnosis
Cancer Detection, Prognosis
Precision Medicine
Drug Discovery
Electronic Health Records
Clinical Trial and Research
MACHINE LEARNING: FIELD OF APPLICATIONS IN HEALTHCARE

Figure 3: Machine learning in healthcare – field of applications.

Machine learning has a wide panorama of applications in healthcare including:

- Disease diagnosis
- Cancer detection

- Prognostic implications
- Precision medicine
- Drug discovery
- Electronic health records
- Clinical trials and research.

Disease diagnosis

Machine vision and other machine learning technologies can greatly enhance the efforts of pathologists and radiologists in identifying minor discrepancies. Facial recognition software is combined with machine learning aid in the diagnosis of rare diseases. Patient photographs are analyzed using facial analysis and deep learning to detect phenotypes that correlate with rare genetic diseases.

Cancer detection

Machine learning algorithms and deep learning have been employed to recognize dysplastic and neoplastic changes at the tissue level comparable to trained physicians [11]. ML algorithms have been successfully used to identify neoplastic histopathological specimens, and detect various cancers including head and neck cancers(HNCs),breast cancer, small cell carcinoma of the lung and colorectal cancer etc.

Prognostic implications

ML algorithms have a profound role in cancer prognosis and risk stratification which directly influences the treatment planning [13-15]. ML algorithms have been used to predict mean survival rate, level of risk in various cancers and potentially malignant disorders.

Precision medicine

ML has been instrumental in changing the face of precision medicine as states in various research articles and evidences in the literature:

- "AI techniques have been applied in cardiovascular medicine to explore novel genotypes and phenotypes in existing diseases, improve the quality of patient care, enable cost-effectiveness, and reduce readmission and mortality rates. Over the past decade, several machine-learning techniques have been used for cardiovascular disease diagnosis and prediction" [19].
- "Referring to huge amounts of epigenetic data coming from biological experiments and clinic, machine learning can help in detecting epigenetic features in genome, finding correlations between phenotypes and modifications in histone or genes, accelerating the screen of lead compounds targeting epigenetics diseases and many other aspects around the study on epigenetics, which consequently realizes the hope of precision medicine" [20].
- "Machine learning approaches for clinical psychology and psychiatry explicitly focus on learning statistical functions from multidimensional data sets to make generalizable predictions about individuals. The goal of this review is to provide an accessible understanding of why this approach is important for future practice given its potential to augment decisions associated with the diagnosis, prognosis, and treatment of people suffering from mental illness using clinical and biological data" [21].

Drug discovery

ML algorithms largely aid in simplifying the understanding of drug-target interactions through computational techniques and thus facilitate rapid advancements in drug discovery. "Machine learning (ML) approaches provide a set of tools that can improve discovery and decision making for well-specified questions with abundant, high-quality data. Opportunities to apply ML occur in all stages of drug discovery. Examples include target validation, identification of prognostic biomarkers and analysis of digital pathology data in clinical trials. Applications have ranged in context and methodology, with some approaches yielding accurate predictions and insights" [22,23].

Electronic health records

Data from Electronic Health Records can be utilized for predicting various imperative situations in a healthcare setting such as the risk of emergency admission [24], Intensive care unit readmission [25], post-partum depression [26] etc.

Clinical trials and research

ML has the greatest role in fast-tracking research in various branches of medical sciences enabling us to device therapeutics that would have otherwise reached us centuries later. ML algorithms can also aid in identifying clinically relevant evidences in the literature, saving the time spent in manual literature mining.

ML Algorithms in common use

Commonly used machine learning algorithms include:

- R packages (e.g. caret, randomForest, e1071, rpart, glmnet)

- Python libraries (e.g. TensorFlow, scikit-learn, Theano, Pylearn2, Pyevolve), and
- MATLAB statistics
- Machine leaning toolbox (e.g. SVM, KNN, PCA, Ensemble, Decision trees)
- Naïve Bayes model

Challenges with ML

The three challenges with using ML based technologies include:

- **Quality, structure, and amount of the training data:** Large amounts of training datasets are essential to achieve clinically and statistically significant accuracy in ML based diagnostics. Smaller datasets correspond to poor diagnostic performance.
- **Training on artifacts of sample collection or processing:** It is imperative to follow standardized methods of sample collection and processing in order to avoid bias and consequent attainment of erroneous results and conclusions.
- **Time and expense involved in sampling:** As large training datasets are required for acceptable levels of diagnostic accuracy in ML based diagnostic devices, sampling becomes a cumbersome, expensive and tie consuming procedure.

Conclusion

Machine learning has the ability to identify signatures of specific disease states in dynamically evolving datasets with constant addition of new molecular information. This has widespread implications in 'precision diagnostics', microchip-based diagnostics and point-of-care devices.

Using 'multidimensional biomarkers' that is by cumulative assessment of multiple molecular biomarkers, machine learning can improve drastically the diagnostic performance compared to manually chosen biomarker(s) using the same input dataset. Moreover, very often individual biomarkers do not have any significant predictive value independently, but the cumulative interpretation of many weak predictors can synergize to a strong correlation with specific disease states. Machine learning algorithms simultaneously evaluate effects of many biomarkers and can discover higher-order interactions among biomarkers, which is extremely tedious to carry out by conventional manual methods. This can open new possibilities for developing highly accurate diagnostic devices and act as a turning point in the timeline of precision medicine and diagnostics.

Bibliography

1. Ko J., *et al*. "Machine learning to detect signatures of disease in liquid biopsies - a user's guide". *Lab on a Chip* 18.3 (2018): 395-405.
2. P Zhang., *et al*. "Integrated biomedical data analysis utilizing various types of data for biomarkers identification". 2017 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), Kansas City, MO (2017): 1469-1475.
3. Shendure J and Aiden E. "The expanding scope of DNA sequencing". *Nature Biotechnology* 30 (2012): 1084-1094.
4. Goecks J., *et al*. "How Machine Learning Will Transform Biomedicine". *Cell* 181.1 (2020): 92-101.
5. Mamoshina P., *et al*. "Machine Learning on Human Muscle Transcriptomic Data for Biomarker Discovery and Tissue-Specific Drug Target Identification". *Frontiers in Genetics* 9 (2018): 242.
6. Kourou K., *et al*. Computational and Structural Biotechnology Journal 13 (2015): 8-17
7. Pratik Shah., *et al*. "Artificial intelligence and machine learning in clinical development: a translational perspective". *Digital Medicine* (2019).
8. ADAM Genomics Schema - Extension for Precision Medicine Research DH '18: Proceedings of the 2018 International Conference on Digital Health (2018)
9. Blatti C., *et al*. "Knowledge-guided analysis of "omics" data using the KnowEnG cloud platform". *PLoS Biology* (2020).
10. Beerenwinkel N., *et al*. "Computational Cancer Biology: An Evolutionary Perspective". *PLoS Computation Biology* 12.2 (2016): e1004717.
11. Litjens G., *et al*. "Deep learning as a tool for increased accuracy and efficiency of histopathological diagnosis". *Scientific Report* 6 (2016): 26286.
12. Mariani M., *et al*. "Integrated Multidimensional Analysis Is Required for Accurate Prognostic Biomarkers in Colorectal Cancer". *PLoS ONE* 9.7 (2014): e101065.
13. Kourou K., *et al*. "Machine learning applications in cancer prognosis and prediction". *Computational and Structural Biotechnology Journal* 13 (2014): 8-17.

14. Cruz JA and Wishart DS. "Applications of machine learning in cancer prediction and prognosis". *Cancer Information* 2 (2007): 59-77.

15. Bashiri A., *et al*. "Improving the Prediction of Survival in Cancer Patients by Using Machine Learning Techniques: Experience of Gene Expression Data: A Narrative Review". *Iranian Journal of Public Health* 46.2 (2017): 165-172.

16. Huang S., *et al*. "Artificial intelligence in cancer diagnosis and prognosis: Opportunities and challenges". *Cancer Letter* 471 (2020): 61-71.

17. Azencott CA. "Machine learning and genomics: precision medicine versus patient privacy". *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* 376.2128 (2018): 20170350.

18. Grapov D., *et al*. "Rise of Deep Learning for Genomic, Proteomic, and Metabolomic Data Integration in Precision Medicine". *OMICS* 22.10 (2018): 630-636.

19. Krittanawong C., *et al*. "Artificial Intelligence in Precision Cardiovascular Medicine". *Journal of the American College of Cardiology* 69.21 (2017): 2657-2664.

20. Fan S., *et al*. "Machine Learning Methods in Precision Medicine Targeting Epigenetic Diseases". *Current Pharmaceutical Design* 24.34 (2018): 3998-4006.

21. Dwyer DB., *et al*. "Machine Learning Approaches for Clinical Psychology and Psychiatry". *Annual Review of Clinical Psychology* 14 (2018): 91-118.

22. Vamathevan J., *et al*. "Applications of machine learning in drug discovery and development". *Nature Reviews Drug Discovery* 18.6 (2019): 463-477.

23. Lima AN., *et al*. "Use of machine learning approaches for novel drug discovery". *Expert Opinion on Drug Discovery* 11.3 (2016): 225-239.

24. Rahimian F., *et al*. "Predicting the risk of emergency admission with machine learning: Development and validation using linked electronic health records". *PLoS Medicine* 15.11 (2018): e1002695.

25. Rojas JC., *et al*. "Predicting Intensive Care Unit Readmission with Machine Learning Using Electronic Health Record Data". *Annals of the American Thoracic Society* 15.7 (2018): 846-853.

26. Wang S., *et al*. "Using Electronic Health Records and Machine Learning to Predict Postpartum Depression". *Studies in Health Technology and Informatics* 264 (2019): 888-892.