International Journal of Scientific Research in Computer Science, Engineering and Information Technology
© 2019 IJSRCSEIT | Volume 5 | Issue 1 | ISSN : 2456-3307
doi : https://doi.org/10.32628/CSEIT2063149

A Survey of Machine Learning models for the wide Spectrum of Computational Biology

Divya Ebenezer Nathaniel*, Sonia Panesar

Assistant Professor Computer Science Engineering, Babaria Institute of Technology, Gujarat Technology University, Gujarat, India
* Corresponding Author: divya.mtech2010@gmail.com, leosonia@gmail.com

ABSTRACT

With the Advent of advancement in the field of Artificial Intelligence the computer is made more intelligent and can enable to think and make prediction accurately. The machine learning being a subfield of Artificial Intelligence is used in numerous research works. Different analysts feel that enormous data generated in field of biology have to be sorted in an intelligent way to yield best model. There are numerous kinds of Machine Learning Techniques like Unsupervised, Semi Supervised, Supervised, Reinforcement, and Evolutionary Learning and Deep Learning. These learning's are used to classify huge data at a rapid pace. This paper discusses about the wide spectrum of Biology and the process of pre-processing data and the best suitable Machine learning model for each of them.

Keywords : Computational biology, genome, phenotype, Neural network, Clustering, Prediction.

I. INTRODUCTION

Data science and machine learning are quickly making inroads not only into all aspects of our life and businesses, but also in biomedical field. The term Biomedical consists of two terms Biology and medicine. It is a vast field which when combine with science or engineering give scope to various area of research. Due to advancement in technology, there are techniques for analysing and storing of large amount of data, Health-care systems generate a large amount of biomedical data including electronic health records, medical imaging, multi omics data, etc. In recent years, the data collected in biomedical fields, deeply analysed and provides a detailed understanding of the data which helps us to focus on advanced technology that improve human health at different levels

Machine learning has become a vital tool for many projects in computational biology, biomedical engineering and pharmacology. A machine learning model consists of computational method based upon statistics, implemented in software, able to learn hidden knowledgeable patterns in a dataset, and then make reliable statistical predictions about similar new data.

A survey of diversified spectrum of computational biology with possible machine learning models is reviewed in the paper. Following fields in computational biology are gaining popularity listed below

The Steps involved in the Machine learning models include

1. **Data Collection**: With the advent of internet era there are many online databases which provide a quality of dataset to train and evaluate the model in the vast field of computational biology the data set representation varies with the field and different data bases has to be referred for the same further quantity & quality of your data decides the accuracy of the model

2. **Data Preparation**: Mostly the data generated in field of computational biology belong to category of images or it can be a matrix so in order to train a machine learning model the data set has to be revenant, concise and free of noise. Split into training and evaluation sets

3. **Choose a Model**: Different algorithms are available in machine learning models and according to input data set and the accuracy of expected output the apt algorithm has to be selected the best algorithm is discussed under each section for different biomedical fields in the following paragraphs.

Disease Prediction: As the name suggests Machine learning model once trained with existing patient's dataset, predict viability of unseen patient data to acquire the disease in future. When a new patient visits the health care centre [6]

(a) Health care professional collects the patient's vital report as well as genome - a complete sequence of genes presents in cell which aids in the well-being of an individual. This collected data set is compared with a database for the process of identifying mutations and disease-causing genes, furthermore many

(b) Laboratory tests are conducted using tissue samples

(c) The patient's Habits and lifestyle information need to be processed.

The patient's vital information in all perspective like genome, behaviour and lifestyle, family history, are collected and stored in systematised databases of biomedical knowledge. Finally, a machine learning algorithm is devised in such a way to forecast the probability of the patient chance to develop a particular disease in near future. In order to make accurate prediction, the machine learning models are trained with multiple patients' data of various characteristics. For disease prediction, supervised learning model is recommended to classify whether the patients will fall in to the category of "Disease" or "No Disease" Strata.

Support Vector

The support vector algorithm is a supervised machine algorithm [14] where the data items are plotted as points on n dimension and those points are grouped on a dimensional plane. The grouping is done by creating a hyper plane that separates the groups with a margin that is as wide as possible. This helps with the classification. the support vector machine also tends to separate the non linear data using kernel function and can provide even a multidimensional view for the given problem

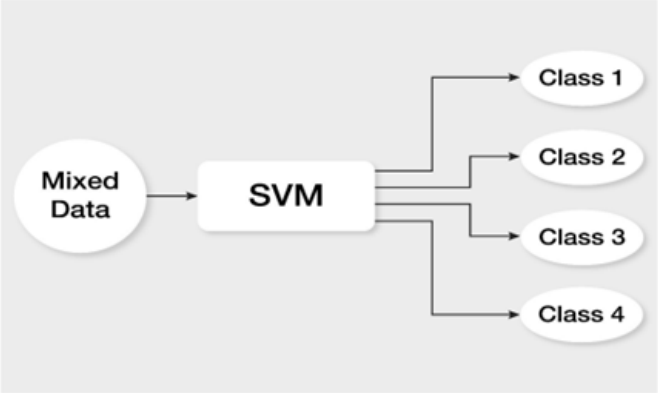

Figure-I SVM

Gene Expression Analysis (Epi genetic): All living beings are made up of cells- the atomic unit of a biological life and each cell holds the key component of life called the gene which embody the DNA. The Gene is responsible for many phenotypic characteristics of an organism. Such as colour, intelligence, immunity, weight, height etc. The study

of epigenetic [1] –means by altering the sequence of DNA many absurd characteristics can be ruled out or it can be suppressed from exhibition. DNA is made up chemical compound called as nucleotides and each nucleotide follows the structure which has a phosphate group, a sugar group and a nitrogen base. The nitrogen bases are of four types and they are adenine (A), thymine (T), guanine (G) and cytosine (C).The DNA's instructions or genetic code are determined by the sequence of the bases. The study of various genes inside the DNA is called genomics [3,4,5] the information embedded in the gene are transferred to the cells through the translation processes occurs in RNA and the RNA synthesis the protein responsible for metabolism in the cellular activities the study of generation and synthesis of protein inside the cell is called proteomics. In wet labs the experiments conducted by adding some chemical compound such as methyl group, phosphorous, acetylene. One example of an epigenetic change is DNA methylation [10] is the process of the addition of a methyl group, or a "chemical cap," to part of the DNA molecule, which prevents certain genes from being expressed. Another example is histone modification, phosphorylation, acetylation, by alternation in the sequence of DNA the undesirable Phenotypes can be suppressed or its occurrence can be nullified, this extensive process includes a large set with data and carrying out the task using machine learning algorithm will give good outcome

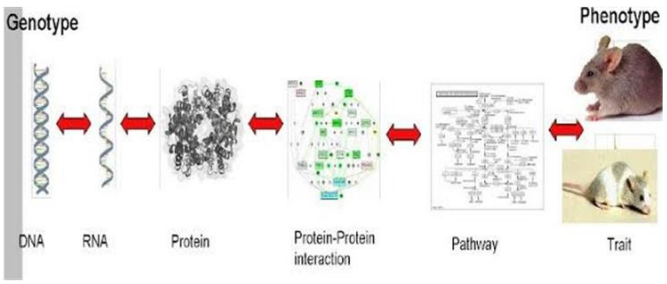

Figure-II Gene Expression Analysis

Neural Network

There are various classification algorithms but artificial neural network gives competitive result they consist of different layers for analysing and learning data. Every hidden layer tries to detect patterns. Once a pattern is detected the following hidden layer is activated and so on. Depending on the number of layers, it will be able to define prediction the more layers in a neural network, the more is learned and the more accurate the pattern are detected. Neural Networks learn and attribute weights to the connections between the different neurons each time the network processes data

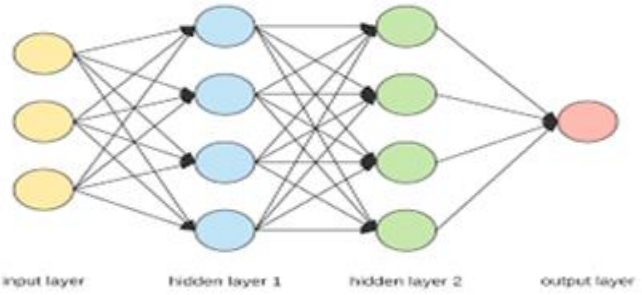

Figure -III Neural Networks

Protein-protein interaction prediction: proteomics is to analyse networks of physical interactions between proteins.

The protein – protein interaction also termed as PPI influences many phenotype characteristic of an organism [7]. Proteomics-based approaches, which investigate proteins, present inside the any tissue or cell and it complement the genomes that are widely used to address biomedical questions. Proteins are the main functional output, and the genetic code cannot always indicate which proteins are expressed, in what quantity, and in what form. For example the environmental factors or multigenic processes such as ageing or disease cannot be determined only by evaluating the genome

This illustrates the dominance of proteomics in several areas of biomedical research, ranging from pathogenesis of neurological disorders to drug and vaccine design.

The recent headway in field of pharmacy, significant number of PPIs are identified and number keeps surging. To combat this situation and to get a complete knowledge of PPIs and their characterisation at the network level. Protein sequence and structural information, which is the core of PPIs, the computational method to analyse PPI has become widely popular. The suitable algorithm to study the different types of proteins and their similarity the affinity to the drugs can be modelled using

Clustering algorithm: is placed under the category of unsupervised learning method. An unsupervised learning method is a method in which the references are drawn from data sets consisting of input data without labelled responses. Clustering **[14]**is the activity of dividing the population or data points into a number of groups such that data points in the same groups are more alike to each other and dissimilar to the data points in other groups. It is inherently a procedure to create a group of data points with similar features.

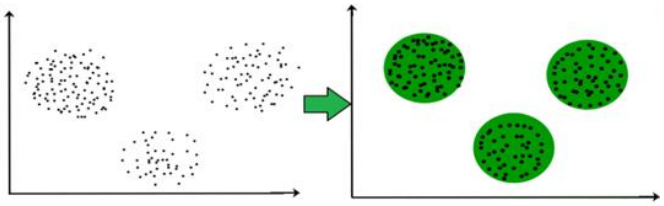

Figure-IV Clustering

Computational pharmacology

The main focus of computational pharmacology is to use the collected data for drug prediction, for better understanding of how drugs affect the human body, support decision making in the drug discovery process, improve clinical practice and avoid unwanted side effects [9].The drug characteristics and their reaction inside the human body are depicted in many ways and measured at the physicochemical, pharmacological, and phenotype levels. The affinity or the attraction between a drug and its target proteins are measured by following parameters such as the change in a cellular state or gene expression, binding strength, kinetic activity. Furthermore, the information such as drugs used in a treatment of diseases, its side effects, and its reaction with other combinational drugs can be represented in a mathematical model which are then analysed to aid drug discovery. ML explores the relationship and discovers the hidden knowledge that can be used for better result.

Drug-target interaction prediction (DTI)

Drug-target interactions are the prime process for both new drug discovery and old drug repurposing. With the constrain of time and money factors the limited number of DTI are uncovered in wet-lab

The huge gap between known and unknown drug-target pairs has compelled interest in DTI prediction. With the recent evolution, the computational model can more efficiently predict potential interaction candidates. And furthermore, DTI is the interaction of drug with the protein available in the cell of a human body for example, if the administered drug 95% binds to the protein and the 5% free drug which remain unbound to any protein inside the cell may cause some pharmacological side effects.

Machine learning is a cost effective technology for drug-target interaction prediction.

Ligand-based methods [13] which describe the nature of liaison between the like molecules with the like protein, this is one of the Quantitative Structure Activity Relationship (QSAR) Model. Precisely, in these methods prediction are carried out by understanding interactions of a new ligand to known proteins ligands. However, in a scenario where the number of known ligand is less then ligand-based methods behave poorly.

The docking simulation methods [14] requires the three-dimensional (3D) structures of proteins for simulation and becomes inapplicable when the 3D structure of proteins is unavailable

Chemogenomic approaches [15] play a vital role in drug discovery and drug repositioning. The major

criteria involved in DTI prediction are gene, protein, disease, and side effect. The drug-target pair prediction, works by invoking the methods which coalesce the chemical space of elements and the genomic space of target proteins into a combined space namely pharmacological space. An ample amount of biological data is fully utilized by Chemogenomic approaches for better prediction.

The chemogenomic approaches can be stratified into the following methods such as machine learning based methods, network-based methods and graph-based methods [16] the accurate prediction of machine learning-based methods paved way for its vouge.

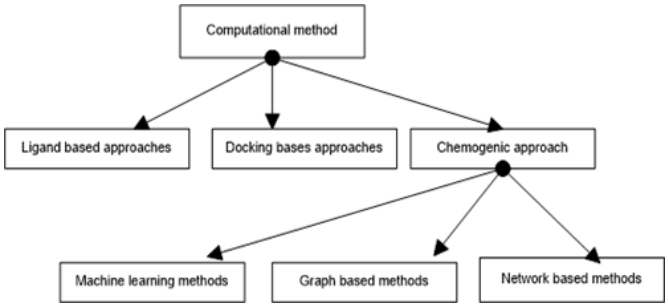

Figure V Chemogenomic approaches

Drug-drug interaction and drug combination prediction

Many patients are prescribed multiple drugs at the same time to treat complex diseases or co-existing conditions [10]. A drug combination consists of multiple drugs, each of which is used to treat a specific symptom in the patient population the different drugs in a drug combination can kindle the activity of distinct proteins, even though the drug combinations can improve the therapeutic efficiency, but the major consequence of a drug combination for a patient is higher risk of side effects. It is practically infeasible to test all possible pair of drugs and its side effect in a clinical evaluation. For example, given n drugs, there are $n\ (n - 1)\ /2$ pair wise drug combinations and many higher-order combinations. To address this combinatorial explosion of candidate drug combinations, computational methods were adopted to identify drug pairs that potentially interact. The drug –drug interaction is predicted by estimating the scores which render the overall strength of interacting drug pair

The prevalent approaches such as classification- or similarity-based can do well. The Classification-based approaches handle drug-drug interaction prediction as a binary classification problem. The method first extracts the feature representation of each drug pair by dimensionality reduction algorithm then feature vectors of individual drugs are grouped together to form integrated feature vectors of drug pairs.

The Machine learning models such as logistic regression classifier, support vector machine, or neural network are trained on the reduced data set for the feature representations of drug pairs.

Drug repurposing

Drug repurposing alias "drug re positioning" adopts computational methods to discover different uses for existing drugs [12,13] the main objective is to speculate a drug that might treat any given disease methods such as network approaches [12,13], similarity-based methods, and matrix factorization are well known among drug repurposing

The drug repurposing has the following two observations. First, many drugs have multiple target proteins and hence a multi- target drug might be used for more than one purpose. Second, different diseases share genetic factors, molecular pathways, and symptoms and hence drug acting on such overlapping factors might be beneficial to more than one disease. Drug repurposing approaches can be categorized into four groups:

(1) On the basis of protein target interaction networks the method predict different uses for currently existing drugs

(2) Predictions are made by analysing gene expression which are activated by the various drug administration

(3) Methods that make predictions based on drug side effects,

(4) The similarity relationship between the drug and disease are explored to assemble drug-drug similarity network disease-disease similarity network and a

drug-disease interaction network,. Based on observation similar drugs to treat similar.

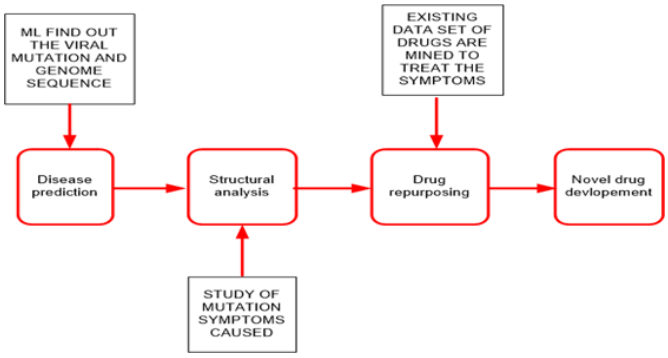

Figure VI- Drug repurposing

Disease is stated. Machine learning models perform well for drug-target interaction prediction. ML model uses the guilt-by-association principle, a principle that portrays like drugs tend to bind like target proteins and vice versa. Using this postulates prediction can be formulated as a binary classification task, which aims to predict whether a drug target interaction is present or not.The drug-target interactions as positive label, chemical structure of drugs and DNA sequence of target proteins as input features (or kernels) given to the classification models. DTI utilises graph neural networks to obtain deeper representations for drug–target activity prediction, based on structural information of both molecules and proteins.

Challenges in data integration for biology and medicine

The challenges in emerging machine learning approaches to integrate biomedical data are listed below.

1)Biological and medical datasets from the medical or biological source are complex and huge data set, incomplete, sporadic, multidimensional , distorted divergent, changing, and noisy.

2) Next important challenge exhibits from limitations of measurement technology [15], natural and physical constraints [11,15], and investigative biases. For example, information on which type of chemical compounds binds to which type of genes is available for only several thousands of genes, [15].

In the field of biomedical the data are stacked and they range from molecules, pathways, cells, tissues, organs, patients, and populations that enclose the large scale of species and their timelines. But For better understanding detailed representation which captures atomic details of molecule to characteristics of sample population is required.

3) The biomedical outcomes are inconsistent and they keeps transforming with time, so machine learning models synthesis the outcomes needs to be updated for active change . For example, bacteria, and viruses mutate promptly and they become drug resistance, to combat this situation the drugs have to progress, failure in adaptation and neglecting the dynamics of drug response can steer to poor performance in prediction of drug efficacy and toxicity.

4) A notable challenge in biomedical data science lies in formulation of new knowledge outside existing domain knowledge such as mapping a drug response from an animal model to a human patient.

Because of the partial datasets the model behaves poorly with the new data.

II. CONCLUSION

Machine learning models are inherently blended with the modern biomedical analysis, notably may approaches have been evolved to coalesce data from various source of biomedical dataset These procedures focus to connect the gap between our ability to generate vast amounts of data and our understanding of biomedical systems.

On-going systematic developments and emerging applications of machine learning promise an exciting future for biomedical data integration, though there is no single method that will perform best for all

problem. The methods are to be designed in such a way to handle different types of biomedical outcomes, domain-specific models, and specific types of data. In this Review various approaches are listed that can currently be executed to perform robust consolidated analyses.

III. REFERENCES

[1]. B. Linghu, E.S. Snitkin, Z. Hu, Y. Xia, C. DeLisi, Genome-wide prioritization of disease genes and identification of disease-disease associations from an integrated human functional linkage network, Genome Biol. 10 (9) (2009) R91.

[2]. ENCODE Project Consortium, An integrated encyclopedia of DNA elements in the human genome, Nature 489 (7414) (2012) 57.

[3]. A. Lundby, E.J. Rossin, A.B. Steffensen, M.R. Acha, C. Newton-Cheh, A. Pfeufer, S.N. Lynch, S.-P. Olesen, S. Brunak, P.T. Ellinor, et al., Annotation of loci from genome-wide association studies using tissue-specific quantitative interaction proteomics, Nat. Methods 11 (8) (2014) 868–874.

[4]. A. Kundaje, et al., Integrative analysis of 111 reference human epigenomes, Nature 518 (7539) (2015) 317–330.

[5]. M.D. Ritchie, E.R. Holzinger, R. Li, S.A. Pendergrass, D. Kim, Methods of integrating data to uncover genotype-phenotype interactions, Nat. Rev. Genet. 16 (2) (2015) 85–97.

[6]. M. Zitnik, B. Zupan, Data imputation in epistatic MAPs by network-guided matrix completion, J. Computation. Biol. 22 (6) (2015) 595–608.

[7]. M. Costanzo, B. VanderSluis, E.N. Koch, A. Baryshnikova, C. Pons, G. Tan, W. Wang, M. Usaj, J. Hanchard, S.D. Lee, et al., A global genetic interaction network maps a wiring diagram of cellular function, Science 353 (6306) (2016). aaf1420

[8]. Zeng, J.; Li, D.; Wu, Y.; Zou, Q.; Liu, X. An empirical study of features fusion techniques for protein-protein interaction prediction. Curr. Bioinform. 2016, 11, 4–12.

[9]. X. Li, J. Dunn, D. Salins, G. Zhou, W. Zhou, S.M.S.-F. Rose, D. Perelman, E. Colbert, R. Runge, S. Rego, et al., Digital health: tracking physiomes and activity using wearable biosensors reveals useful health-related information, PLoS Biol. 15 (1) (2017). e2001402

[10]. R.A. Hodos, et al., In silico methods for drug repurposing and pharmacology, Wiley Interdiscip. Rev. Syst. Biol. Med. 8 (3) (2016) 186–210.

[11]. N. Zong, H. Kim, V. Ngo, O. Harismendy, Deep mining heterogeneous networks of biomedical linked data to predict novel drug–target associations, Bioinformatics (2017). btx160

[12]. Luo, Y.; Zhao, X.; Zhou, J.; Yang, J.; Zhang, Y.; Kuang, W.; Peng, J.; Chen, L.; Zeng, J. A network integration approach for drug-target interaction prediction and computational drug repositioning from heterogeneous information. Nat. Commun. 2017, 8, 573. CrossRefPubMed]

[13]. Camacho, D.M.; Collins, K.M.; Powers, R.K.; Costello, J.C.; Collins, J.J. Next-generation machine learning for biological networks. Cell 2018, 173, 1581–1592. CrossRefPubMed]

[14]. Marinka Zitnika,∗ , Francis Nguyenb,c , Bo Wangd,Jure Leskoveca,e,∗, Anna Goldenbergf,g,h,MichaelM.Hoffmanb,c,g,h,∗M achine learning for integrating data in biology and medicine: Principles, practice, and opportunities Science Direct 201

Cite this Article :

Divya Ebenezer Nathaniel, Sonia Panesar, "A Survey of Machine Learning models for the wide Spectrum of Computational Biology ", International Journal of Scientific Research in Computer Science, Engineering and Information Technology (IJSRCSEIT), ISSN : 2456-3307, Volume 5 Issue 1, pp. 604-611, January-February 2019. Available at doi : https://doi.org/10.32628/CSEIT2063149
Journal URL : https://ijsrcseit.com/CSEIT2063149

Authors :

Prof Divya Ebenezer Nathaniel working as an Assistant professor in Department of Computer Science and Engineering, Babaria Institute of Technology, Vadodara, Gujarat. She has a teaching experience of 15 years she guided several under graduate projects and her area of interest include image processing, distributed system, Machine learning she acquired her MTech from Anna university .
Email: divya.mtech2010@gmail.com

Prof. Sonia Panesar is working as a Assistant Professor in Computer Science Engineering at Babaria Institute of Technology,Vadodara She has an academic experience of 4 years. Her research interest includes areas of Machine Learning, Deep Learning, Object Detection, NLP etc.
Email:leosonia@gmail.com