JOURNAL OF MECHANICS OF CONTINUA AND MATHEMATICAL SCIENCES
www.journalimcms.org

ISSN (Online) : 2454 -7190, Special Issue, No.-7, February (2020) pp 86-94 ISSN (Print) 0973-8975

# AN EMPIRICAL SCIENCE RESEARCH ON BIOINFORMATICS IN MACHINE LEARNING

**Sindhu V[1], Nivedha S[2], Prakash M[3]**

[1]Assistant Professor, Department of IT, Karpagam college of Engineering, Coimbatore

[2]Assistant Professor, Department of CSE, Karpagam college of Engineering,Coimbatore

[3]Professor, Department of CSE, Karpagam college of Engineering, Coimbatore

[1]v.sindhu.velu@gmail.com, [2]nivedha.gs6@gmail.com, [3]salemprakash@gmail.com

https://doi.org/10.26782/jmcms.spl.7/2020.02.00006

## Abstract

*The subset of Artificial Intelligence (AI) is Machine Learning. Machine Learning (ML) has a rapid growth in all fields of research such as medical, bio-surveillance, robotics and all other industrial applications. Improvements in accuracy and efficiency of ML techniques in bio-informatics have steadily increased for solving problems in medicine. The aim of this paper is to give brief note about applications of ML in bio-informatics and science research. Bioinformatics involves the interaction of biology, computer science and statistics. In bioinformatics, Data were extracted, analyzed and classified for the prediction of various diseases. This process is time consuming and expensive. To reduce the cost and time, traditional techniques for extracting and analyzing the data were replaced by machine learning techniques.*

**Keywords:** Machine Learning, Deep Learning, Bioinformatics

## I. Introduction

Machine Learning (ML) is a type of AI that "learns" as it recognizes new patterns from the given data. It provides machines with the ability to gain knowledge from the data without explicit programming. ML focuses on the transformation of programs, when exposed to new data. It involves computer to get trained using a given data set, and use this training to predict the properties of a given new data. For



*The Paper Presented at 14th International Conference on Intelligent System and Control (ISCO'20) Organized by The Department of Computer Science and Engineering, Karpagam College of Engineering, Coimbatore, India*

example, computer can be trained by feeding it with 100 images of dog and 100 more images which are not dog by informing each time, whether it is dog or not. After the training, when the computer is exposed with a new image, computer can able to decide whether this new image is horse or not. Machine learning methods are used for making predictions or classifications from the given raw data. These methods are the algorithm that describes how the classifications or predictions are made using the given data and can allow for a larger number of predictors, referred to as high-dimensional data. These methods can identify the relationships and interactions within the variables and can be applied to predict continuous outcomes, generally referred to as regression type problems.

## II. Machine Learning and Human Learning

People gain knowledge from the ingestion of facts as well as social norms. Human brain senses the data it process, and categorizes it. Brain tries to understand and make sense of the data, when the processing data is compared with new data. Similar to brain, Algorithms identify the patterns in the information. Basic learning process is same for both human and machine/computer. It includes data, abstraction, generalization and evaluation. There are number of ways humans can learn, including observation, input and examples. Decisions are made based on the input or observations received, which is combined with the memories to build up the knowledge. Outcome or skill can be obtained using observation and learning process (decisions). For humans, input/observation occurs through sensing and distinguishing the things around them. Since machines do not have physical sensing ability like human, it collects data through things speech recognition etc. Machine also gathers or collects information through learning by example. When provided with set of data with answers. Gradually the machine will differentiate between the answers that produce correct outcome. Learning by algorithm refers to program that instructs the computer. Output will be result of machine interaction around it like navigation, generation of speech etc.

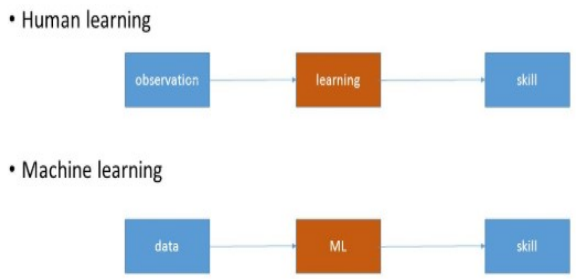


**Fig.1:** Human Learning Vs. Machine learning

## III. Artificial Intelligence vs. Machine Learning vs. Deep Learning

In the research of big data, ML and AI are used conversely. But they are not same, it is essential to understand, how ML and AI are enforced differently.

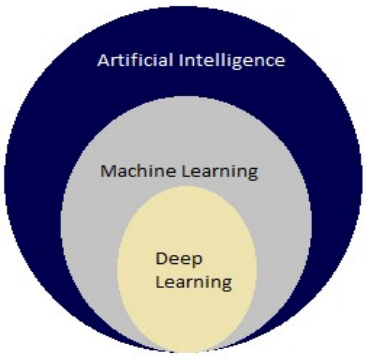


**Fig.2:** AI vs. ML vs. DL

**Artificial Intelligence and Machine Learning :**

AI is a wider research area than ML, which refers to the use of machines to imitate the functions of brain. AI makes the computers/devices to perform tasks in a "smart" way based on algorithm. While ML is a part of AI, ML focuses on machines' ability to abstract and generalized the data as they alter the algorithm based on the data it is processing. ML algorithms can be used to prepare the machines to think like humans.

**Deep Learning:**

Deep Learning (DL) is a subfield of ML concerned with algorithms imitates the structure and function of the human brain. DL goes further deeper than ML. DL networks need to review large quantity of data-items. To get trained, instead of programming, the machine is exposed to millions of data. DL need not to programmed, they have to identify the edges when exposed to new data.

## IV. ArchitectureWorkflow of Machine Learning

The following are the list of stages for developing and managing a ML,

- Data Preparation
- Develop the ML model.
- Train an ML model
- Get Prediction
- Monitor the predictions on an ongoing basis.
- Manage the model

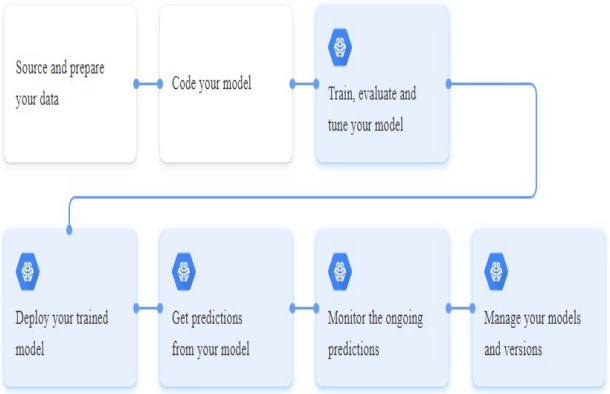


**Fig.3:** Stages in ML Workflow

Given dataset is spitted into three subsets; one part is for training the data, second for evaluation and last is used for testing the data. The training data includes features or attribute, based on which other features can be inferred. Data preparation involves data analysis and preprocessing. In data analysis, identify the patterns and features of the data, clean the contents that are not related to the features. Data preprocessing is the transformation of clean data into format that suits the model. Normalizing, reduce the data redundancy, applying formatting rules are some of the features of data preprocessing. Develop model using ML by defining operations. Model should be trained and evaluated with the given model. Training data can be executed to predict the output values. Based on the predictions, model can be adjusted to get better data and to predict the outcomes more accurate. For evaluation, input would be the trained data with target values. Finally test the data and deploy it for further predictions. Monitor the results of prediction for further analysis.

## V. Applications

Progression in ML can be clearly viewed based on the real-time applications that are developed. A brief description on some of the ML algorithm based applications are listed below,

- **Speech recognition:** Present Speech based recognition system that are in use, applies ML methods to train the model for more accuracy.
- **Computer vision:** Facial recognition systems, which are able to identify the images of cells, are computer vision systems. It employs ML methods for more accuracy. For example, Post office in US uses CV for handwriting analysis to sort the handwritten addresses.
- **Bio-surveillance:** Outburst of diseases can be identified using ML. Automated Learning methods can be handle this complex dynamic data.

- **Empirical science research:** ML is implemented in intensive science disciplines. In genetics, ML algorithms are used to discover unusual celestial objects in Neuroscience and psychological analysis.

## VI. Empirical Science Research

In Science Research, ML algorithm is used in the field of bio-informatics for protein structure prediction, microarray, gene expression data etc. Bioinformatics is the interaction of biology, computer science and statistics.

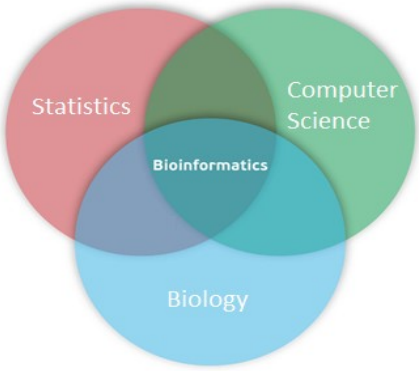


**Fig.4:** Interaction of computer science- biology – statistics

**Genomics:** It is about the study of genome. Genome is a DNA sequence of a living organism. It refers to the complete set of genes within a particular organism. It provides the researches with the ability to track the genetic sequence that directs all the activities of an organism. Genomics is related to personalized medicine (Precision medicine), an approach for patient care which includes genetics, activities and surrounding with the purpose of employing patient–specific treatment. Due to high cost and limitation in technology, researchers prefer ML learning techniques. Application of ML in genomics includes Genome sequencing, Direct-to-consumer genomics and Gene-editing. Genome sequencing involves in predicting the individual's probability of having diseases and therapies. Direct-to-consumer genomics includes analysis of genetic information. Gene-editing refers to the specific alteration to DNA at cellular level. Micro-array is used for monitoring the gene expression, which is helpful for diagnosing different types of cancer. ML provides a solution for identification based on its classifying methods.

**Micro arrays:** It is a collection of huge amount of biological data. Here, ML algorithms can be applied in classification, identifying the pattern and generic

network generation. Some of microarray applications include analyzing genes under different circumstances, predicting the responses of genes on certain therapies, responses based on surrounding stress. In Microarray, ML can be used to analyze gene based on the data, differentiate gene stages and predict future stages of genes for the prevention of diseases.

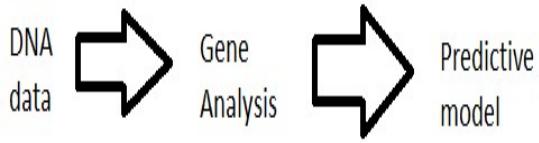


**Fig.5:** ML process on Gene Analysis

**Systems biology**: It is about the study of components such as DNA, proteins, RNA, metabolites. It includes the behaviors from interaction of biological components. ML plays a vital role in interacting with biological system in domain such as signal transduction networks, genetic networks, and metabolic pathways. The most commonly used methods are probabilistic graphical models, transcription factor binding sites, genetic algorithm

**Stroke Diagnosis:** Stroke can be diagnosed by using ML. It will analyze through neuro-imaging data, CNN (Convolutional Neural Network) and SVM (Support-Vector Machines) methods are used .It is generally in three-dimensional format.

**Transcryptomics**: It is the complete arrangement of RNA transcripts that are created by the genome under explicit conditions or in a particular cell. Bioinformatics is utilized for transcriptome investigation where mRNA articulation levels can be resolved.

**Cheminformatics**: It includes applying various ML methods in chemistry to predict molecule's structure, shape, chemical properties etc. It refers to the prediction of relationship between similarity and properties of molecules. Recurrent Neural networks (RNN) and Graph convolutional Networks (GCN) are the popular DL architecture to work on molecular structure. RNN allows generating new sequences which finds the molecules with desirable properties. GNN takes the features and molecule graph as input and convert it into matrix form to merge it with ML model.

**Drug Discovery:** The application of bioinformatics cut across the entire process of drug discovery, thereby

- Reducing the risk of drug failure
- Making it a bit cheaper
- Reducing the time spent in the discovery
- It automates the entire process, thereby reducing human intervention.

**Text mining:** It allows gathering of unstructured data from various source, detect abnormalities in data based on the pre-processing operations. Convert the information to structured format (unstructured data which is extracted to structure data). Analyze the pattern based on the information. Store the information for enhanced decision – making. Text mining techniques involves information extraction, retrieval, categorization and clustering. In bioinformatics, identification of entities that allows organizing medical records and extracting the linked medical research can be done used text mining. It allows the user to gather information from various pieces of text and construct a detailed medical graph related to it. There exists lot of text recognition ML algorithms which are trained by the training data. Data includes labels for each part in the text. Initially, word-level features were extracted and trained with the model. Later character-level extraction which encodes the lexical information observed in the characters. These pre-trained characters are trained on large volume of medical texts.

**Proteomics:** It refers to the study of proteomes. Proteins are the group of amino acids, which gain their function from protein folding. A proteome is a set of proteins produced in living organisms. It differs from cell to cell and changes over the time. It contain significant biological information which includes Interaction of proteins (tumor suppressor protein), proteins localized in sub-cellular compartment etc. ML algorithms where used for prediction of problems based on the information. Prior to ML, researches needed to test the prediction of protein structure manually. This process is time-consuming and expensive. Now the protein secondary structure prediction is made using deep convolutional neural networks.

## VII. Conclusion

This paper attempts to study and provide a brief knowledge about the difference between AI, ML and DL. Most common approach of ML is bio-informatics. To solve complex biological problems, bio-informatics merges with the field of mathematics, information technology, statistics and computer science. This survey gives the knowledge about different classification methods and Bioinformatics approach. The approach has been of great importance to develop fast and accurate target identification and prediction method for discovery.

**References**


I. Aerts, S., Van Loo, P., Moreau, Y., & De Moor, B. (2004). A genetic algorithm for the detection of new cis-regulatory modules in sets of coregulated genes. Bioinformatics, 20(12), 1974-1976.

II. Alom, M. Z., Taha, T. M., Yakopcic, C., Westberg, S., Sidike, P., Nasrin, M. S., ...&Asari, V. K. (2019). A state-of-the-art survey on deep learning theory and architectures. Electronics, 8(3), 292.

III. Alpaydin, E. (2009). Introduction to machine learning. MIT press.

IV. Bhaskar, H., Hoyle, D. C., & Singh, S. (2006). Machine learning in bioinformatics: A brief survey and recommendations for practitioners. Computers in biology and medicine, 36(10), 1104-1125.

V. Bockhorst, J., Craven, M., Page, D., Shavlik, J., &Glasner, J. (2003). A Bayesian network approach to operon prediction. Bioinformatics, 19(10), 1227-1235.

VI. Buskirk, T. D., Kirchner, A., Eck, A., &Signorino, C. S. (2018). An introduction to machine learning methods for survey researchers. Survey Practice, 11(1), 2718.

VII. Das, K., &Behera, R. N. (2017). A survey on machine learning: concept, algorithms and applications. International Journal of Innovative Research in Computer and Communication Engineering, 5(2), 1301-1309.

VIII. Ferrucci, D., Brown, E., Chu-Carroll, J., Fan, J., Gondek, D., Kalyanpur, A. A., ...&Schlaefer, N. (2010). Building Watson: An overview of the DeepQA project. AI magazine, 31(3), 59-79.

IX. Gentleman, R., Carey, V., Huber, W., Irizarry, R., &Dudoit, S. (Eds.). (2006). Bioinformatics and computational biology solutions using R and Bioconductor. Springer Science & Business Media.

X. Inza, I., Calvo, B., Armañanzas, R., Bengoetxea, E., Larrañaga, P., & Lozano, J. A. (2010). Machine learning: an indispensable tool in bioinformatics. In Bioinformatics methods in clinical research (pp. 25-48). Humana Press.

XI. Kaur, S., & Jindal, S. (2016). A survey on machine learning algorithms. Int J Innovative Res AdvEng (IJIRAE), 3(11), 2349-2763.

XII. Larranaga, P., Calvo, B., Santana, R., Bielza, C., Galdiano, J., Inza, I., ...& Robles, V. (2006). Machine learning in bioinformatics. Briefings in bioinformatics, 7(1), 86-112.

XIII. Mathé, C., Sagot, M. F., Schiex, T., &Rouzé, P. (2002). Current methods of gene prediction, their strengths and weaknesses. Nucleic acids research, 30(19), 4103-4117.

XIV. Manyika, J. (2011). Big data: The next frontier for innovation, competition, and productivity. http://www.mckinsey.com/Insights/MGI/Research/Technology_and_Innovation/Big_data_The_next_frontier_for_innovation.

XV. Mitra, S., Datta, S., Perkins, T., &Michailidis, G. (2008). Introduction to machine learning and bioinformatics. Chapman and Hall/CRC.

XVI. Parmigiani, G., Garrett, E. S., Irizarry, R. A., &Zeger, S. L. (2003). The analysis of gene expression data: an overview of methods and software. In The analysis of gene expression data (pp. 1-45). Springer, New York, NY.

XVII. Sapp, C. E. (2017). Preparing and architecting for machine learning. Gartner Technical Professional Advice, 1-37.

XVIII. Witten, I. H., Frank, E., Hall, M. A., & Pal, C. J. (2016). Data Mining: Practical machine learning tools and techniques. Morgan Kaufmann.