ISSN:0975-9646

Machine Learning Approach to Detect Tuberculosis in patients with or without HIV co-infection - A Survey

Ashwini D.V.[1]and Dr. Seema S.[2]

[1]*Department of Computer Science Engineering*
M.S. Ramaiah Institute of Technology, Bangalore, India.

[2] *Associate Professor, Department of Computer Science Engineering*
M.S. Ramaiah Institute of Technology, Bangalore, India.

***Abstract*—Data mining technique has seen the tremendous increase in their application recently. The biggest challenge has been related to the creation of meaningful data model and an iterative process of using data transformation mechanism. These challenges are coupled with the unsolved problems that arise while treating data that results in high inaccuracy by using conventional methods. Huge amount of data is generated in health care transaction which is increasingly becoming difficult to understand, manage and analyse manually. In order to analyse clinical data effectively and keep up with the technology producing high intensity data, Machine Learning techniques are involved to solve this complex problem. A machine learning algorithm learns from the previous experience and automatically takes decisions. Gene Expression levels change over time. These profiles can distinguish between cells that are actively dividing, or show how the cells react to a particular treatment. The knowledge acquired though the data analysis can't be handled explicitly for future use. They may also contain missing data value which makes analysis more difficult. Therefore sets of genes that are co-ordinately expressed in different diseases and defined as specific modules, often demonstrating coherent functional relationships through unbiased literature profiling. There are many papers who have applied different data mining and machine learning techniques to solve these challenges. This paper highlights the main challenges involved in analysing and identifying the patients and classifying them into HIV with TB and those having HIV without traces of TB disease. This paper is a brief survey on the techniques used to perform the classification of HIV/TB co-infected patients.**

***Keywords*— Gene Signature, MTB, HIV-TB co-infection, SVM**

I. INTRODUCTION

Understanding biological information and there analysis has given rise to new field called Bioinformatics. Bioinformatics is an application of computer technology to the management of biological information. Computers are used to collect the data, store them effectively, analyze for meaningful information and integrate biological and genetic information which can then be applied to gene-based drug discovery and treatment of patients in future. The need for research in bioinformatics has been precipitated by the explosion of publicly available genomic information resulting in many Human Genome Projects.

Biological data are being produced at a phenomenal rate. As a result there is a need to study these data and probe the complex dynamics observed in nature. The aims of bioinformatics are threefold; First, at its simplest bioinformatics organizes data in a way that allows researchers to access existing information and to submit new entries as they are produced. The second aim is to develop tools and resources that aid in the analysis of data. The third aim is to use these tools to analyse the data and interpret the results in a biologically meaningful manner [1].

Human Immunodeficiency Virus type (HIV) and Tuberculosis have been associated with each other over the past few decades, the severity of this co-infection has been extensively examined and analysed in clinical studies and many researches have been involved in this assisting the World Health Organization to reduce the rapid increase rate of the co-infection. The most common infection in HIV-positive affected patients is closely linked with Tuberculosis (TB), which is also considered as most common reason for death in HIV/AIDS patients. By producing a progressive decline in CD4 cell count affecting the immunity of the patients, HIV alters the pathogenesis of TB, greatly increasing the risk of disease from TB in HIV co-infected individuals and leading to more frequent extra pulmonary involvement, atypical radiographic manifestations, and paucibacillary disease, which can slow down by timely diagnosis [2]. The World Health Organization (WHO) estimates that one third of the world's population is infected with Mycobacterium tuberculosis, resulting in an estimated nearly 9 million new cases of active TB in 2010. Worldwide, 14.8% of TB patients have HIV co-infection, and as many as 50-80% have HIV co-infection in parts of sub-Saharan Africa [2].

The Drastic increase in the HIV infection in different parts of the world has resulted in the notifiable growth in the TB overlapping up to 10 times. In some of the developed countries with well-equipped TB control treatments, it is seen that there is new patients with TB are recorded. TB rate cannot be decreased until the HIV control is achieved. An imperative requirement is there should be an effective control program to decrease the increasing rate of HIV

patients as there is close synergy between TB and HIV/AIDS.

The T-cells of the TB infected patients will be primary target of attack by the HIV bacteria causing decline in the immune system of the body. CD4+ cells of T-cell will act as a secrete lymphokines that increases the capacity of macrophages in killing the mycobacteria. In most of the patients it will only be as latent TB which will not be exhibited for long years. In HIV infected patients, the virus first target for depletion and dysfunction of CD4+ cells.

On an average the HIV infected patient survives eight to ten years without the treatment, but with the improvement of modern technologies in medicine it has been extended considerably. HIV infection can be associated with various diseases ranging from normal fever to various neurological symptoms which are mostly self-limiting in nature. The HIV patients have high chances of co-infection with TB. An HIV infected person co-infected with Tuberculosis has a 50 percent lifetime risk of developing TB disease, but the risk is less in-case of HIV-negative patients. There is a need of keen importance to be given to two epidemic diseases by both TB as well as HIV/AIDS control programmers. There are various techniques adopted to find out and examine TB/HIV co-infection which is discussed below.

II. Growing Burden of TB infection

Globally there are approximately 9 million new active tuberculosis cases and 1.4 million deaths annually (Tuberculosis, 2014). Blood transcriptional signature can be used to differentiate between active and latent TB in HIV infected patients. Human Immunodeficiency Virus Infection is a potential risk factor for Tuberculosis. There are chances of re-activation of TB Infection. The risk of re-infection in person with MTB around the world in his lifetime is only 10-20%, whereas co-infection of HIV/MTB is around 10% annually. In this paper we focus on relationship between TB and HIV, and on the growing magnitude of TB problem. Various reasons for the rapid replication rate of virus can also be Drug resistant HIV virus, drug resistant tests are conducted which results to be time taking and costly to afford.

One third of the global population lacks reliable access to needed medicines. The situation is even worse in the poorest countries of Africa and Asia, where as much as 50% of the population lacks such access. While some 10 million lives a year could be saved by improving access to essential medicines and vaccines – 4 million in Africa and South-East Asia alone – a major obstacle to achieving this has been price (EL, 2003).There is a need for a simple, low-cost, point-of-care assay for tuberculosis is most required.

III. Related Work for HIV/TB infection

HIV/TB prediction has a major challenge because of huge genes to be compared together, in machine learning this problem is referred as feature selection. For the efficient gene selection noise has to be reduced to increase performance, and reduce the computational cost also finding more interpretable features. Among the large gene number, only few have biological relationship with the targeted diseases. Monitoring gene for the particular disease is referred to Gene Expression analysis. Here mRNA is analysed instead of proteins. Gene Expression analysis helps to determine the fate and functions of the cell.

A. Gene Expression Analysis in HIV/TB co-infection

Gene expression profiling is the measurement of the activity (the **Error! Bookmark not defined.**) of thousands of genes at once, to create a global picture of cellular function. These profiles can distinguish between cells that are actively dividing, or show how the cells react to a particular treatment. It was shown that transcriptional signature in blood correlates with extent of disease in patients with active TB, and reflects changes at the site of disease. The transcriptional signature was diminished in patients with active TB after 2 months, and completely extinguished by 12 months after treatment, reflecting radiographic improvement.

The 393-gene active TB signature may also reflect common inflammatory responses are developed during many other diseases. Therefore TB related separate 86-gene whole-blood signature was considered for analysis, which was compared with patients possessing active TB. This 86-gene signature was also tested against patients normalized to their own controls from seven independent data sets by class prediction (k-nearest neighbours). To identify functional components of the transcriptional host response during active TB, we used a modular data-mining strategy, using sets of genes that are co-ordinately expressed in different diseases and defined as specific modules, often demonstrating coherent functional relationships through unbiased literature profiling.

When comparing the Gene signature of active TB patients with other infections like SLE infection demonstrated decreased over-representation of the IFN-inducible module also with decrease in the abundance of B-cells and T-cells but plasma-cell-related module which was present in SLE infected patients but absent in TB infection. The blood modular signature of active TB patients were also compared with group A Streptococcus where there was almost no change in IFN-inducible module but marked over-representation of the neutrophil-related module (M2.2), distinguishing these diseases from TB. By this we can differentiate other infections for M-TB infection.

By the cytometry test it was evident that there was significant decrease in T-cells carrying CD4 and CD8 antigens, but there was increased cell count of myeloid transcripts which was less pronounced. Type I IFN signalling is crucial for defence against viral infections but there was gene downstream of both IFN-γ and type I IFN-αβ receptor signalling in active TB patients. Thus complete description of the human blood transcriptional signature of TB is provided. The signature of active TB, observed in 10–20% of patients with latent TB, may identify those individuals who will develop active disease, facilitating targeted preventative therapy.

B. HIV/TB co-infection analysis using Support Vector Machine

Support Vector Machine has the properties like duality, ability to incorporate kernels, margin maximization, convexity and sparseness. Using the CD4 cell count would be more valuable for physicians to determine and treat HIV-positive patients.

Pre-processing of dataset: The various steps involved before feeding into the machine learning algorithm are:

Consider Subtype B consensus protease (PR) genome sequences, CD4 count, viral load and the number of weeks from the baseline measure of CD4 count for each patient sample was determined by joining individual datasets using the sample identifier (the unique number that identifies a sample) and date for pre-processing.

PR dataset is further processed to reduce amino acid positions. The number of mutation occurred is also calculated at each position. Positions with mutation less than 20% are removed from the dataset.

Data of Patient's genome sequences and associated viral load and CD4 count data at different time points are also calculated.

Dataset was further processed to determine the change in CD4 count between these measurements. 65 random data elements were removed from each dataset and formed a testing set. Training was done on the remaining data elements of the protease datasets.

Machine Learning Algorithm: Three different inputs are given to Machine Learning algorithm separately: Only genome sequence, genome sequence along with current viral load and No. of weeks from current CD4 count to baseline CD4 count.

Output of the Algorithm: The classification model was build based on the CD4 count changes. The four categories were grouped on the changes observed:

$$Classification = \begin{cases} Output1, & \Delta CD4 < 0 \\ Output2, & 0 \le \Delta CD4 \le 50 \\ Output3, & 51 \le \Delta CD4 \le 100 \\ Output4 & \Delta CD4 > 100 \end{cases}$$

The SVM will be used to find the mapping between the three input groups and four output groups. The results obtained by this experiment shows us that the time component is the valuable parameter in analysis of CD4 cell count as there was not much difference between e RBF, linear and quadratic SVM's with respect to input models 1 and 2. But significant difference was observed between input model 1 and 2 with input model 3.

According to the study there are mainly three reasons for the drug resistance HIV are [5]: 1. Due to the mutations, the structural conformation of the HIV-1 protease changes, and therefore directly affects the interactions between the inhibitors or substrate and the HIV-1 protease at the active site 2. Mutations indirectly change the ability of the protease to bind inhibitor 3. Mutations at the dimer interface may decrease the stability of the protease, and thus weaken the enzymatic function of the HIV-1 protease.

The molecular structure can also be used to predict the drug resistance value to the mutations to certain drug/inhibitor [5].A Delaunay tessellation derived four-body statistical potential mutagenesis method together with support vector machine (SVM) and random forest classification methods is applied to predict the drug resistance for HIV-1 reverse transcriptase inhibitor.

C. HIV/TB analysis using Gini Co-relation Co-efficient

The entire document should be in Times New Roman or Times font. Type 3 fonts must not be used. Other font types may be used if needed for special purposes.

Understanding the human gene system using Gene co-expression networks (GCN) is become a research topic in recent years. In the GCN, each node represents a gene, and the edge links two co-expressed genes. The edge weight is usually determined with the similarity of gene expression profiles using the Pearson correlation coefficient (PCC) method.

The inequality or imbalance in the symbiosis and pathogenesis duality is the major reason for the microbial infection. HIV/AIDS is also microbial diseases which is very complex and is difficult and almost no curing aids. The three stages of HIV-1 infection are: acute, asymptotic and AIDS stage. Using Gini method the inequalities of the connectivity and edge weight in HIV-1 stages represent GCN's. The steps followed to find out the different patterns of gene during all these stages of HIV infected patients are:

The microarray dataset considered for analysis contains Affymetrix gene expression profiles of human lymphatic tissues from both infected and uninfected subjects. Finally 908 probes (704 genes) was considered for the construction of GCNs. Considering the log2 transformed gene expression of these 704 gene, four GCN's were constructed with one for uninfected patients and three as different stages of HIV infection. Similarity was obtained by using PCC method. For example if we consider gene A and gene B of uninfected patients:

$$\mathrm{PCC} = \frac{\sum (a_i - a_m)(b_i - b_m)}{\sqrt{\sum (a_i - a_m)^2 \sum (b_i - b_m)^2}},$$

, where ai = log2-transformed gene expression of gene A in the ith subject,

bi= log2 transformed gene expression of gene B in the ith subject,

am = mean of log2-transformed gene expression of gene A,

bm represents the mean of log2- transformed gene expression of gene B.

The significance level of PCC value is estimated with the statistic result of

$t = \mathrm{PCC} \sqrt{((n-2)/(1-\mathrm{PCC} \cdot \mathrm{PCC}))}$

PCC value is assigned as the edge weight of these two genes.

A well-defined method to identify the inequalities in the population is Gini- method which uses Gini-coefficient to measure the inequalities in GCNs. For the given variable X Gini coefficient can be computed as:

$$G = \frac{\sum_{i=1}^{n} (2i - n - 1) X_{(i)}}{(n-1) \sum_{i=1}^{n} X_{(i)}}$$

Where n (n ≥ 2) = number of considered variable in the population

X(i) = ith value of considered variable sorted in increasing order, $0 \leq X(1) \leq X(2) \leq \cdots \leq X(n)$.

The components considered in GCNs are positive and negative connectivity. The Gini correlation of the positive connectivity can be represented by:

$$R_p = \frac{\sum_{i=1}^{k}(2i-k-1)P_{[i]}}{\sum_{i=1}^{k}(2i-k-1)P_{(i)}}$$

Where (Pi, Ni) = positive and negative connectivity of the ith gene in GCN,

k = number of analysed genes.

Rn is also calculated is similar way

P(i) and P[i] are obtained by two different ways. For P(i), the positive connectivity's of analysed genes are firstly sorted in an ascending order, and then the P(i) is used to represent the ith positive connectivity sorted in this order. Whereas for P[i], the connectivity's of analysed genes are firstly sorted in an increasing order, then the P[i] is used to represent the concomitant positive connectivity of ith connectivity (Chuang Ma, 2012). If Gini co-efficient is less than zero which implies that positive connectivity decreases and thus decreasing the inequalities, where as if the Gini-correlation is higher, positive connectivity increases thus inequalities also increases.

For a given gene i in two GCNs (N1, N2), we can calculate ΔG+ and Δ G- respectively

$$\Delta G_{+}\left(\frac{N_2}{N_1}, i\right) = G_{+}(N_2, i) - G_{+}(N_1, i)$$

$$\Delta G_{-}\left(\frac{N_2}{N_1}, i\right) = G_{-}(N_2, i) - G_{-}(N_1, i)$$

Where G+ (Nj,i) and G- (Nj, i) = Gini coefficients of positive and negative edge weight in the Nj(j = 1 or 2) respectively.

Here common permutation method is used to determine the statistical significance of ΔG+ and Δ G-

The result can be discussed with respect to three aspects:

Dual Positive and Negative Connectivity in GCNs of HIV-1 Infection: The positive connectivity varies from uninfected patients to HIV infected patients in all three stages. Thus the experiment shows that statistical analysis of positive and negative connectivity's can be used and could be helpful in understanding pathogenesis mechanism exhibited in HIV infected patients.

Connectivity Inequality in GCNs of HIV-1 Infection: There were also remarkable differences between the Gini coefficients of positive and negative connectivity in Nacut and Nasym. The dynamic changes between positive and negative inequalities also had significant differences. Further it was observed that Gini positive and negative inequalities also contributed for overall inequality in GCNs. Negative connectivity had major changes in different stages of HIV-1 infection over positive connectivity.

Edge Weight Inequality in GCNs of HIV-1 Infection: With the help of permutation test method as mentioned above we identified a set of genes with significant ΔG+ or ΔG− between GCNs of the uninfected subjects and infected patients at different stages. Compared with the number of gene with significant ΔG+, the number of genes with significant ΔG− is relatively large, also indicating the differences in inequality between positive and negative co-expression links.

D. HIV/TB analysis using Artificial Neural Network

The set of processing units called neurons forms, artificial neural network. There are various types of Neural Network, pictographically shown in the figure1.

- **Feed Forward Neural network:** The information in these types of networks move in one direction from input nodes to hidden nodes and finally output node.
- **Recurrent Networks:** In these types of networks, the data flow is bi-direction. The data can be flowing back to earlier stages when required.
- **Modular Neural Networks:** Modular NN are the series of independent neural network moderated by some node. Every node performs a separate subtask, the output of one task is considered as input for other node, finally providing network of the module as a whole.
- **Associative Neural Network:** These types of networks are the combination of feed-forward neural networks and the k-nearest neighbour technique.

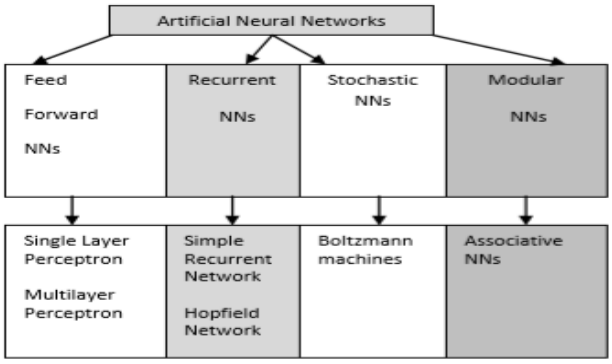

Figure1: Different types of Neural Networks

Neural Networks process large number of simple units in parallel. The ability of identifying the active TB could be improved using computer analysis using Neural Networks. The patients were isolated for analysis of active TB. The patients with HIV infection has CD4 count greater than 200cells/μL.

A General Regression Neural Network (GRNN) was used for the development. GRNN does not require a priori specification of the structure of the regression equations, and also the regression surface is not constrained to specific form. In this research GRNN parameters are divided into three layers of: Input layer, Hidden layer and Output Layer. The Hidden Units finds the higher-order inputs features in the input layer and sends the output to other neurons. The output layer determines the estimate of the active TB in the patients selected for the analysis. The 10 fold cross-

validation is implemented. The Data set was divided into 10 subsets, out of which 9 are pooled and used in training and the 10 subset were used as evaluation set. The mean square error was computed for each of the 10 neural networks on the entire derivation data set. The artificial Neural network with which had mean square root closest to the average value was selected.

IV. CONCLUSIONS

This Survey paper mainly talks about various machine learning algorithms, in analysing the gene expression of HIV infected patients. This survey studies within the most popular sub-branch of Health informatics. The focus was on analysing the various co-infection associated with HIV infected individuals. Death rate with respect to HIV co-infected patients is increasing dramatically worldwide. Various publicly available datasets are utilized in the above discussed machine learning algorithms in answering various medical questions eventually to improve the healthcare of patients.

ACKNOWLEDGMENT

I would like to thank Dr. Seema S. Associate Professor MSRIT Department of Computer Science, for her valuable support, guidance. I would take this opportunity to express my profound gratitude and deep regard for her exemplary guidance, valuable feedback and constant encouragement throughout.

REFERENCES

[1] N.M. Luscombe,D. Greenbaum,,M. Gerstein "What is bioinformatics? An introduction and overview" Review Paper Department of Molecular Biophysics and Biochemistry Yale University New Haven, USA.

[2] Annie Luetkemeyer" Tuberculosis and HIV" HIV Insite Knowledge Base Chapter Janauray 2013.61.

[3] [Online] www.who.int/mediacentre/factsheets/fs104/en/

[4] Corbett EL, Watt CJ, Walker N, Maher D, Williams BG, Raviglione MC, Dye C"The Growing Burden of Tuberculosis- Global trends and interactions with the HIV epidemics", Arch Intern Med. 2003 May 12;163(9):1009-21

[5] Yashik Singh and Maurice Mars "Support vector machines to forecast changes in CD4 count of HIV-1 positive patients" Scientific Research and Essays Vol. 5(17), pp. 2384-2390, 4 September, 2010.M. Wegmuller, J. P. von der Weid, P. Oberson, and N. Gisin, "High resolution fiber distributed measurements with coherent OFDR," in *Proc. ECOC'00*, 2000, paper 11.3.4, p. 109.

[6] Ali A. El-Solh, Chiu-Bin Hsiao, Susan Goodnough, Joseph Serghani, Brydon J. B. Grant "Predicting Active Pulmonary Tuberculosis Using an Artificial Neural Network" CHEST / 116/4/ OCTOBER, 1999.

[7] Chuang Ma and Xiangfeng Wang "Application of the Gini Correlation Coefficient to Infer Regulatory Relationships in Transcriptome Analysis" Plant Physiology, September 2012, Vol. 160, pp. 192–203, www.plantphysiol.org 2012 American Society of Plant Biologists.

[8] Supawan Prompramote, Yan Chen1 and Yi-Ping Phoebe Chen1,2 "Machine Learning in Bioinformatics" Springer.*FLEXChip Signal Processor (MC68175/D)*, Motorola, 1996.

[9] Xiaxia Yu "HIV Drug Resistant Prediction and Featured Mutants Selection using Machine Learning Approaches" ScholarWorks @ Georgia State University 2014

[10] Matthew P. R. Berry1, Christine M. Graham1*, Finlay W. McNab1*, Zhaohui Xu6, Susannah A. A. Bloch3, Tolu Oni4,5, Katalin A. Wilkinson2,4, Romain Banchereau9, Jason Skinner6, Robert J. Wilkinson2,4,5, Charles Quinn6, Derek Blankenship7, Ranju Dhawan8, John J. Cush6, Asuncion Mejias10, Octavio Ramilo10, Onn M. Kon3, Virginia Pascual6, Jacques Banchereau6, Damien Chaussabel6 & Anne O'Garra1 "An interferon-inducible neutrophil-driven blood transcriptional signature in human tuberculosis" Vol 466|19 August 2010|doi:10.1038/nature09247.

[11] Isabelle Guyon, Jason Weston, Stephen Barnhill, "Gene Selection for Cancer Classification using Support Vector Machines" Machine Learning, 46, 389–422, 2002 Kluwer Academic Publishers.

[12] *Wireless LAN Medium Access Control (MAC) and Physical Layer (PHY) Specification*, IEEE Std. 802.11, 1997.