A Study On Data Mining Approach To Precision Medicine

N.Narmadha[1], J.Divyaa[2], A.Kalyani[3], M.S.Geetha Devasena[4]
[1, 2, 3, 4] Department of Computer Science and Engineering
[1, 2, 3, 4] Sri Ramakrishna Engineering CollegeCoimbatore

Abstract- *The data mining is the task of analysing of large quantities of data to derive previously unknown, interesting patterns such as groups of data record, unusual record and dependencies. It has the ability to turn raw data into useful information. There is a pressing clinical need to improve early prevention and clinical management of type 2 diabetes and its complications. Predict the data which are exceeding the normal level and also remedies for type 2 diabetes patient with phenotypically similar but genotypically and molecular dissimilar disease. It uses precision medicine approach to characterize the complexity of diabetes patient identifying three distinct subgroup of type 2 diabetes. The efficient method for type 2 diabetes is neural network algorithm. This method works relatively fast and enables accurate unit. In this study it is used to see how to train neural network data and identifying subgroup of T2D.*

Keywords- Data mining; Precision medicine; T2D; Neural network.

I. INTRODUCTION

Everybody gather information and store large data set in system. It is an interdisciplinary subfield of computer science. The overall goal of data mining process is to extract information from a dataset and transform it into an understandable structure for further use. It has the ability to turn raw data into useful information. Type 2 diabetes is a heterogeneous complex disease. It is a progressive condition in which the body becomes resistant to the normal effect of insulin and gradually loses the capacity to produce enough insulin in the pancreas. Everyone needs some glucose in their blood but if it's too high its complications. Predict the data which are exceeding the normal level and also predict the genomic type for finding possibilities of diseases. It use precision medicine approach to characterize the complexity of diabetes patient identifying three distinct subgroup of type 2 diabetes from topology-based patient-patient networks. Subtype 1 was characterized by T2D complication diabetic nephropathy and diabetes retinopathy; Subtype 2 was enriched for cancer malignancy and cardiovascular diseases and Subtype 3 was associated with cardiovascular disease, neurological diseases, allergies and HIV infection. The approach demonstrates the utility of applying the precision medicine paradigm in T2D and promise of extending the approach to the study of other complex, multifactorial diseases can damage your body over time. Clinicians have understood the patients who carry the type 2 diabetes diagnosis have a variety of phenotypes and susceptibilities to diabetes related.

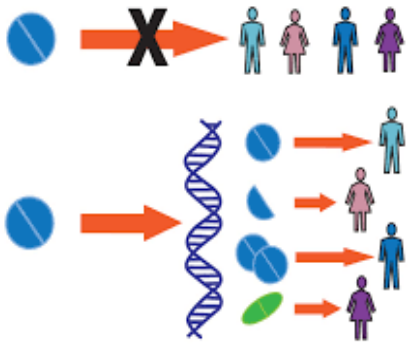

Fig 1. Precision Medicine

II. LITERARY SURVEY

David C. Klonoff proposes a Genes signifying increased risk for both type 1 and type 2 diabetes have been identified. Genomewide association studies have identified over 50 loci associated with an increased genetic risk of type 1 diabetes. Several T1D candidate genes are increased to risk of developing type 1 diabetes have been suggested or identified within these regions, but the molecular basis which they contribute to islet cell inflammation and beta cell destruction is not fully understood [1]. Also, several candidate genes for increased risk of developing type 2 diabetes have been identified, including the peroxisome proliferator-activated receptor gamma (PPARγ2), angiotensin converting enzyme (ACE), methylene tetrahydrofolatereductase (MTHR), fatty acid binding protein-2 (FABP2), and fat mass and obesity associated gene (FTO)[2].

The conclusions of a "Workshop on Metformin Pharmacogenomics," is sponsored by the National Institute of Diabetes and Digestive and Kidney Diseases, were published in 2014[3]. The meeting was intended to review metformin pharmacogenomics and identify the both novel targets and more effective agents for diabetes. The idea behind the

meeting was that understanding the genes and pathways that determine the response for metformin has the potential to reveal new drug targets for the treatment of diabetes. The group noted that there have been few genes associated with glycemic control by metformin, and the most reproducible associations to have been in metformin transporter genes. They acknowledged that nongenetic factors also contribute to response to metformin and that broader system biology approached will be required to model the combined effects of multiple gene variants and their interaction with nongenetic factors. They concluded that the overall challenge to the field of precision medicine as it's relate to antidiabetes treatment is to identify the individualized factors that can lead to improved glycemic control.

Erwin P. Bottinger proposes type 2 diabetes (T2D) is a heterogeneous complex disease affecting more than 29 million Americans alone with a rising prevalence trending toward steady increases in the coming decades. There is a Pressing clinical need to improve early prevention and clinical management of T2D and its complications. We used a precision medicine approach to characterize the complexity of T2D patient's population based on high-dimensional electronic medical records (EMRs) and genotype data from 11,210 individuals. We successfully identified three distinct subgroups of T2D from topology-based patient-patient networks. We performed a genetic association analysis of the emergent T2D subtypes to identify subtype-specific genetic markers are identified 1279, 1227, and 1338 and single-nucleotide polymorphisms (SNPs) that mapped to 425, 322, and 437 unique genes specific to subtypes 1, 2, and 3, respectively. By assessing the human disease-SNPs association for each subtypes, the enriched phenotype and biological functions at the gene level for each subtype matched with the disease comorbidities and clinical differences that we identified through EMRs.

Michael S. Atkins proposes a neural network is a system of hardware and software patterned after the operation of neurons in the human brain. Neural networks also called artificial neural networks are a variety of deep learning technologies. Commercial application of these technologies it's generally focus on solving complex signal processing or pattern recognition problems.

Neural networks usually involve a large number of processors operating in parallel and arranged in tiers. The first tier receives raw input information analogous to optic nerves in human visual processing. Each successive tier receive the output from the tier preceding it, rather than from the raw input in the same way neurons further from the optic nerve receive signals from those closer to it. The last tier produces the output of the system. The tiers are highly interconnected, which means each node in tier n will be connected to many nodes in tier n-1 its input and in tier n+1, which provide input for those nodes. There may be one or multiple nodes in the output layer.

Neural networks are notable for being adaptive, which means they have been modify themselves as they learn from initial training and subsequent runs provide more information about the world. The most basic learning model is too centered on weighting the input streams, which is how each node weights the importance of input from each of its predecessors.

III. EXISTING SYSTEM

An electronic medical record (EMR) it is a digital version of the traditional paper-based on medical record for an individual. The EMR represents a medical record within a single facility, such as a doctor's office or a clinic. The Disadvantages of Electronic Medical Records is much skill required, minimal error could mean big loss, privacy is key, better have a backup plan and also high start-up costs, substantial learning curve, confidentiality and security issues, lack of standardized terminology, system architecture and indexing. In order to overcome this problem it use neural network algorithm.

IV. PROPOSED SYSTEM

Neural Network

A neural network is a system of hardware and software patterned after the operation of neurons in the human brain. Neural networks also called artificial neural networks are a variety of deep learning technologies. Commercial application of these technologies it's generally focus on solving complex signal processing or pattern recognition problems.

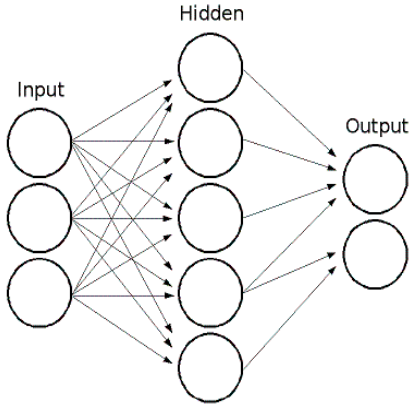

Fig 2. Neural Network

Neural networks usually involve a large number of processors operating in parallel and arranged in tiers. The first tier receives raw input information analogous to optic nerves in human visual processing. Each successive tier receive the output from the tier preceding it, rather than from the raw input in the same way neurons further from the optic nerve receive signals from those closer to it. The last tier produces the output of the system. The tiers are highly interconnected, which means each node in tier n will be connected to many nodes in tier n-1 its input and in tier n+1, which provide input for those nodes. There may be one or multiple nodes in the output layer.

Neural networks are notable for being adaptive, its means they modify themselves as they learn from initial training and subsequent runs provide more information about the world. The most basic learning model is too centred on weighting the input streams, which is how each node weights the importance of input from each of its predecessors.

The mathematical process through this network achieves *"learning"* can be principally ignored by the final user. In this way, the network can be viewed as a *"black box"* that receives a vector with *m* inputs and provides a vector with *n* outputs.

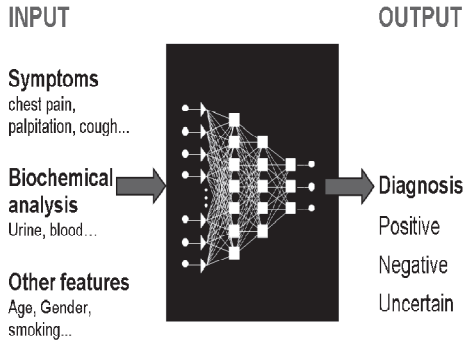

Fig 3.Details of input and output item concerning ANNs-based diagnosis (ANN architecture is often hidden and it is indicated here as a black box).

This method works relatively fast and enables accurate unit.

Structure of the training database

As stated above, the network must be trained by using a suitable database. The database is a table (or *matrix*) of data concerning patients for whom the diagnosis (Positive or negative) about a certain disease is already known. Each row of the matrix is referring to one patient. The first *m* elements of the row are medical data and the last *n* elements represent the output (*diagnosis*). The term *"medical data"* indicates the biochemical, nuclear magnetic resonance (NMR), laboratory data, and symptoms and other information provided by the medical specialist. An example of such training matrix with one output variable ($n = 1$) that may can assume two possible values (*positive* or *negative*).

Patient code	MEDICAL DATA	DIAGNOSIS
1	$data_{1,1}$ … $data_{1,i}$ … $data_{1,m}$	POSITIVE
2	$data_{2,1}$ … $data_{2,i}$ … $data_{2,m}$	POSITIVE
3	$data_{3,1}$ … $data_{3,i}$ … $data_{3,m}$	POSITIVE
…	……	
k	$data_{k,1}$ … $data_{k,i}$ … $data_{k,m}$	NEGATIVE
k+1	$data_{k+1,1}$ … $data_{k+1,i}$ … $data_{k+1,m}$	NEGATIVE
…	……	
n	$data_{n,1}$ … $data_{n,i}$ … $data_{n,m}$	NEGATIVE

Fig 4.Ex of training database structure. Each row refers to a different patient labelled with a numerical code. The element data *k, i* refer to the i^{-th} medical data of the k^{-th} patient.

Building the database

The neural network is trained to using a suitable database of "example" cases. An "example" is provided by one patient whose values for the selected features have been collected and evaluated. The quality of training the data and the resultant generalization, and therefore the prediction ability to the network, strongly depend on the database used for the training. The database should contains a sufficient number of reliable "examples" (for which the diagnosis is known) to allow the network to learn by extracting the structure hidden in the dataset and then use this "knowledge" to "generalize" the rule to new cases. In addition, the clinical laboratory data should be in a form that is readily transferable to programs for computer-aided diagnosis.

Data cleaning and preprocessing

Data in the training database must be preprocessed before its evaluation by the neural network. Several approaches are available for this purpose. Data are normally scaled to lie within the interval [0, 1] because its most commonly used to transference function is the socalled logistic one. In addition, it has been demonstrated that cases for which some data are missing should be removed from the database to improve the classification performance of the network. A decrease in the classification performance of the network is observed for imbalanced databases.

GPU

A graphics processing unit (GPU), is also called visual processing unit (VPU), is a specialized electronic circuit designed to rapidly manipulate and alter memory to accelerate the creation of images in a frame buffer intended the output to a display device. GPUs are used in embedded systems, mobile phones, personal computers, workstations, and game consoles. Modern GPUs are very efficient at manipulating computer graphics, image processing, and their highly parallel structure makes them more efficient than general-purpose CPUs for algorithms where the processing of large blocks of data is done in parallel. It is also a single-chip processor primarily used to manage and boost the performance of video and graphics. GPU features include:

- 2-D or 3-D graphics
- Digital output to a flat panel display monitors
- Texture mapping
- Application supports for high-intensity graphics software such as AutoCAD
- Rendering polygons
- Support for YUV color space
- Hardware overlay
- MPEG decoding

These features are designed to less the work of the CPU and produce faster video and graphics.

A GPU is not only used in a PC on a video card or motherboard; it's also used in mobile phones, display adapters, workstations and game consoles.

V. MODULE

MODULE 1: Document collection and preprocessing

- Data Collection
  Collecting the data sets of patient molecular levels and their gene for detecting database.
- Data Preprocessing
  Predict the data which are exceeding the normal level and also predict the genomic type for finding possibilities of diseases.

MODULE 2: Implementation of Neural Networks

- Technique
  -Neural networks.
  -It is a machine learning technique.
- Advantage
  -Train and test subsets

MODULE 3: Implementation of Neural Networks using cuda with R

-Neural network implementation.

-Cuda with R will be used to reduce executing time for large set of data.

MODULE 4: Performance Analysis and Documentation

-Finding overall accuracy for subtype classification.

-Comparison on CPU with GPU for finding accuracy and execution time.

VI. FUTURE ASPECTS

The future work is to improve the execution time faster for that we are using GPU (Graphical Processing unit). Type 2 diabetes patients with phenotypically similar but genotypically and molecularly dissimilar diseases. Predictive diagnosis for Subtype 1 was characterized by T2D complication diabetic nephropathy and retinopathy; Subtype 2 is cancer malignancy and cardiovascular diseases and Subtype 3 is most strongly with cardiovascular disease, neurological diseases, allergies and HIV infection. Based on that individuals or given with a different dosage.

VII. CONCLUSION

The study has lead to the conclusion that data mining allows the use of different approaches to build different models based on the data and objectives. Data mining method can be predictive. These techniques can be applied to extract information from patient to store data record. The prediction can be used to predict the data which are exceeding the normal level and also remedies for type 2 diabetes patients with phenotypically similar but genotypically and molecularly dissimilar diseases. We use precision medicine approach to characterize the T2D patient. The efficient method for T2D is neural network algorithm. To train the data is fast and enables accurate unit.

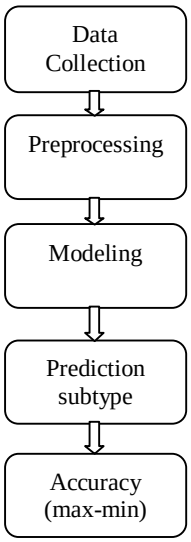

Fig 5. The architecture diagram of the proposed system

REFERENCES

[1] Santin I, Eizirik DL. Candidate genes for type 1 diabetes modulate pancreatic islet inflammation and beta cell apoptosis. Diabetes obesmetab. 2013; 15(suppl3):71-81. [PubMed]

[2] Abbas S, Raza ST, Ahmed F, et al. Association of genetic polymorphism of ACE, MTHFR, FABP-2 and FTO genes in risk prediction of type 2 diabetes mellitus. J Biomed Sci. 2013; 20(1): 80. [PMC free article] [PubMed]

[3] Pawlyk AC, Giacomini KM, McKenon C, et al. Metformin pharmacogenomics: current status and future directions. Diabetes. 2014; 63(8): 2590-2599. [PMC free article] [PubMed]