RESEARCH ARTICLE OPEN ACCESS

Application of Machine Learning Technique for Disease Cluster Detection and Functional Characterization

Mahalakshmi
*(Master of Computer Application, Visvesvaraya Technological University, and Gulbarga
Email: Mahalakshmikhemankar@gmail.com)

--************************----------------------------------

Abstract:

Human disease have been made disclose with the development of the biomolecular process. It is valuable in prevention, diagnosis and treatment. It reveals common functional changes that accompany a particular disease and recommend treatment that can be proper for the disease. This inadequacy can be potentially overcome by looking for common biological processes rather than only specified gene matches between diseases. In this work we uses the three types of disease heart disease, diabetes disease and COVID disease. The use of connection between biological processes to estimate disease similarity could enhance the identification and characterization of disease similarity. In this work we maintain a database were we store the number of records which we make the matches to the trained dataset .Machine learning make a contribution were we can predict the result in a low cost. In this project we are using the two algorithm linear regression for comparison and random forest algorithm which helps to predict the detection of disease and get the output. Here we take the parameters for predicting and make the future result of disease. It is beneficial to make a usage of this particular attribute to save a time and it's a low cost. We can say this is one of the best algorithms to predict the disease detection as we are getting 51% of accuracy.

--************************----------------------------------

I. INTRODUCTION

Cluster Disease Can Be Impart From One Generation To Another Generation. This Pattern Contributes To Building The Relation Between Phenotypes And Genes. The Number Of Disease We Have Seen In Daily Life From That We Are Using Here The Different Kinds Of Disease Which Can Make A Prediction. We Are Using The Heart Disease, Diabetes Disease And Covid Were We Uses The Parameters To Predict The Particular Outcome For Particular Disease. The Interaction Among Different Disease Has Been Discovered By Clinical Cases. In This Work We Maintain A Database Through Which Can Predict The Disease Availability To The Certain Age With The New Data. As We Have Used Here The 2 Libraries Sklearn And Pandas There Is A Some Built In Function To Perform The Prediction Pandas We Want This Libraries. Sklearn Is Used To Retrieve The Various Data Is Mainly Used For Statistical Calculation. Now, We Upload The Data What We Entered The Data And The Trained Data For Matching That Data We Want The Sklearn Package. The Used Algorithm In This We Have A Decision Tree ,Linear Regression,Random Forest.Here We Have A Number Of Disease Parametes Through Which We Can Identify The Prediction Of Whether The Patients Is Having A Getting Chances Of Disease Or Not We Measure Through The Trained Dataset Which We Using In This.

II. PAGE LAYOUT

In[1] they have presented a systematic approach to evaluate the benefit of using an intellectual approach for inter disease similarity. they have measured to represent the similarity between terms in an ontology that combines the information from occurrence in annotation and intellectual hierarchy there performance is compared to well known similarity using a subset of KEGG pathways as the standard the performance of a gene – based approach to disease similarity with a biological-process based by using a predefined set of disease as the standard. it is important to measure a similarity between the accuracy prediction. .they have see some multiple ways to classify the entities such as anatomy, symptoms and etiology. This can result in different structure connected. A similarity metric can therefore yield different estimates for the similarity of a pair of entities based on the particular ontology used. Further, intellectual are often a work in progress with terms and relationships both added and deleted over time (e.g., GO has grown in size by an order of magnitude in the decade since its inception). Thus, similarity metrics that are based on the blood relatives of two terms in the ontology will, by definition, be unable to capture similarity between terms that are related but topologically far apart in the intellectual. Such methods are also sensitive to errors in theoretical. Hybrid methods which use graph structure and node information are also inadequate as they will fail to capture similarity between terms that are far apart in the ontology.

III. PAGE STYLE

In[2]they said about the exomes sequence it is a promising method for diagnosis with a complex phenotype.There will some challenging scenario for the patient phenotype .they reveals many possibly damaging variants were have individually assessed to clear association with patients phenotype. the algorithm order genes by the same computed between phenotype descriptor with each gene and those describing by the patient. It is necessary to correlate a mutated gene to the patient phenotype to reach a clinical diagnosis.

TABLE I

Processor	Pentium IV 2GHz and Above
RAM	4 GB RAM
Monitor	15” Color Monitor
Keyboard	
Mouse	

Table1: Hardware Requirements

System	Windows /Linux
Technology	Django 3.0(MVC)
Database	SqLite
Coding Language	Python
Front End Tool	Sublime Text Operating

Table 2 : Software Requirement

OBJECTIVE OF THE STUDY

The most simple and direct approach is to cluster diseases based on their phenotypic similarity through clinical observation.

To collect the data for observing the disease percentage chances.

To collect the data for predicting the disease accuracy.

To propose a model which predicts the result for disease detection

1.4 SCOPE OF THE STUDY

There are numerous applications of Machine Learning. the trained dataset is our creation in which Predicting diseases perhaps one of the most promising areas of testing the accuracy of Machine Learning outcomes. Here, the usage of linear regression and random forest algorithm in which it provides the accuracy .

1.5 METHODOLOGY USED

REGRESSION TREE

The decision tree is a known practical formation for supervised learning .It is used for both classification and regression appraisal. The decision tree is a tree-structured classifier that consists of three types of nodes such as the root node, interior node, and leaf node. The root node is the initial node that represents the whole illustrative. The interior nodes represent the characteristics of a data set. Lastly, the root nodes provide the prediction. For a particular data point, the decision tree is run by answering true/false questions until they reach the leaf node. The final result is calculated by finding the average value of a dependent variable in a specific leaf node. In this way, the tree can predict a proper value for the data point through number of iterations. The decision tree is a useful because it is simple to understand and requires less data cleaning. Like ridge and lasso regression, decision tree regression may have over fitting problems. An group of decision trees (e.g., the RF algorithm) can overcome these problems.

RANDOM FOREST

Random forest is an grouped algorithm that builds a set of independent and non-identical decision trees following the idea of randomization .This algorithm is used for both classification and regression purposes, and it is a combination of tree predictors. Each decision tree employs a random vector as a parameter randomly chooses the attributes of samples, and it then finally chooses the sample subset as the training dataset .The generalization error of a forest of trees depends on the forest's individual trees strength and correlation. However, deep decision trees might suffer from over fitting RF prevents over fitting by generating random subsets of attributes and constructing trees using these subsets .

2.LITERATURE SURVEY

In[3] the number of large scale is used to defined connection between genes and proteins in various species. Some attempts have been made to classify them in connection at the phenotype level. It is unknown whether they carry any biological meaning to it or not. they used the text mining to classify over 5000 human phenotypes contained in OMIM database. they find the gene which is related. they have same positive correlation with a number of measure of gene function which relates the level of protein sequence, protein motifs ,functional annotation and direct PPI.

In [4] Genetic factors are strongly affect sensitivity to common diseases and also determine disease-related numerical feature. Identifying the applicable genes has been difficult, in part because each causal gene only makes a small contribution to overall innate. Genetic association studies offer a potentially powerful approach for depict causal genes with adequate effects, but are limited because only a small number of genes can be studied at a time.

"Construction of a genetic linkage map in man using restriction

fragment length polymorphisms," American journal of human

In[5] The work of the OMIM is derived the entire produce from the biomedical literature and its modernize it daily.the currently it contains around 18,961 entries explained about phenotype and genes.2239 genes have been newly causing disease. With the examining the genes ,the focus of OMIM.With the addition of kind of genes, connection of phenotype and genes. Over the many years, systemize for Y-linked and organelle phenotypes and genes were added and the text had been changed to 'A Catalog of Human Genes and Genetic Disorders. The same basic cooperative has been continued to grow to include traits of the consequence of difference and repetitive and deletions/microdeletions and duplications/microduplications.

In[6]The patients with diabetes mellitus (DM)account for around one-quarter of all patients who undertake of relating to pericardial procedures each year, and they experience worse outcomes compared with non-diabetic patients. The clinical trials of percutaneous transluminal coronary angioplasty (PTCA) versus coronary artery bypass grafting (CABG) that included diabetic patients have been reviewed.

In[7]The diabetic outmost neuropathy is insufficient treated ,the role of improving body mass control initially in type-2 diabetes remains unclear.the international clinic guidelines has been recommended many treatments.therapy includes tricyclic antidepressants, First-line therapies include tricyclic antidepressants, serotonin–noradrenaline reuptake inhibitors, and anticonvulsants that act on calcium channels. Other remedy include narcotics and topical agents such as capsaicin and lidocaine. The study widespread are to review current guidelines for the pharmacological management of DPN and to study relevant to the further development of pharmacological recommendations for the treatment of diabetic neuropathy. The Diabetic neuropathy is a highly prevalent, disabling condition, the management of which is associated with importance costs. Evidence supports the use of specific anticonvulsants and antidepressants for pain management in patients with diabetic peripheral neuropathy.

In[8] A number of large-scale efforts are underway to define the relationships between genes and proteins in various species. But, few attempts have been made to systematically classify all such relationships at the phenotype level. Also, it is unknown whether such a phenotype map would carry biologically meaningful information.

In[9]In the study of genetic chemical science and molecular genomics have raised our knowledge in the basic element of human biology and disease .At the same time the more importance of the networks in between those biological components is building the skills.In recent the technologies and advance summary,a new way called the network medicine,an approach to know the human disease from the network point-of view is about to appear

In[10] This research has been made an deep learning approach which integrates Bagging Ridge (BR) regression with Bi-directional Long Short-Term Memory (Bi-LSTM) neural networks used as base regressors to become a Bi-LSTM BR approach. Bi-LSTM BR was used to divine the exchange rates of 21 currencies against the USD during the pre-COVID-19 and COVID-19 periods. To reveal the effectiveness of our proposed model, they predict the performance based on several traditional machine algorithm ,such as the regression tree, support vector regression, and random forest regression, and deep learning-based algorithms such as LSTM and Bi-LSTM. Our their work ensemble deep learning approach outperformed the compared models in forecasting exchange rates in terms of prediction error. However, the performance of the model significantly different during non-COVID-19 and COVID-19 periods across currencies, indicating the important role of prediction models in periods of highly volatile foreign currency markets. By providing an update prediction performance and identifying the most seriously affected currencies, this reearch is useful for foreign exchange traders and other stakeholders in that it offers opportunities for potential trading profitability and for reducing the impact of increased currency risk during the epidemic.

In[11] to describe a new basis for the construction of a genetic connecting map of the human genome. The basic principle of the mapping scheme is to develop, by reintegration DNA techniques, random single-copy DNA probes capable of detecting DNA sequence polymorphisms, when hybridized to restriction digests of an individual's DNA. Each of these probes will define a locus. Loci can be expanded or diminish to include more or less polymorphism by further application of recombinant DNA technology. Suitably polymorphic loci can be tested for linkage relationships in human pedigrees by established methods; and loci can be arranged into linkage groups to form a true genetic map of "DNA marker loci." Pedigrees in which inherited traits are known to be segregating can then be analyzed, making possible the mapping of the gene(s) responsible for the trait with respect to the DNA marker loci, without requiring direct access to a specified gene's DNA. For inherited diseases mapped in this way, linked DNA marker loci can be used predictively for genetic counseling.

In[12] The intention of these studies is to feature the development of pandemic in certain regions and, thus, to provide for the requirements acquire to contain the impurity by virus and allocation of resources. a mathematical model based on SEIR model (Susceptible - Exposed - Infectious - Recovered) for forecasting the transferal change of Covid-19 in Korea, is proposed. This study is able to predict the final size and the timing of the end of outbreak as well as the maximum number of accessible individuals using daily confirmed cases comparing epidemiological parameters between the national level and the Daegu/Gyeongbuk area. the role of asymptomatic carriers in transmission poses challenges for control of the Covid-19 pandemic, is labeled.

2.1EXISTING AND PROPOSED SYSTEM

EXISTING SYSTEM

In existing system how we analyze the system .If we suppose we have 100 records and 60 peoples suffering from diabetes 20 suffering from like that we do analyze and will take individual person to interact with the data and verifying concept we are not using. The main missing part in this is taking the information and checking with our database .So this is the main existing thing which is not there in this we don’t need analyze part .one individual person information through him .we collect the information and we check the response which is not there in the existing system here we are implementing .

PROPOSED SYSTEM

In the proposed system, the persons data basic parameters like symptoms of different disease will take all the basic information of the patient will put in our maintained database . we compare the trained dataset and with the person data we compare and make the prediction with the particular disease. In this work the trained database we compare with the diabetes disease were it has the basic parameters like sugar and blood pressure and sugar level of the person .we are some other parameters through which we can detect the data and make the prediction and we can identify whether the person is have if the disease or not and how much percentage chances of getting disease. Comparing to the heart and covid it has particular basic parameter with which we identify the percentage of getting the disease or not. The graph analyze we predict through and 51% of accuracy.

2.2FEASIBILITY STUDY

Feasibility depends on the outcome of the result investigation overiew is expanded to make more research on the study. It is a test system to propose according to its work which impacts on the coordination,potential to meet the requirement and being constructive,on the used of weapon.

During feasibility analysis for this weapon, following small areas of absorption are to be value. Exploring and generating ideas about a new system does this. Steps in feasibility analysis eight steps involved in the feasibility analysis are:

Form a project team and appoint a project leader.
Prepare system flowcharts.
List potential proposed system.
Define and identify characteristics of proposed system.
Determine and examining performance and cost effective of each proposed system.
Weight system performance and cost data.
Select the best-proposed system.
Prepare and report final project directive to management.

A study of weapon availability that may affect the ability to achieve an acceptable System. This inquiry determines whether the technology needed for the proposed System is available or not.

This is concerned with specifying tools and software that will successfully satisfy the user requirement. The technical needs of the system may include:

Front - end and back-end selection

An important issue for the development of a project is the selection of suitable front-end and back-end. When we decided to develop the project we went through an considerable Study to determine the most suitable platform that suits the needs of the organization as Well as helps in development of the project.

The aspects of our study included the following factors.

Front-end selection:

1. It must have a graphical user interface that assists employees that are not from IT Background.
2. Scalability and extensibility.
3. Flexibility.
4. Robustness.
5. According to the organization requirement and the culture.
6. Must provide excellent reporting features with good printing support.
7. Platform independent.
8. Easy to debug and maintain.
9. Event driven programming facility.
10. Front end must support some popular back end like SQL.

According to the above stated features we selected JAVA as the front-end for
developing our project.

Back-end Selection:

1. Multiple user support.
2. Efficient data handling.
3. Provide innate features for security.
4. Efficient data recover and maintenance.
5. Stored procedures.
6. Popularity.
7. Operating System cooperative.
8. Easy to install.

9. Various drivers must be available.
10. Easy to implant with the Front-end.

According to above stated features we selected SQL SERVER as the backend.

The technical feasibility is frequently the most difficult area experience at this stage. It Is essential that the process of analysis and definition be conducted in parallel with an Assessment to technical feasibility. It centers on the existing computer system (Hardware, software etc.) And to what extent it can support the proposed system.

The following are the major factors to consider while doing a feasibility analysis:

TECHNICAL FEASIBILITY

Economic Feasibility:- The proposed software will save lots of paper work and easy captivating record keeping there by reducing the costs sustain on above heads. This reduction in cost cause the hospital to go for such computer-based system.

Operational Feasibility:- User-friendly: Users can register themself easily and perform login operation and can check for the result with a simple joining by provided minimal information.

Reliability: The package wills pick-up current transaction online. Regarding the old transaction, User will enter them in to the system and with the help of dataset.

Security: Application is built on Python which has very high security and MVC architecture.

Availability: This software will be available always.

Technical Feasibility - As the saying goes, "to error is human". Keeping in view the above fact, now a day all company are automating the repetitive and monotonous works done by humans. The key process areas of current system are nicely manageable to voluntary and hence the technical feasibility is proved beyond doubt.

Organizational Feasibility - Organizational feasibility aims to assess the effectiveness of management and limitation of resources to bring a product. The company should evaluate the ability of its management team on areas of interest and execution. Typical measures of management include appraise the founders' passion for the business idea along with industry expertise, educational background, and professional experience. Founders should be honest in their self-assessment of ranking these areas.

CONCLUSIONS

The version of this template is V2. Most of the formatting instructions in this document have been compiled by Causal Productions from the IEEE LaTeX style files. Causal Productions offers both A4 templates and US Letter templates for LaTeX and Microsoft Word. The LaTeX templates depend on the official IEEEtran.cls and IEEEtran.bst files, whereas the Microsoft Word templates are self-contained. Causal Productions has used its best efforts to ensure that the templates have the same appearance.

Causal Productions permits the distribution and revision of these templates on the condition that Causal Productions is credited in the revised template as follows: "original version of this template was provided by courtesy of Causal Productions (www.causalproductions.com)".

ACKNOWLEDGMENT

The good scores can be achievable by using the machine learning it can have an improvement. The addition of more data which helps to algorithm to know more bitterly. Overactive optimization is another method of tuning the availables to get the better result on the performance were the data set has been provided. It can be implemented using the Scikit-learn machine learning library. The two famous algorithms which can be used are:

1) Grid Search: In this method, a search space as a surrounded department of active values is defined then random points are sampled within the surrounded authority.

2) Random Search: In this method, a search space as a surrounded department of available values is defined, and then random points are sampled in the surrounded areas.

The future proposed can be building a machine learning application were the user can use a stock of dataset were they can get a aproporiate result with the more accuracy .the machine should take a proper dataset correctly and choose the algorithm which gives the low error rate.the result can be printed on the window screen.the user can interact easily and make a user friendly for the learners.this application can be can be deployed on the servers Heruko in which the user can see the model in action .

REFERENCES

[1] [1]Sachin. Mathur, et al.,"Finding disease similarity based on implicit semantic similarity" Journal of Biomedical Informatics 45 (2012) 363–371 ,2011.

[2] [2] Aaron J Masino1 , Elizabeth T Dechene "Clinical phenotype-based gene prioritization: an initial study using semantic similarity and the human phenotype ontology" BMC Bioinformatics 2014.

[3] [3] Marc A van Driel et. al.,"Human phenome analysis European Journal of Human Genetics (2006) 14, 535–542,2006

[4] [4]J. N. Hirschhorn and M. J. Daly, "Genome-wide association studies for common diseases and complex traits," Nature reviews genetics, vol. 6, pp. 95-108, 2005.

[5] [5] J. Amberger, C. A. Bocchini, A. F. Scott, and A. Hamosh, "McKusick's online Mendelian inheritance in man (OMIM®)," Nucleic acids research, vol. 37, pp. D793-D796, 2009.

[6] In[6] Colin Berry, MD, PHD, Jean-Claude Tardif, MD, FACC, Martial G. Bourassa, MD, FACC Coronary Heart Disease in Patients With Diabetes Vol. 49, No. 6, 2007

[7] In[7] Maher R. Khdour "Treatment of diabetic peripheral neuropathy: a review" , pp. 863–87272 (2020),

[8] In[8] Marc A van Driel1 , Jorn Bruggeman2 , Gert Vriend1 , Han G Brunner*,3 and Jack AM Leunissen"A text-mining analysis of the human phenome," 14, 535–542 , 2006

[9] In[9] Kwang-Il Goh and In-Geol Choi Exploring the human diseasome: the human disease network . VOL 11. NO 6. 533-542,2012

[10] In [10]Mohammad Zoynul Abedin et.al Deep learning-based exchange rate prediction during the COVID-19 pandemic, 2021

[11] [11] DAVID BOTSTEIN et. al.Construction of a Genetic Linkage Map in Man Using Restriction Fragment Length Polymorphisms, Am JHum Genet 32:314-331, 1980

[12] [12] Daiana Caroline dos Santos Gomes" Machine Learning Model for Computational Tracking and Forecasting the COVID-19 Dynamic Propagation" S, VOL. 25, NO. 3,2021.