International Journal of Innovative Technology and Exploring Engineering (IJITEE)
ISSN: 2278-3075 (Online), Volume-9 Issue-5, March 2020

Diabetes and its Complication Prediction using Multi-Task Learning

Shubharthi Dey, Bagish Choudhury, S. Sharanya

Abstract: ***Diabetes is a long-term disease that ends up in multiple side-effects. It has now become a reticent exterminator in society because it doesn't reveal any signs hitherto to the patients until it's too late. It leads to many complications to other organs, such as kidney, cardiovascular, liver or blood pressure [1]. This work tends to apply a unique multitask learning [2] to synchronously map the relation between manifold complications wherever every task conforms to risks of modelling of complications [3]. It also uses feature selection to reduce the set of risk factors from high-dimensional datasets. Then using the concept of correlation, it finds the degree of relativity among various side-effects. The proposed method is able to identify the possible future health hazards identified with the diabetes patient. This will enable us to explain medical conditions and can improves healthcare applications which would help to improve disease prediction performance.***

Index Terms: Diabetes Risk; Feature Selection; Healthcare; Multitask Learning

I. INTRODUCTION

Diabetes Mellitus is a sickness that influences almost a large portion of the world. The most commonly perceived typecast is the Type 2 Diabetes and it embodies mass of the cases. It is outlined by hyperglycemia— anomalously risen blood glucose levels and quite often connected with various intricacies. After some time, blood vessel impair that is prompted by the interminable rise of blood glucose levels, which thus prompts related complications including kidney ailments visual impairment stroke coronary episode and in extreme cases also demise. Concurrently, over the previous decades the debit of diabetes care has been expanding ex-peditiously. T2DM handling obligates consistent medicinal consideration with techniques outside glycemic ability to control. Gist of T2DM handling is placating the peril of complications. Numerous investigations have been directed to pick up information related to hazard elements and the determination of its side- effects. Be that as it may not many examinations have been directed to assess the diabetes com-plexity infections particularly its hazard factors.

Revised Manuscript Received on March 30, 2020.
* Correspondence Author
Shubharthi Dey*, Department of Computer Science and Engineering Department, SRM Institute of Science and Technology, Kattankulathur, Chennai, India.
Bagish Choudhury, Department of Computer Science and Engineering Department, SRM, Institute of Science and Technology, Kattankulathur, Chennai, India.
S. Sharanya, Department of Computer Science and Engineering Department, SRM Institute of Science and Technology, Kattankulathur, Chennai, India

© The Authors. Published by Blue Eyes Intelligence Engineering and Sciences Publication (BEIESP). This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/)

From one viewpoint T2DM complications incorporate serious infections for example renal impairment and coronary failure and thus involve costly medicinal strategies. Additionally, doctoring of its complexities involves a majority of its overall expense. Framing of risk of T2DM complexities is basic for healthcare practitioners to suitably adjust customized healing plans for sufferers and lessening cost. There is a newfangled chance to apply prescient investigation to polish this process owing to the fresh plethora of electronic health records and electronic pharmaceutical claims. It can be employed for an inclusive scope of medicinal services applications such as sickness beginning forecast, disease progression tolerant stratification, hospital readmission forecast and mortality expectation.

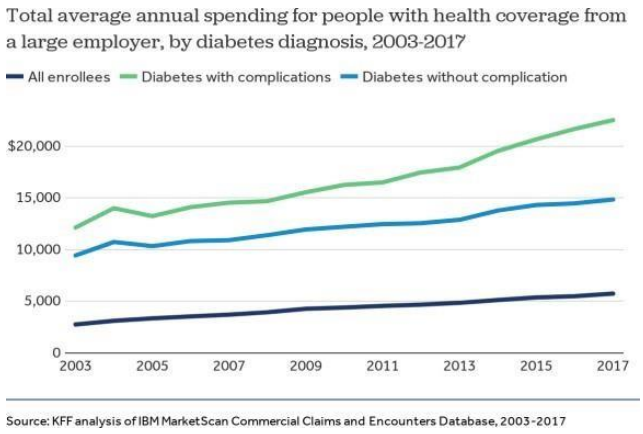


Fig. 1. Expenditure of Diabetic Patients

This paper has the following flow of discussion: section II is dedicated to the necessary related work of prediction strategies and section III illuminates the intended system and section IV gives the dissection results while section V gives the conclusions and prospective works.

II. RELATED WORKS

The authors in [4] use an approach supported by the com-bined use of a genetic algorithmic program (GA) and a nearest neighbour's classifier for the choice of the features that are powerfully associated with the incidence of fatal and nonfatal disorder (CVD) in patients with T2DM. Skevofilakas [5] have devised a Decision Support System (DSS) together with digital hospital record to foresee retinopathy risk. It is a structure consisting of a Rule Clas- sifier, Neural Network and a Decision tree and many im- provements have been developed to surmise the prospects of developing the corollary.

The role of Adaboost and conduct escalating methods discussed in [6] using J48 decision tree as the basis for classifying the diseases and concerned sufferers, based on risk perils.
Results achieved after the experiment proves that, Adaboost technique outclasses the conduct of any regular decision tree. Enforcement of the approach by the authors in [7], classifies the chances of contracting diabetes. To fulfill the intent, authors have employed four following renowned methods: Neural Networks, Regression, Decision tree and Naive Bayes. For improving the cogency of designed model conduct escalating aptitudes are used.
A scheme for preventing possible risk situations at DM- II patients through continuous glucose monitoring and subsequent prediction-based tele consults modelling was presented by the author of [8]. The subsequent output goes to a continuous observing framework so as to identify and update those situations where patients would need quick support. The results of simulations have indicated that drug therapy, adequate diet, and opportune tele consults might be crucial to enclose glucose values within permissible ranges. Prof. M. A. Pradhan and his team [9]. conceived a familial scheduling-based classifier blueprint that will help the health- care professionals to reaffirm the concern's diagnosis towards the climax of the malady. The devised system promises to give quicker and better yields.

III. PROPOSED WORK

A. Overview

The work is based on multi-task learning (MTL), which means to mutually become familiar with various undertakings utilizing a common portrayal with the intent that counsel acquired from one task. As of late, MTL archetypes have been generally utilized in the medicinal services area, for example, disease amelioration, lethality forecast, risk profiling, elderly patient consideration. MTFL and MTRL are the two most broadly utilized MTL techniques. MTFL hypothesizes that task affiliation is discharged through. The fundamental thought of MTFL approaches is to get familiar with a couple of tasks common over the assignments utilizing diverse sparsity procedures. MTRL hypothesizes that the task affiliation is uncovered in the mold of the collegial network. Proposed method is a build-up on these existing methods, aiming to further increase the efficiency of these method- ologies by boosting their performance using various associ- ated concepts such as using feature selection, the dataset is trimmed leading to an even better prediction, as irrelevant and negligible factors are removed. Furthermore, the Correlation Coefficient is used to find the degree of relativity among the various factors and this degree is used in the final prediction. As it happens, our adduced method is great for wellness pro- grams where we do not just get better prediction exhibitions, yet additionally infer clinically significant bits of knowledge about the connections amidst the various complications and also, among the diverse features.

B. Feature Selection

Patient medicinal data is usually multi-dimensional with many applicable features. We are keen on distinguishing an instructive subspace of coefficients, which mirror the confer- ring features. Filter based feature selection has become critical in numerous characterization settings, recently confronted with including learning techniques that start a huge number of cues. We use Infinite Latent Feature Selection (ILFS) [10] method to trim the available dataset to a smaller, precise and accurate set of features that adequately define the future complexities of the disease. Considering an array of factors as a path among feature conveyances and letting these tend out for an infinite number allows the examination of the significance (relevance) into a subjective arrangement of cues. Ranking the significance individuates applicant features [11], which end up being viable from a characterization perspective.

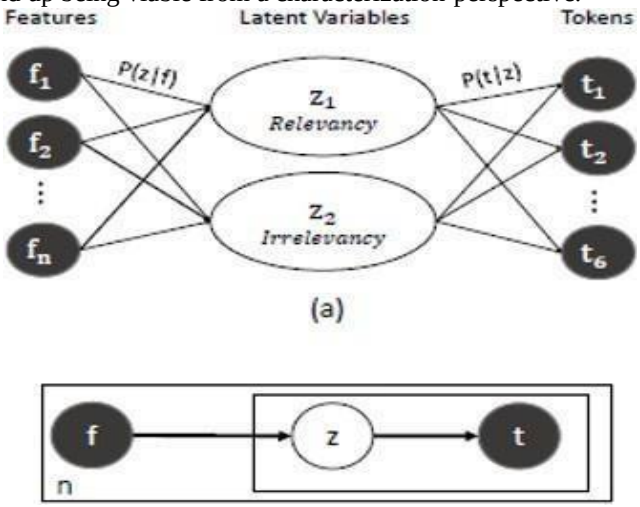


Fig. 2. The intermediate layer that links the features and the tokens

C. Correlation Analysis

Correlation is a factual caliper that computes the quality of the connection among the overall developments of two factors. The qualities mediate between 1.0 and 1.0. Pearsons relationship coefficient is the covariance of the two compo- nents isolated by the aftereffect of their standard deviations. The sort of the definition incorporates an item minute that is the mean the primary moment about the reason for the conse- quence of the mean-adjusted unpredictable components; thus, the modifier thing minute in the name. Pearsons correlation is utilized to quantify whether two informational collection are in a line which is utilized to gauge the separation of direct connections between factors. When two factors are typical nonstop factors and there is a direct injunction among them their connection is communicated by item minute connection. This approach is used for finding any reciprocity mediating the occurrence of various side effects of diabetes.

It gives insight into whether there exists any reciprocity between the anticipation of these complications whether one influences the occurrence of another. Pearson correlation 'r' is given by,

$$r = \frac{N\sum x_i y_i - \sum x_i \sum y_i}{\sqrt{N\sum x_i^2 - (\sum x_i)^2}\sqrt{N\sum y_i^2 - (\sum y_i)^2}}$$

Where 'x' and 'y' are the two variables under consideration. The degree of reciprocity is defined based on the numerical

Correlation Coefficients	Correlative Degree
0.8-1.0	Strongest
0.6-0.8	Strong
0.4-0.6	Moderate
0.2-0.4	Weak
0.0-0.2	Weaker or none

Fig. 3. Table for the Correlated Coefficient and Correlative Degree

value obtained after engaging the x and y uncertain into the reciprocity formulae. Spearman's Correlation determines the relativity between two ranked. It calculates the mono- tonic injunction between two datasets unlike, Pearson sim- ulation which takes linear relationships under consideration. Spearman's formula calculates whether the two datasets are unswervingly proportional to each other. The Spearman rank injunction test doesn't convey any presumptions about the dis- persion of the known and is the fitting injunction examination when the agents are estimated on a horizon that is in any event ordinal. There are two strategies to ascertain Spearman's injunction relying upon whether:

- The Information doesn't have equal positions

$$\rho = 1 - \frac{6\sum d_i^2}{n(n^2-1)}$$

di= Discrepancy between the ranking of paired information
n= Number of cases

- The information has equal positions where i = paired score

$$\rho = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_i (x_i - \bar{x})^2 \sum_i (y_i - \bar{y})^2}}$$

D. Predictive Analysis

The next step in the process is the splitting of the dataset into training and testing data. 80% of the data is used to train the model whereas the rest of the 20% is used to test the model. Fi- nally, we use the existing Multitask learning (MTL) method to get the input i.e. the trimmed dataset and predict the probable complications associated with diabetes. We have already men- tioned about the advantages of using multi-learning approaches in prediction-based case studies, like ours. As of late, MTL models have been generally utilized in the medicinal services area, for example, disease progression, mortality forecast in intensive care, risk profiling, elderly patient consideration, and diabetes. The instinct behind MTL is that a joint learning strategy representing task connections is more productive than learning each undertaking independently. The MTL uses the training data to learn about the various covariances among the factors and then consequently uses the testing data to test its knowledge and predict the output with the help of the reciprocity coefficient.

IV. EXPERIMENTAL RESULTS

After the successful fabrication of the proposed method to test the virtuosity and output of the method. The method was validated using a model diabetic datahub easily procurable on the internet. The dataset contained clue whether the patient is diabetic or not. It has 11 factors and a total of 503 entries. Also, it contains additional instruction about the various side- effects linked with the advent of diabetes in a patient. Multiple side-effects and their risk factors are listed in this dataset and using these knowledge the future health hazards can be predicted. We test our proposed tactic using this available data. We then create a confusion grid to determine the success of the method. The efficiency grades that are under consideration are Precision and Recall.

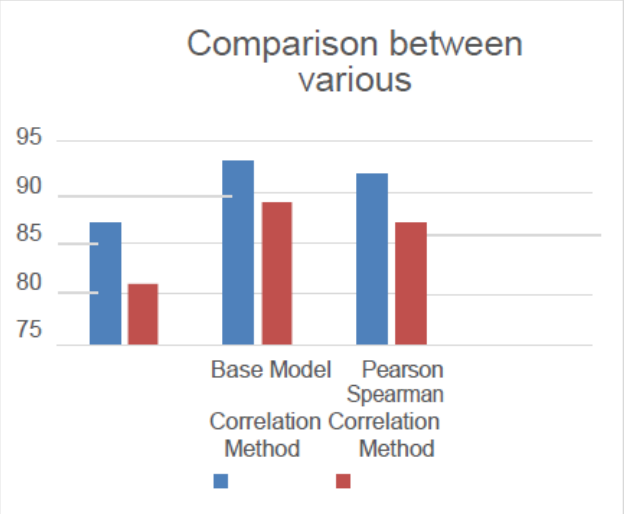
Exactness alludes to adjacency of the estimations to a particular worthwhile precision alludes to the adjacency of the estimations to one another.

$$precision = \frac{true\ positives}{true\ positives + false\ positives}$$

True positives are variable points termed Bonafede by the system that genuinely are certain, and False negatives are variable points focuses the model distinguishes as wrong. Recall is described as a model's retention to discover all the material focal points in a dataset.

Recall is given by,

$$recall = \frac{true\ positives}{true\ positives + false\ negatives}$$



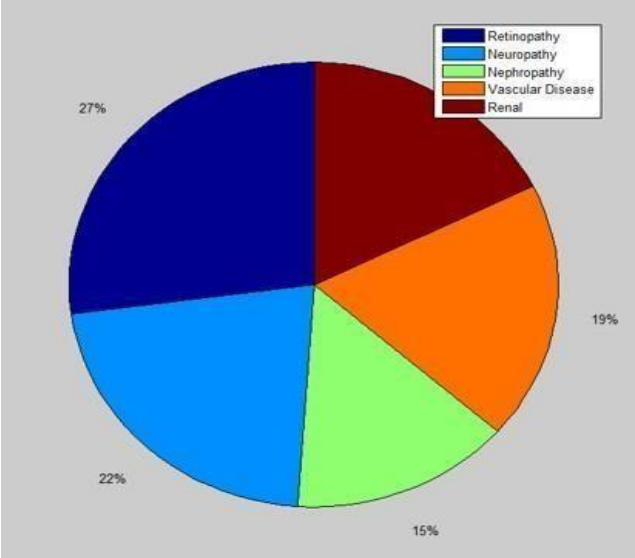


Fig. 4. Percentage of Diabetic Complications

From the exercises on the dataset we conclude that while a wide array of side-effects is associated with the advent of diabetes in a particular patient, but Retinopathy (27%) holds the top spot and poses the highest risk for the patients. It is then followed by Neuropathy (22%) and Vascular Diseases (19%). Next comes Nephropathy (15%) and lastly, we have Renal diseases (17%). The proposed method using Pearsons Correlation analysis was successfully able to predict the future side-effects of various patients with an accuracy of 93.2%, which is definitely a big gain on the traditional predicting approaches that only offer an accuracy of 87% approximately. Additionally, the proposed method using Spearman Correlation analysis was successfully able to predict the future side-effects of various patients with an accuracy of 91%. Recall was 89% for Pearson Correlation analysis whereas it was 87% for Spearmans. This proposed system can be used for the prediction and customisation of treatments for patients in healthcare.

V. CONCLUSION

Complication profiling has come a long way in the last few years. With emerging technologies, we have been able to use these new knowledges to gain more information about diseases, their symptoms and prevention. Our work involves using some existing methods in a new and innovative way to try and predict the advent of side-effects associated with diabetes. If left untreated, these complications can cause even more harm to the patient. Thus, timely detection and treatment is the best way to surely protect the patient from further distress. Our works has many limitations and also opens the door to many interesting future works. Some of the limitations include

- Many side-effects are unpredictable and may or may not surface in a given patient
- Better factor representation can help include the latent data associated with various side-effects

We are also intrigued to see how our proposed model can be utilised to predict complications for various other chronic diseases.

REFERENCES

1. J. M. Forbes and M. E. Cooper, “Mechanisms of diabetic complications,” Physiological reviews, vol. 93, no. 1, pp. 137–188,2013.
2. J. Zhou, L. Yuan, J. Liu, and J. Ye, “A multi-task learning formulation for predicting disease progression,” in Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining, pp. 814–822, 2011.
3. B. Liu, Y. Li, S. Ghosh, Z. Sun, K. Ng, and J. Hu, “Complication risk profiling in diabetes care: a Bayesian multi-task and feature relationship learning approach,” IEEE Transactions on Knowledge and Data Engineering, 2019.
4. L. Sacchi, A. Dagliati, D. Segagni, P. Leporati, L. Chiovato, and R. Bellazzi, “Improving risk-stratification of diabetes complications using temporal data mining,” in 2015 37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC) pp. 2131–2134, IEEE, 2015
5. K. V. Dalakleidi, K. Zarkogianni, V. G. Karamanos, A. C. Thanopoulou, and K. S. Nikita, “A hybrid genetic algorithm for the selection of the critical features for risk prediction of cardiovascular complications in type 2 diabetes patients,” in 13th IEEE International Conference on BioInformatics and Bioengineering, pp. 1–4, IEEE, 2013.
6. M. Skevofilakas, K. Zarkogianni, B. G. Karamanos, and K. S. Nikita, “A hybrid decision support system for the risk assessment of retinopathy development as a long-term complication of type 1 diabetes mellitus,” in 2010 Annual International Conference of the IEEE Engineering in Medicine and Biology, pp. 6713–6716, IEEE, 2010.
7. C. J. Steele, A. H. Marshall, A. Kouvonen, F. Kee, and R. Sund, “Modelling the time taken to experience a type 2 diabetes related complication using a survival tree in order to advise general practitioners,” in
8. 2016 IEEE 29th International Symposium on Computer-Based Medical Systems (CBMS), pp. 271–272, IEEE, 2016.
9. C. Fiarni, E. M. Sipayung, and S. Maemunah, “Analysis and prediction of diabetes complication disease using data mining algorithm,” Procedia Computer Science, vol. 161, pp. 449–457, 2019.
10. S. Perveen, M. Shahbaz, A. Guergachi, and K. Keshavjee, “Performance analysis of data mining classification techniques to predict diabetes,” Procedia Computer Science, vol. 82, pp. 115–121, 2016.
11. N. Nai-arun and R. Moungmai, “Comparison of classifiers for the risk of diabetes prediction,” Procedia Computer Science, vol. 69, pp. 132–142, 2015.

12. H. Nieto-Chaupis, M. Caballero, H. Matta-Solis, R. Perez-Siguas, S. Blas, E. Carranza-Manrique, E. Contreras, G. Quispe, S. Ramirez, and J. Rocha, "Preventing risk situations at type-ii diabetes mellitus patients through continuous glucose monitoring and prediction-based teleconsults," in 2015 IEEE 28th International Symposium on Computer- Based Medical Systems, pp. 27–28, IEEE, 2015.
13. G. Thangarasu and P. Dominic, "Prediction of hidden knowledge from clinical database using data mining techniques," in 2014 International Conference on Computer and Information Sciences (ICCOINS), pp. 1–5, IEEE, 2014.
14. M. Pradhan, G. Bamnote, V. Tribhuvan, K. Jadhav, V. Chabukswar, and V. Dhobale, "A genetic programming approach for detection of diabetes," International Journal of Computational Engineering Research, vol. 2, no. 6, pp. 91–94, 2012.
15. G. Roffo, S. Melzi, U. Castellani, and A. Vinciarelli, "Infinite latent feature selection: A probabilistic latent graph-based ranking approach," in Proceedings of the IEEE International Conference on Computer Vision, pp. 1398–1406, 2017.
16. G. Roffo and S. Melzi, "Ranking to learn," in International Workshop onNew Frontiers in Mining Complex Patterns, pp. 19–35, Springer, 2016

AUTHORS PROFILE

Shubharthi Dey, is currently pursuing his Bachelor's degree in Computer Science and Engineering from SRM Institute of Science and Technology, Chennai, to be completed in the year 2020. His area of interest includes Artificial Intelligence, Web development and Machine learning and its various applications with healthcare and/or business analytics. He aims to study the various ways Machine learning can be used to further improve upon the existing approaches using real world evidence and various learning techniques. His recent focus has been on leading research efforts to develop advanced machine learning, data mining for deriving data-driven insights from real world healthcare data to facilitate learning health systems.

Bagish Choudhury, is currently pursuing his Bachelor's degree in Computer Science and Engineering from SRM Institute of Science and Technology, Chennai, to be completed in the year 2020. His research interest spans topics in Machine learning and its intersection with healthcare applications and/or business analytics. He aims to study the various ways Machine learning can be used to further improve upon the existing approaches using real world evidence and various learning techniques. His recent focus has been on leading research efforts to develop advanced machine learning, data mining and visual analytics methodologies for deriving data-driven insights from real world healthcare data to facilitate learning health systems. His current work primarily includes developing statistical and Machine learning methods to generate insights in the healthcare.

S. Sharanya, M.E., is working as Assistant Professor in the Department of Computer Science and Engineering at SRM Institute of Science and Technology, Chennai. She is a member of IEI and ISC, reviewer for peer reviewed journals and author of few technical books. She has completed her Master's degree in Computer Science and Engineering in the year 2010 and Bachelor's degree in the same discipline in the year 2008. She is now currently pursuing her Ph.D. in the area of Machine learning. Her recent focus has been on leading research efforts to deriving data-driven insights from real world healthcare data to facilitate learning health systems.