# Identification of the Most Important Physical and Pathological Factors for Diabetics Prediction Accurately at Early Stage

## Introduction

Diabetes is one of the most common diseases in every region of the world. In 2019, about 463 million people, 8.8% of the total population were affected by diabetes and it caused about 4.2 million death in the world[1]. Due to diabetes, the chance of death is increased two times[2]. Diabetes can create more complexity in the case of kidney diseases, cardiovascular diseases, foot ulcers, loss of eyesight, damage to the nerve system, etc[3]. That's why it is called the seventh foremost cause of human death in the world[4]. And, the treatment of diabetes is very costly. But, early-stage identification of diabetes can minimize these complexities greatly.

## Problem Statement

In most cases, patients come to the doctor when the blood sugar level is very high and facing many physical complexities. But, identification in the early stage by observing some physical factors can prevent these problems. Also, the clinical process of diabetes prediction is costly and time-worthy.At present there is no trustworthy process of diabetes prediction at early stage.

## Objective

Our research objective is to predict diabetes in the early stage by analyzing minimum numbers of physical and pathological tests.

## Literature Review

Many researchers are working in the fields of diabetes prediction and drug designing for diabetes. Machine Learning and Data Mining are now very common in diabetes prediction.

One of the best works has been done by Kumar and Velide. They have used a dataset consists of 865 samples with nine attributes and applied different machine learning algorithms like Naïve Bayes, Jrip, J48, Decision Trees, Artificial Neural Networks. Perfect classification accuracy has been achieved with the J48 algorithm[5]. Agarwal and Dewngan have studied the Pima Indian Diabetes dataset and suggested that the Support Vector Machine and Linear Deterministic Analysis algorithm together perform better[6]. Singh, Leavline, and Baig have also studied the Pima Indian Diabetes dataset and concluded that Random Forest outperforms for classification[7]. Joshi and Chawhan have considered only 7 attributes and identified that the support vector machine performs the best[8]. Recently, Islam has studied a dataset containing 520 samples with 16 attributes and achieved 99% classification accuracy with the Random Forest algorithm[9]. Maniruzzaman has analyzed a dataset consists of 6561 samples with 7 factors and achieved 94.25% with the combination of RF-based classifier and LR-based feature selection[10].

Most of the researches are focused on classification. A few have tried to minimize the number of factors but their classification accuracies are not noteworthy. Our focus will be on both factor minimization and classification accuracy. Because, less number of factors with better accuracy will help us to predict diabetes at the early stage.

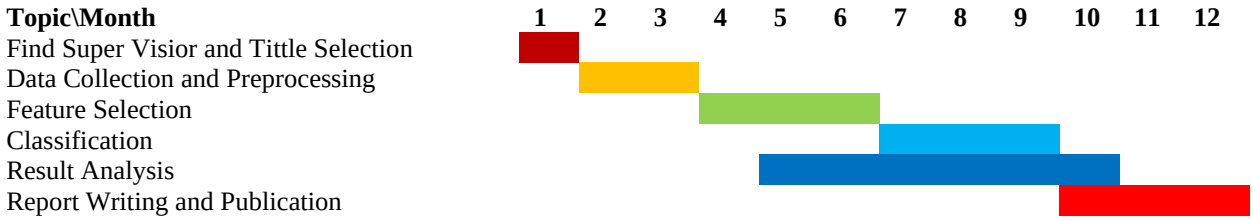
## Methodology

First of all, we will have to collect the dataset. Then, data preprocessing like missing value handling, data normalization, outlier detection, etc will be performed. After that, feature selection techniques like Chi-square test, LASSO, Recursive Feature Elimination, mRMR, Random Forest, Boruto, Genetic Algorithms, etc will be

applied to identify the most important factors. In the next, the dataset with selected features will be classified with classification algorithms like Support Vector Machine, Decision Trees, Random Forest, Naïve Bayes, Logistic Regression, K Nearest Neighbor, Neural Networks, etc. Finally, results will be analyzed and concluded.

## Time Schedule

As the M.Tech research duration is normally 1 year ( 2 Semesters ) at IIT, Kharagpur, I have scheduled the research within 12 months.




## Budget

The estimated budget has been tabulated below:

| Item | Price (USD) |
|---|---|
| Dell Inspiron 15 Plus Laptop ( Intel Core I7-10750H, NVIDIA® GeForce GTX® 1650 4GB, 16 GB DDR4, 512 GB SSD) | 1029 |
| Publication | 150 |
| Total= | 1179 |

N.B. The publication fee is not exact and the laptop price is accessed on 18th May 2021 from Dell's Offical website.

## Reference


[1]. R. Thomas, S. Halim, S. Gurudas, S. Sivaprasad, and D. Owens, "Idf diabetes atlas: A review of studies utilizing retinal photography on the global prevalence of diabetes-related retinopathy between 2015 and 2018," Diabetes research and clinical practice, vol. 157, p. 107840, 2019.

[2]. W. H. Organization et al., "Diabetes fact sheet n 312. October 2013," Archived from the original on, vol. 26, 2013.

[3]. E. Saedi, M. R. Gheini, F. Faiz, and M. A. Arami, "Diabetes mellitus and cognitive impairments," World journal of diabetes, vol. 7, no. 17, p. 412, 2016.

[4]. U. Diabetes and H. Lobby, "What is diabetes," Diabetes UK, 2014.

[5]. V. Kumar and L. Velide, "A data mining approach for prediction and treatment of diabetes disease," Int J Sci Invent Today, vol. 3, pp. 73–9, 2014.

[6]. P. Agrawal and A. Dewangan, "A brief survey on the techniques used for the diagnosis of diabetes-mellitus," Int. Res. J. of Eng. and Tech. IRJET, vol. 2, pp. 1039–1043, 2015.

[7]. D. Singh, E. J. Leavline, and B. S. Baig, "Diabetes prediction using medical data," Journal of Computational Intelligence in Bioinformatics,vol. 10, no. 1, pp. 1–8, 2017.

[8]. T. N. Joshi and P. Chawan, "Diabetes prediction using machine learning techniques," Ijera, vol. 8, no. 1, pp. 9–13, 2018.

[9]. M. F. Islam, R. Ferdousi, S. Rahman, and H. Y. Bushra, "Likelihood prediction of diabetes at early stage using data mining techniques," in Computer Vision and Machine Intelligence in Medical Image Analysis. Springer, 2020, pp. 113–125.

[10]. Maniruzzaman M, Rahman MJ, Ahammed B, Abedin MM. Classification and prediction of diabetes disease using machine learning paradigm. Health information science and systems. 2020 Dec;8(1):1-4.