# Prediction of Diabetes by KE Sieve Algorithm

V.Rohith
M.Tech,CSE Dept.
Sreenidhi institute of science and technology
Ghatkesar Hyderabad, Telangana-501301
India
vulli.rohith@gmail.com

G.Priyanka
M.Tech,CSE Dept.
Sreenidhi institute of science and technology
Ghatkesar Hyderabad, Telangana-501301
India
priyankagummad@gmail.com

Dr. Prasanta Kumar Sahoo
Professor in CSE Dept.
Sreenidhi institute of science and technology
Ghatkesar Hyderabad, Telangana-501301
India
prasantakumars@sreenidhi.edu.in

Dr. K. Eswaran
Professor in CSE Dept.
Sreenidhi institute of science and technology
Ghatkesar Hyderabad, Telangana-501301
India
Kumar.e@gmail.com

***Abstract*:** Diabetes mellitus commonly referred as diabetes, is a complex condition which impairs the body's ability to produce or respond to insulin, leading to high blood sugar levels. The diagnosis of diabetes is of great importance due to its severe long-term complications like cardiovascular disease, stroke, chronic kidney disease and damage to the eyes. Over the past few years, plenty of research studies has been done on diabetes identification, most of them are based on the Pima Indian diabetes data set. In this paper, we use a new non-iterative algorithm called KE Sieve to predict the presence of diabetes.
***Keywords:*** *KE Sieve, Pima Indians Dataset.*



## 1. INTRODUCTION

Diabetes has become a serious health problem worldwide, it is a disease which occurs when body does not produce or respond to insulin, resulting in high blood sugar levels than normal. When a person consumes food (carbohydrates), body breaks down the complex carbohydrates into glucose which enters into the blood stream. Glucose is an important source of energy for body cells and provides nutrients to the organs, muscles and nervous system. Glucose cannot directly enter into cells, beta cells in pancreas produces a hormone called insulin that helps body cells to consume glucose [1].

There are three types of diabetes seen in patients namely type 1, type 2, gestational diabetes. In case of type 1, body does not produce insulin, person has to inject the insulin as supplement when needed. In case of type 2, cells resist insulin, this is called insulin resistance. Gestational diabetes occurs during pregnancy in women.

According to WHO, number of people with diabetes in 1980 is around 108 million. In 2014 it is recorded that 422 million are affected, global prevalence of diabetes among adults has increased from 4.7% in 1980 to 8.5% in 2014. Over 205 million women are now living with diabetes, Half of the Women in low income countries die prematurely due to high blood glucose. On an average each year 1.6 million deaths are recorded. Rise over the years is due to type 2 diabetes, major factors are obesity and overweight among adults [2].

Medical diagnosis is one of the most important and challenging task. Hospitals contain huge amounts of patient's data, to diagnose a disease, the analysis of the data, decision making depends on the doctor's knowledge and experience. If any mistake done by inexperienced Doctor may lead to incorrect results. In Today's digital time, application of computer based machine learning techniques on medical field has been showing great results in diagnosing diseases. So, the risk caused by manual data analysis can be replaced by machine learning.

In this paper we use a new algorithm [3, 4] for the classification of diabetes and show how it gets best results. The idea of algorithm is to separates N data points of dimension (d), by hyper-planes (q). The number of hyper planes required to separate the points is approximately log2 (N).

## 2. LITERATURE SURVEY

Many researchers have applied different machine learning techniques on Pima Indian diabetes identification. Sidong Wei1, Xuejiao Zhao, Chunyan Miao [5] applied different preprocessors on Deep Neural Network (DNN), support vector machine (SVM), they have shown that among those techniques, DNN gave best accuracy as 77.86%.

In a research by authors Aakanksha Mahajan, Sushil kumar, Rohit Bansal [6] have used k-nearest neighbors (KNN) classifier and PSO to pre-process the dataset. The accuracy achieved by them is 77.0%.

In paper [7] author Asma A. AlJarullah studied the dataset clearly and followed two step pre-processing, first step was focused on attribute selection, some irrelevant attributes with missing values were removed and in second step numerical discretization is applied on data obtained from step 1, now this processed data is given to decision tree classifier and achieved 78.17% accuracy.

## 3. METHODOLOGY

### 3.1 KE SIEVE ALGORITHM

This algorithm separates each n dimensional point from every other point by at least one hyper plane [3, 4]. The algorithm can be briefly explained as

1) Consider a set of N train points in an n-dimensional space, assume this as 'X' space. Consider another n-dimensional space as 'T' with no points in it. Draw some initial planes in T space, plane equation is represented as

1+a1x1+a2x2+.....+a n x n = 0 [4]

2) Transfer one point at a time from X space to T space. Before placing the points in T space, orientation vector [4] (Ov) for each point is calculated w.r.t all the planes. Orientation vector gives the information about whether the point lies positive or negative side of the plane [4].

3) For the first point, Ov is calculated and placed in T space. From second point onwards Ov is compared with Ov of existing points in T space, If Ov of the two points does not match then the point is placed which means point is separated by plane. If Ov of two points match then point is said to be falling in the same quadrant. So, point is not placed, those pair of points are called neighbours [4] and are kept aside.

4) Whenever the neighbours collection count reaches n (dimension), a new plane is drawn in T space by passing through midpoints of each and every neighbours pair. With a single plane all the neighbours will be separated in T space.

5) Also whenever a new plane is added, orientation vectors of all the existing points in T space are also updated.

6) After the separation of all the train points is done, when a new test point is given, compute the dot product of test point orientation vector with all the train points orientation vectors. Take x% of max value dot product results and calculate the Euclidean distance for those train points with test point, assign the label of the train point with minimum distance to the test point.

This algorithm also has the functionality of restarting [2] i.e. if the new points are given for training, algorithm starts separating these points from where it has previously stopped.

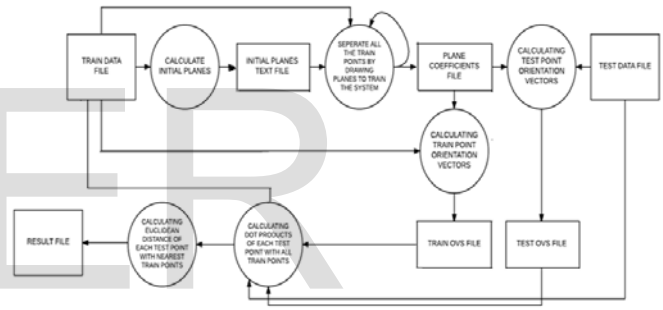


**Figure 1: Process Flow Diagram**

### 3.2 PIMA INDIANS DATA SET

This data set is taken from UCI machine learning repository [8], originally data was collected by National Institute of Diabetes and Digestive and Kidney Diseases. Dataset consists of 768 samples of women aged at least 21 with 8 attributes.

**Attribute Information**:

Pregnancies: Number of times pregnant
Glucose: Plasma glucose concentration a 2 hours in an oral glucose tolerance test
Blood Pressure: Diastolic blood pressure (mm Hg)
Skin Thickness: Triceps skin fold thickness (mm)
Insulin: 2-Hour serum insulin (mu U/ml)
BMI: Body mass index (weight in kg/ (height in m) ^2)
Diabetes Pedigree Function: Diabetes pedigree function

Age: Age (years)
The outcome is determined by 0 (absence of diabetes) or 1 (presence of diabetes).

## 4. RESULTS

Pima Indians dataset consists of 768 instances with 8 dimensions. We have divided the dataset with 40-60 split i.e. 40% (308) instances for training and 60% (460) instances for testing.
Proposed algorithm is applied and got the following results:

i) Initially 3 planes were taken.
ii) Total train points were separated by 19 planes.
iii) 13% of orientation vectors dot product and 12 nearest neighbors is taken for each test point while testing.
iv) Time taken for training and testing is within a second and accuracy is 83.6%.

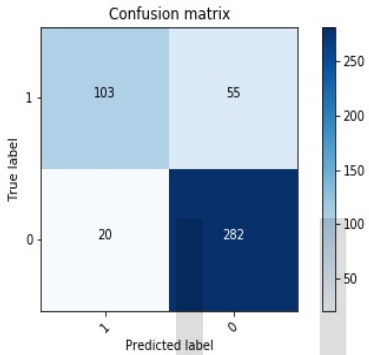


**Figure 2: Confusion Matrix**

Specificity = TN / (TN+FP) = 282 / (282+20) = 93%
Sensitivity = TP / (TP+FN) = 103 / (103+55) = 65%
Accuracy = (TP+TN) / Total = (103 + 282) / 460 = 83.6%

## 5. CONCLUSION

In this paper KE Sieve's algorithm is applied on Pima Indians dataset and achieved the best accuracy of 83.6% when compared to accuracy achieved by KNN classifier [6].

## REFERENCES

[1] https://www.livescience.com/62673-what-is-blood-sugar.html
[2] http://www.who.int/diabetes/global-report/en/
[3] K.Eswaran, "A non iterative method of separation of points by planes in n dimensions and its application" in https://arxiv.org/abs/1509.08742v5 October 23 2015
[4] Eswaran, K. (2017). On non-iterative training of a neural classifier Part-I: Separation of points by planes. 10.1109/IntelliSys.2017.8324238.
[5] S. Wei, X. Zhao and C. Miao, "A comprehensive exploration to the machine learning techniques for diabetes identification," 2018 IEEE 4th World Forum on Internet of Things (WF-IoT), Singapore, 2018, pp. 291-295.
[6] R. Bansal, S. Kumar and A. Mahajan, "Diagnosis of diabetes mellitus using PSO and KNN classifier," 2017 International Conference on Computing and Communication Technologies for Smart Nation (IC3TSN), Gurgaon, 2017, pp. 32-38.
[7] A. A. Al Jarullah, "Decision tree discovery for the diagnosis of type II diabetes," 2011 International Conference on Innovations in Information Technology, Abu Dhabi, 2011, pp. 303-307.
[8] Pima Indian Diabetes Database, Url: www.ics.uci.edu/~mlearn/MLRepository.html