# Predictive Analysis for Claims in Insurance Industry using Machine Learning

**Seema Jamal[1], Dr Kamal Shah[2]**

*[1]Seema Jamal PG Student, Thakur College of Engineering and Technology, Kandivali (E), Mumbai*
*[2]Dr. Kamal Shah Prof of I.T Department, Thakur College of Engineering and Technology, Kandivali (E), Mumbai*

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

**Abstract -** *Machine learning is turning into a preferred and vital approach within the field medical analysis. The relative performance of various machine learning methods such as Logistic Regression, Support Vector Machine, Random Forests and Naïve Bayes, for predicting diabetes complications has made the life easier by decreasing early mortality rate of patient suffering from diabetes for longer duration. This may help health insurance company to improve health by providing health service. Type 2 diabetes is the most common form of the disease with complications such as heart, kidney, vision, foot conditions. This study aimed at predicting some important complications of Type 2 diabetes such as heart disease and nephropathy in order to provide useful information for patients. Different variables such as age, glucose in blood, Blood pressure, Hba1c, smoking habit, year of infection were selected for each complication. The results indicate that which algorithm is suitable to predict complications more accurately. The study findings can be useful to develop information systems in the field of health as a decision support system for physicians and help patient to take insurance plan. Companies perform underwriting method to form choices on applications and to cost policies consequently.*

***Key Words***: **Diabetes complications, predictive analysis, machine learning algorithm, feature selection, health insurance scheme.**

## 1. INTRODUCTION

Machine Learning is the most advanced technique used today for pattern and decision rule extraction from a particular dataset. Despite being a branch of Artificial Intelligence, at its core, Machine Learning depends on statistical techniques. This has opened a new horizon for data modelling, data representation, data reasoning and data learning for contemporary computational science. Machine Learning algorithms are being used in various prediction models such as weather prediction, sports result prediction, stock market prediction and to some extent medical condition prediction. However, use of Machine Learning for passive health condition prediction is still rare. Diabetes Mellitus is one such example of health condition.

Diabetes is often called a modern-society disease because widespread lack of regular exercise and rising obesity rates are some of the main contributing factors for it. Problems related to diabetes are many and quite costly. Diabetes is a very serious disease that, if not treated properly and on time, can lead to very serious complications, including death.

Long-term complications of diabetes develop gradually. The longer you have got diabetes and therefore less the controlled your blood sugar the upper the chance of complications. Eventually, diabetes related complications could also be disabling or even life-challenging. With increasing incidence of diabetes, the economic burden of diabetes, given its chronic nature, severe complications, and need for long-term care, has become an important public health issue. In Type 2 diabetes, genetic factors, obesity and lack of physical activity have an important role in infection.
The complications of diabetes fall into two categories, i.e. acute and chronic. The present study focuses on the chronic complications which are also divided into two categories, including microvascular and macrovascular. The microvascular and macrovascular complications of diabetes mellitus account for most of the morbidity and mortality associated with the disease. While poor glycemic control and long duration of illness seem to be the most important risk factors for these complications, evidence suggests that ethnic variability in the susceptibility to the complications might also exist. Microvascular complications are diabetic retinopathy, nephropathy, neuropathy and macrovascular complications are coronary artery disease, peripheral arterial disease and stroke. In particular, the management of microvascular and macrovascular complications accounts for a large portion of costs during diabetic care. According to the structure and content of the data set, the following complications are examined.

- **Heart disease**: cardiovascular diseases are three times more common in diabetic patients than non-diabetics.[3]
- **Nephropathy**: a complication for both types of diabetes, i.e. insulin-dependent and non-insulin dependent. Raised albumin excretion rate is often the first laboratory manifestation of nephropathy. Diabetic nephropathy is a late complication of the disease whose symptoms appear years after the onset of diabetes.[5]. Along with higher Bp, HBA1C, Age and infection year if a person is having a GRE *below* 60 may have kidney disease and GFR of 15 or *lower* may have kidney failure.

A **eGFR** is a number based on your blood test for creatinine, a waste product in your blood. It tells how well your kidneys are working.

## 2. BACKGROUND AND MOTIVATION

### 2.1 Background

Significant advances in biotechnology and more specifically high-throughput sequencing result incessantly in an easy and inexpensive data production, thereby ushering the science of applied biology into the area of big data [1,2].

To date beside high performance sequencing methods, there is a plethora of digital machines and sensors from various research fields generating data, including super resolution digital microscopy, mass spectrometry, Magnetic Resonance Imagery (MRI), etc. Although these technologies turn out a wealth of information, they do not provide any kind of analysis, interpretation or extraction of knowledge. To this end, the area of Biological Data Mining or otherwise Knowledge Discovery in Biological Data, is more than ever necessary and important. In such a hybrid field, one of the most important research applications is prognosis and diagnosis related to human-threatening and/or life quality reducing diseases. One such disease is Diabetes mellitus (DM).

Machine Learning (ML) is expected to bring significant changes to the field of technology. Machine learning is definitely a subfield of AI and software engineering that enables software package to be a lot of correct in predicting results. The prime objective of machine learning technology is to make algorithms that may get input data and leverage statistical analysis to predict an appropriate output.

### 2.2 Motivation

According to IDF Atlas published in 2017, there are around 424.9 million Diabetes patients around the world aged from 20-79 years, of whom 95% suffer from Type 2 Diabetes Mellitus (T2DM). It is predicted that the number will increase to 628.6 million by 2045. Several Machine Learning based models exist that deal with Diabetes Mellitus. However, most of these systems only predict the probability of a person having Diabetes in the near future. Diabetes Mellitus can induce other complications like Nephropathy, Cardiovascular disease, Retinopathy and Diabetic Foot disease. In 2017 alone, 4 million people died all around the world due to diabetic related complications, mostly because they were not monitored closely and warned beforehand. There is a scope to introduce a complete system that can correctly predict onset of complications caused by T2DM using Machine Learning techniques, which can save thousands, if not millions, of lives around the world. This influenced the research work done in this paper.

Migrant Asian Indian populations have a higher prevalence of diabetic nephropathy than native populations of the concerned countries [10]–[11]. Similarly, in a cohort study conducted among patients with diabetes mellitus in the Netherlands, individuals of Asian Indian ancestry had 3.9 higher odds of developing albuminuria, and a 1.45 times higher rate of reduction in glomerular filtration rate than individuals of European ancestry [11]. Not with standing these lower prevalence rates, the numbers of individuals reaching end-stage renal disease as a result of diabetic nephropathy is likely to substantially increase in the near future, on account of the sheer number of people with diabetes mellitus in India. That few of these individuals will be able to afford chronic dialysis or kidney replacement, the only two effective modalities of treatment for end-stage renal disease, is of great concern [12].

The presence of T2DM seems to confer a 3–4 times higher risk of cardiovascular disease to Asian Indian individuals than to their white counterparts, even after adjusting for sex, age, smoking status, hyper tension and obesity [17]. Possible explanations include the atherogenic milieu promoted by high levels of insulin resistance and the high prevalence of ‘atherogenic dyslipidaemia’ characterized by high levels of triglycerides and small dense LDL cholesterol, and low levels of HDL cholesterol [18]. In the Chennai Urban Population Study, a population-based study conducted in two residential colonies in Chennai, in South India, CAD had a prevalence of 21.4% among individuals with T2DM, compared with 9.1% among those with normal glucose tolerance and 14.9% among those with impaired glucose tolerance [19].

## 3. METHODOLOGY

This was an experimental study that was collected by several features including age, gender, blood pressure, and so forth. The goal is, predicting the implications of Type 2 diabetes such as heart disease, retinopathy, neuropathy, and nephropathy. Seasonable prediction has been done to inform the patient about the future of their disease through timely prediction of the complications. The study aimed at classifying patients with Type 2 diabetes and finds a model between experimental signs of patients, family history and their daily routine with the complications observed in patients with diabetes. Identifying these factors can help control the disease.

### 3.1 Data

Data used in the study is collected doctor’s clinic. The data contains the patient who is diagnosed with type 2 diabetes. Patients' records contained age, BP, blood glucose, HBa1c, bmi, smoking habit, eGFR, infection year, family history. And treatment process as well as observed complications. The final data set included 268 records (268 patients) and 10 particular features. There are two types of dependent and independent variables based on independent variables dependent variables changes. Data set is divided into train and test data in which 70% 0f data is used to train algorithm and 30% of data is used to test algorithm.

### 3.2 Classification

We used different regression and classification algorithm to predict the complications due to long term diabetes. Jupiter notebook is used to implement different algorithm in which different library is used for coding. Four algorithms (logistic regression, Support vector machine, Random forest and Naïve byes) is used to predict complication due to diabetes.

Logistic Regression is a supervised learning algorithm that trains the model by taking input variables(x) and a target variable(y). In Logistic Regression the output or target variable is a categorical variable, unlike Linear Regression, and is thus a binary classification algorithm that categorizes a data point to one of the classes of the data [20]. The general equation of Logistic Regression is:

$$log\left(\frac{p(X)}{1-p(X)}\right) = \beta_0 + \beta_1 X$$

Where, p(X) is the dependent variable, X is the independent variable, β0 is the intercept and β1 is the slope co-efficient.

Support Vector Machines, SVM, is a supervised learning model with associated learning algorithms that analyze data used for classification, regression analysis and outlier detection [21, 22]. It is a non-probabilistic binary linear classifier, but can be manipulated in a way that it can perform non-linear and probabilistic classification as well, making it a versatile algorithm. An SVM model is a representation of the instances as points in space mapped so that they can be categorized and divided by a clear gap. New instances are then mapped into the same space and predicted in which category it might be in based on which side of the gap they fall in. The main advantage of SVM is the fact that it is effective in high dimensional spaces. Additionally, it is also memory efficient since it uses a subset of training points in the decision function.

Random Forest is a versatile, easy to use machine learning algorithm that produces, even without hyper-parameter tuning, a great result most of the time. It is additionally, one among the foremost used algorithms, because it's simplicity and the fact that it can be used for both classification and regression tasks. To say it in easy words: Random forest builds multiple decision trees and merges them along to get a lot more correct and stable prediction. Random Forest is used to increase the performance by avoiding overfitting and bias through aggregation of several trees.

Naïve Bayes is a supervised learning algorithm that depends on Bayes theorem for classification. Bayes theorem uses conditional probability which in turn uses prior knowledge to calculate the probability, that a future event will take place. The formula for Bayes Theorem is:

$$P(H|E) = \frac{P(E|H) * P(H)}{P(E)}$$

Here, P(H|E) is the posterior probability, the probability that a hypothesis (H) is true given some evidence (E). P(H) is the prior probability, i.e., the probability of the hypothesis being true. P(E) is the probability of the evidence, irrespective of the hypothesis. P(E|H) is the probability of the evidence when hypothesis is true. Naïve Bayes algorithm is used for binary and multiclass classification and can also be trained on a small dataset which is a huge advantage.

## 4. RELATED WORK

The performances of different algorithms are evaluated in order to predicate the complications to find the best model and algorithm. To evaluate various algorithms, Python 3+, Anaconda and Jupyter notebook is employed.

**Step 1** Data preparation: To start any project of data mining the most tedious task is acquiring and preparing data set. Here we used clinical data. We will perform machine learning algorithm to predict the complications due to diabetes.

**Step 2** Data exploration: When encountered with a data set, first we should analyse and "get to know" the data set. This step is important to familiarise with the dataset, to achieve some understanding regarding the potential features and to check if data cleaning is required.

**Step 3** Data Cleaning: Next phase of the machine learning work flow is the data cleaning. Considered to be one of the crucial steps of the work flow, because it can make or break the model.

There are several factors to consider in the data cleaning process.

- Duplicate or irrelevant observations.
- Bad labelling of data, same category occurring multiple times.
- Missing or null data points.

- Unexpected outliers.

Since we are using a standard data set, we can safely assume that above factors are already dealt with. Unexpected outliers either useful or potentially harmful.

**Step 4** Feature Engineering: Feature engineering is the process of transforming the gathered data into features that better represent the problem that we are trying to solve to the model, to improve its performance and accuracy. In the data set we have the following features.

Age, BP, Blood glucose, HBA1C, Insulin, Smoking habit, eGFR, Infection year, family history.

**Step 5** Model Validation: In many studies, two validation methods are used, namely hold-out method and k-fold cross validation method, to evaluate the capability of the model. According to the goal of every drawback and therefore the size of dataset, we can choose different methods to solve the problem. In hold-out technique, the dataset is divided two parts, training set and test set. The dataset is employed to train the machine learning algorithm is training set and the dataset employed to judge the model is test set. The training set is different from test set. In this study, we used this method to verity the universal applicability of the methods. In k-fold cross validation technique, the whole dataset is used to train and test the classifier. First, the dataset is average divided into k sections, that known as folds. In training method, the method uses the k-1 folds to training the model and onefold is used to test. This method will be repeated k times, and each fold has the chance to be the test set. The final result is the average of all the tests performance of all folds. Though it is more accurate but much slower than train/test split.

The performances of different algorithms are evaluated in order to predicate the complications to find the best model and algorithm. The main criteria for comparison are: accuracy, Precision, Recall, F1 score, False positive, True Positive, AUC, Sensitivity, Specificity.

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN}$$

$$Recall = \frac{TP}{TP + FN}$$

$$Specificity = \frac{TN}{TN + FP}$$

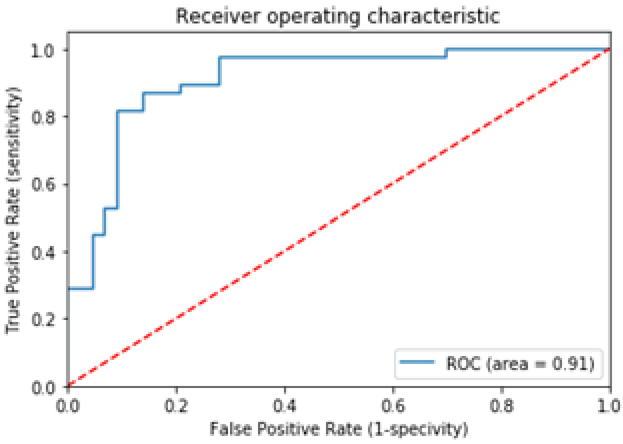
$$F1 = 2 * \frac{Precision * Recall}{Precision + Recall}$$

$$Precision = \frac{TP}{TP + FP}$$

ROC or Receiver Operating Characteristics is a graphical plot of sensitivity against (1Specificity) or in other words, a comparison of true positive rate (TPR) and false positive rate (FPR). It is used to visualize a classifier's performance at different thresholds to determine the best threshold point for the classifier.

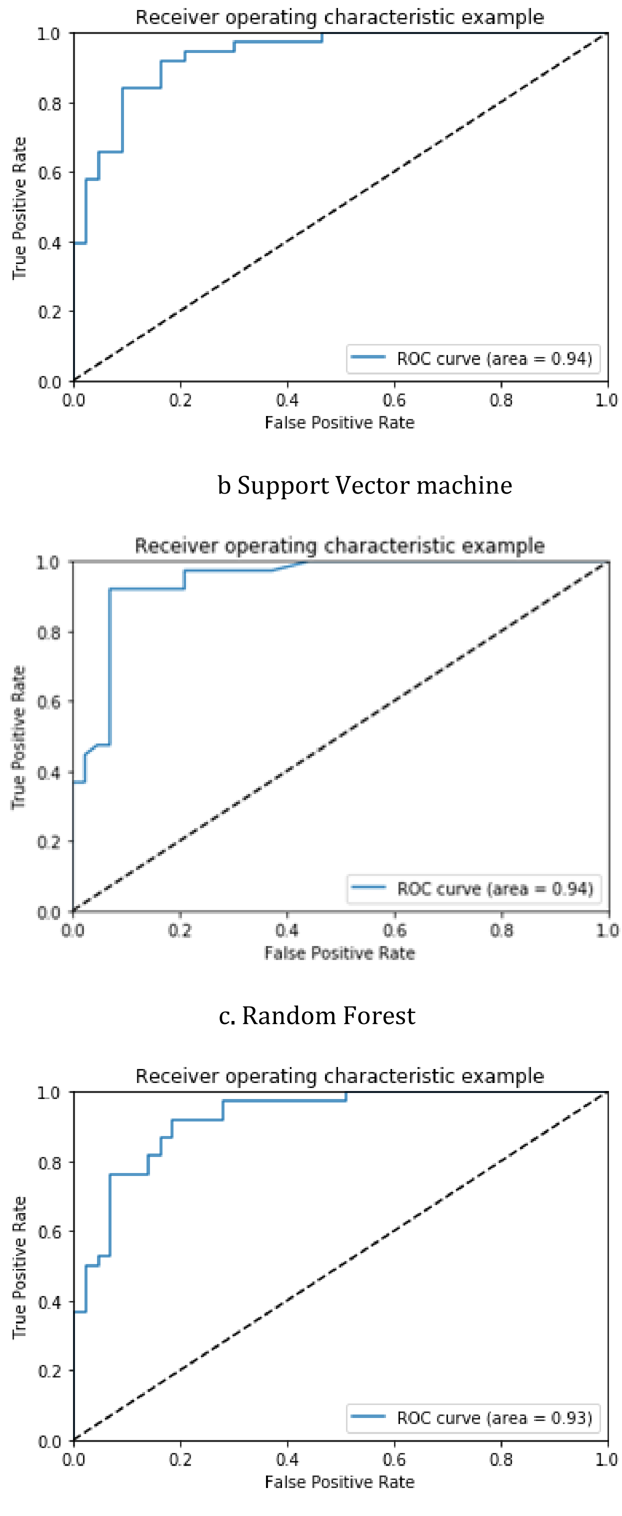
Table 1 RESULTS OF PRETICTING HEART DISEASE.

| | Logistic Regression | SVM | Random Forest | Naïve Bayes |
|---|---|---|---|---|
| Accuracy | 0.86 | 0.85 | 0.9 | 0.84 |
| Precision | 0.86 | 0.85 | 0.9 | 0.84 |
| Recall | 0.86 | 0.85 | 0.9 | 0.84 |
| F1 Score | 0.86 | 0.85 | 0.9 | 0.84 |
| False Positive | 6 | 7 | 5 | 6 |
| True Positive | 37 | 36 | 38 | 37 |
| AUC | 0.91 | 0.94 | 0.94 | 0.93 |
| Sensitivity | 0.87 | 0.83 | 0.92 | 0.81 |
| Specificity | 0.86 | 0.84 | 0.93 | 0.86 |



Roc of Heart disease

a. Logistic regression



b Support Vector machine

c. Random Forest

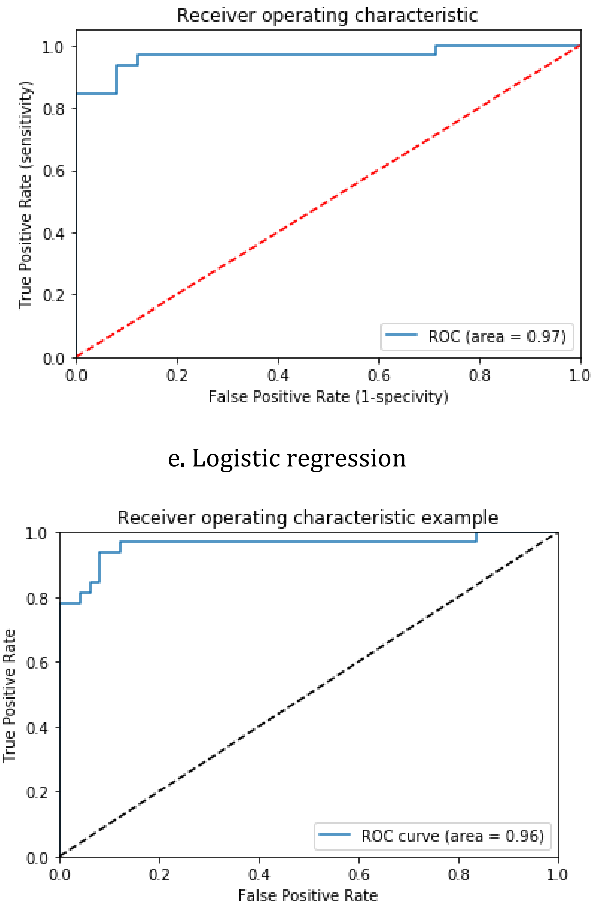
d. Naïve Bayes

**Figure 1: ROC of (a) logistic regression, (b) SVM, (c) Random Forest and (d) Naïve Bayes**

Table 2 RESULTS OF PREDICTING NEPHROPATHY

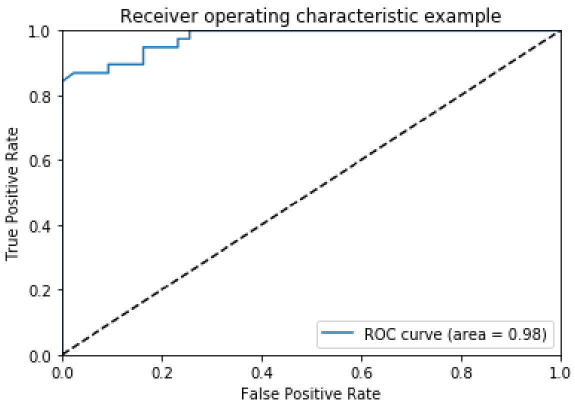
| | Logistic Regression | SVM | Random Forest | Naïve Bayes |
|---|---|---|---|---|
| Accuracy | 0.91 | 0.91 | 0.92 | 0.9 |
| Precision | 0.91 | 0.92 | 0.93 | 0.9 |
| Recall | 0.91 | 0.91 | 0.93 | 0.9 |
| F1 Score | 0.91 | 0.91 | 0.93 | 0.9 |
| False Positive | 2 | 0 | 1 | 4 |
| True Positive | 47 | 49 | 42 | 45 |
| AUC | 0.97 | 0.96 | 0.98 | 0.93 |
| Sensitivity | 0.84 | 0.84 | 0.97 | 0.78 |
| Specificity | 1 | 1 | 0.98 | 0.92 |

**ROC of Nephropathy**

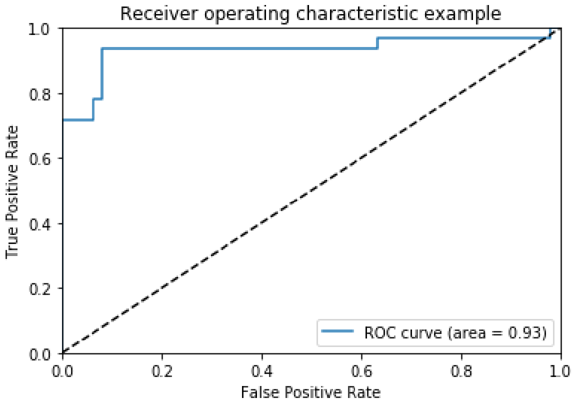


e. Logistic regression

f. Support Vector machine



g. Random Forest



h. Naïve Bayes

**Figure 2: ROC of (e)Logistic regression, (f)SVM, (g)Random Forest, (h)Naïve Bayes.**

All criteria are between 0 and 1. The results are summarized in Tables 1 and 2. As shown in Table 1, it could be said that the Random Forest algorithm has highest accuracy among all algorithms 90% also AUC of Random forest is 94% which highest among all so, for predicting heart disease Random forest is best algorithm. Table 2, also shows Random forest has highest accuracy 92 % and AUC 98%. So Random Forest is best algorithm to predict Nephropathy. Therefore, we can conclude from our dataset and prediction table that Random Forest is best algorithm to predict Heart and Kidney disease.

## 5. CONCLUSIONS

In this study, systematic effort was made to identify and review machine learning data mining approaches applied on DM research. DM is rapidly emerging as one of the greatest global health challenges of 21st century. The study aimed at exploring the best classification algorithms and attempting to create a dataset in order to reduce errors. The complications significantly aggravated expenditures on T2DM. Specific types of complications and also the presence of multiple complications are related with very higher expenditures. Proper management and the prevention of related complications are urgently needed to reduce the growing economic burden of diabetes. Buying health insurance plan for diabetes and its complications is the key to winning the battle against diabetes. A diabetes health insurance plan covers the cost of doctor's appointment, expenses towards diabetic tests, medicine costs, hospitalizations bill are also covered under health insurance. In this study the best algorithms are chosen for the detection of diabetic complications such as heart disease, and nephropathy. This will help the patient and also insurer to choose best insurance plan according to their complications. Insurers are using machine learning to enhance operational potency, from claims registration to claims settlement. Many carriers have already started to automate their claims processes, thereby enhancing the customer experience while reducing the claims settlement time. Machine learning and predictive models may also equip insurers with a much better understanding of claims prices. These insights will facilitate a carrier save a lot of bucks in claim prices through proactive management, fast settlement, targeted investigations and better case management. Insurers may also be a lot of assured regarding what proportion funding they assign to claim reserves.

## REFERENCES

[1]. Marx V. 2013. Biology: The big challenges of big data. Nature. 2013 Jun 13;498(7453):255-60. doi: 10.1038/498255a.

[2]. Mattmann CA. 2013. Computing: A vision for data science. Nature. 2013 Jan 24;493(7433):473-5. doi: 10.1038/493473a.

[3]. A.O. Stirban and D. Tschoepe, "Cardiovascular complications in diabetes," Diabetes Care, vol. 31, pp. S215-S221, 2008.

[4] A. Sopharak, B. Uyyanonvara, and S. Barman, "Automatic exudate detection from non-dilated diabetic retinopathy retinal images using fuzzy c-means clustering," Sensors, vol. 9, pp. 2148-2161, 2009.

[5] J. A. García-Donaire and J. M. Alcázar, "Ischemic nephropathy: detection and therapeutic intervention," Kidney International, vol. 68, pp. S131-S136, 2005.

[6] J. J. Duby, R. K. Campbell, S. M. Setter, and K. Rasmussen, "Diabetic neuropathy: an intensive review," American Journal of Health-System Pharmacy, vol. 61, pp. 160-173, 2004.

[7] Dandona, L., Dandona, R., Shamanna, B. R., Naduvilath, T. J. & Rao, G. N. Developing a model to reduce blindness in India: the International Centre for

Advancement of Rural Eye Care. Indian J. Ophthalmol. 46, 263–268 (1998).

[8]. Narendran, V. et al. Diabetic retinopathy among self reported diabetics in southern India: a population based assessment. Br. J. Ophthalmol. 86, 1014–1018 (2002).

[9]. Raman, R. et al. Prevalence of diabetic retinopathy in India: Sankara Nethralaya Diabetic Retinopathy Epidemiology and Molecular Genetics Study report 2. Ophthalmology 116, 311–318 (2009).

[10]. Samanta, A., Burden, A. C., Feehally, J. & Walls, J. Diabetic renal disease: differences between Asian and white patients. Br. Med. J. (Clin. Res. Ed.) 293, 366–367 (1986)

[11]. Chandie Shaw, P. K. et al. South-Asian type 2 diabetic patients have higher incidence and faster progression of renal disease compared with Dutch-European diabetic patients. Diabetes Care 29, 1383–1385 (2006).

[12]. Mani, M. K. Treating renal disease in India's poor: the art of the possible. Semin. Nephrol. 30, 74–80 (2010).

[13]. Pradeepa, R. et al. Prevalence and risk factors for diabetic neuropathy in an urban south Indian population: the Chennai Urban Rural Epidemiology Study (CURES-55). Diabet. Med. 25, 407–412 (2008).

[14]. Dutta, A., Naorem, S., Singh, P. & Wangjam, K. Prevalence of peripheral neuropathy in newly diagnosed type 2 diabetics. Int. J. Diabetes Dev. Ctries 25, 30–33 (2005).

[15]. Gill, H. K., Yadav, S. B., Ramesh, V. & Bhatia, E. A prospective study of prevalence and association of peripheral neuropathy in Indian patients with newly diagnosed type 2 diabetes mellitus. J. Postgrad. Med. 60, 270–275 (2014).

[16]. Abbott, C. A. et al. Explanations for the lower rates of diabetic neuropathy in Indian Asians versus Europeans. Diabetes Care 33, 1325–1330 (2010).

[17]. Forouhi, N. G., Sattar, N., Tillin, T., McKeigue, P. M. & Chaturvedi, N. Do known risk factors explain the higher coronary heart disease mortality in South Asian compared with European men? Prospective follow-up of the Southall and Brent studies, UK. Diabetologia 49, 2580–2588 (2006).