Minor Research Project - Annual Report
(July 2017 – June 2018)

DEVELOPING PREDICTIVE MODELS TO FORECAST AND PRE - TREAT DIABETES MELLITUS IN TAMIL NADU USING CLINICAL BIG DATA

(UGC Ref. No.: F. MRP-6517/16 (SERO/UGC) June - 2017)

(Link No: 6517 Comcode: TNBD007 UniqueID: RHTSJC)

Submitted to
UNIVERSITY GRANTS COMMISSION
South Eastern Regional Office (SERO), Hyderabad – 500 001

Principal Investigator
Dr. L. AROCKIAM, MCA., M.Tech., MBA., M.Phil., Ph.D.,
Associate Professor in Computer Science

PG & RESEARCH DEPARTMENT OF COMPUTER SCIENCE

ST. JOSEPH’S COLLEGE (AUTONOMOUS)

Special Heritage Status Awarded by UGC

Nationally Accredited at ‘A’ Grade (3rd Cycle) by NAAC

College with Potential for Excellence by UGC

DBT-STAR & DST-FIST Sponsored College

Tiruchirappalli – 620 002

A/c Dy No.	436
Date:	29/6/17

UNIVERSITY GRANTS COMMISSIONS -SOUTH EASTERN REGIONAL OFFICE
5-9-194, CHIRAG ALI LANE, IV FLOOR, A.P.S.F.C. BUILDING, HYDERABAD -500 001
Phones: 040 - 23204735, 23200208 FAX: 040 – 23204734, Website: www.ugc.ac.in, email: ugcsero@gmail.com

No.F MRP-6517/16 (SERO/UGC) Link No:6517. June,2017

The Accounts Officer
UGC-SERO, Hyderabad

Comcode: TNBD007
UniqueID:RHTSJC

Sub: Release of Grants-in-aid to Minor Research Projects for the year 2017-2018.

30 JUN 2017

Sir / Madam,

The has reference to the Minor Research Project proposal submitted by **Dr.L.Arockiam** Department of Computer Science of "St.Joseph's College- Tiruchirapalli-2- District Tiruchirapalli-620002" TIRUCHIRAPPALLI, TIRUCHIRAPPALLI entitled **"DEVELOPING PREDICTIVE MODELS TO FORECAST AND PRE-TREAT DIABETES MELLITUS IN TAMIL NADU USING CLINICAL BIG DATA"**. The subject expert, who evaluated the proposal, has recommended for financial assistance as detailed below.

Sl. No	Item	Amount Allocated (Rs.)	Amount Sanctioned as first installment (Rs.)
1.	Books & Journals	50000.	50000.
2.	Equipment	50000.	50000.
	Total	100000.	100000.
3.	Field work & Travel	50000.	25000.
4.	Chemical & Glass Ware	0 0	0 0
5.	Contingency (incl. Special Needs)	50000.	25000.
6.	Hiring Services	0 0	0 0
	Total	100000.	50000.
	Grand Total	200000.	150000.

1. I am further to convey the sanction of the University Grants Commission to the payment of **Rs.150000.** to the principal, St.Joseph's College- Tiruchirapalli-2- District Tiruchirapalli-620002,TIRUCHIRAPPALLI,TIRUCHIRAPPALLI as first installment (100% Non-Recurring and 50% Recurring grants) towards the above project..

Amount Sanctioned	Head of Accounts	Category
Rs. 100000.	35-CAP-MRP(50)-3(A)2202.03.102.02.01	GEN
Rs. 50000.	31-GIA-MRP(50)-3(A)2202.03.102.02.01	GEN

2. **The above approval is subject to the general conditions of grants prescribed by the UGC for this scheme.**
3. The sanctioned amount is debitable to the **Head of Accounts 35-CAP-MRP(50)-3(A)2202.03.102.02.01 (General), 31-GIA-MRP(50)-3(A)2202.03.102.02.01(General)** and is valid for payment during the financial year 2017-18 only and the amount of the Grant shall be drawn by the Accounts Officer (Drawing and Disbursing Officer) UGC-SERO, Hyd. on the Grants-In Aid Bill and shall be disbursed to and credited to **"The Principal, St.Joseph's College- Tiruchirapalli-2- District Tiruchirapalli-620002, TIRUCHIRAPPALLI, TIRUCHIRAPPALLI by Electronic Mode through PFMS Portal at the following details:**"*(a)Name & Address of Account Holder: The Principal, St.Joseph's College- Tiruchirapalli-2- District Tiruchirapalli-620002, TIRUCHIRAPPALLI, TIRUCHIRAPPALLI (b) Account No: 137501000020012 (c) Name & Address of Bank Branch: IOB CHINTHAMAN (d)IFSC Code:IOBA0001375.*
4. **In case the Principal investigator is having ongoing Major/Minor Research Project OR has been transferred/left/retired from the college, the released amount of Rs.150000./- may be returned to UGC-SERO, Hyderabad immediately, failing which action will be initiated against the College for not adhering with the norms of UGC for the scheme,**
5. The grantee institution shall ensure the utilization of grants -in-aid for which it is being sanctioned/paid. in case of non-utilization /part utilization, **interest @ 10% per** annum as amended from time to time on utilized amount from the date of drawl to the date of refund as per provision contained in General Financial Rules of Govt. of India will be charged.
6. The assets acquired wholly or substantially out of UGC's grants shall not be disposed or encumbered or utilized for

the purposes other than those for which the grant was given, without proper sanction of the UGC and should, at a time the college ceased to function, such assets shall revert to the UGC.

LinkNo.6517.

7. The Principal investigator of the project is required to submit the First year progress report of the work done along with the documents 1) Annual Report of the Project as per Annexure-III 2) Utilization Certificate duly signed by the Principal Investigator, Principal & Chartered Accountant 3) Statement of Expenditure for the approved heads for the sanctioned amount as per Annexure-V duly signed by the Principal Investigator, Principal & Chartered Accountant.
8. The interest earned by the College / Institute on this grants-in-aid shall be treated as additional grant which may be shown in the Utilization Certificate / Statement of Expenditure to furnished by the grantee institution.
9. The college has to send the filled in Acceptance certificate within 15 days of receipt of this letter, else the college may return back the sanctioned amount to this office. Further if the conditions of the acceptance letter is not acceptable or applicable to the P.I/College, the sanctioned amount be refunded back to SERO-UGC, Hyderabad.
10. The guidelines of Minor Research Project have to be followed in toto.
11. The Grant is subject to the adjustment on the basis of Utilization Certificate in the prescribed proforma submitted by the University/Institution.
12. The University/Institution shall maintain proper accounts of the expenditure out of the Grants, which shall be utilized, only on the approved items of expenditure.
13. The Utilization Certificate to the effect that the grant has been utilized for the purpose for which it has been sanctioned shall be furnished to UGC as early as possible after the close of current financial year.
14. The college shall maintain a Register of Assets acquired wholly or substantially out of the grant in the prescribed proforma.
15. The College shall fully implement to Official languages Policy of Union Govt. and comply with the Official Language Act, 1963 and Official languages (use for official purposes of the Union) Rules, 1976 etc.,
16. The approval for the above has been received vide letter No F.7-3/2016(SERO/MRP/RO) dated 6th September, 2016 from UGC, New Delhi.

Yours faithfully.

(Dr.G.Srinivas)
Joint Secretary
30.6.2017

Copy to:

1. The Principal (Along with DD / Funds transferred through E-mode)
   St.Joseph's College- Tiruchirapalli-2- District Tiruchirapalli-620002
   TIRUCHIRAPPALLI, TIRUCHIRAPPALLI - 620002.
2. Dr.L.Arockiam
   Dept. of Computer Science
   St.Joseph's College- Tiruchirapalli-2- District Tiruchirapalli-620002
   TIRUCHIRAPPALLI, TIRUCHIRAPPALLI - 620002.
3. The Dean/Director, College Development Council of affiliating University
4. The Commissioner /Director Collegiate Education, Government of TAMIL NADU
5. The Principal Accounts General (A & E)- Government of TAMIL NADU

(G.K.Pasrija)
Under Secretary

GAR Cap. Sl.No.141. /2017-2018
GAR GIA Sl.No.289. /2017-2018

The sanctioned grant of **Rs.150000. /-** has been transferred to your college Account as mentioned at the Point No. 3 of this Sanction Order by e-payment through PFMS portal vide date............................**You are requested to acknowledge the receipt of the above amount in your account by sending back the enclosed stamped receipt within 7 days.**

(R.Rayappa)
Accounts Officer

Appendix - A1

ANNUAL REPORT OF THE WORK DONE

Title of the Project: Developing Predictive Models to Forecast and Pre-Treat Diabetes Mellitus in Tamil Nadu using Clinical Big Data

(UGC Reference No. : **F. MRP-6517/16 (SERO/UGC) June-2017)**

1. Introduction:

Diabetes mellitus is one of the non-communicable diseases which is becoming a major global health problem. It is a chronic disease that occurs either when the pancreas does not produce enough insulin or when the body cannot effectively use the insulin which it produces. It is found that diabetes causes blindness, amputation and kidney failure. Lack of awareness about diabetes, insufficient access to health services and essential medicines can lead to the above mentioned complications. According to a study by the World Health Organization (WHO), number of diabetic patients will raise to 552 million by 2030, which means that one in 10 adults will have diabetes by 2030. In 2014, the global prevalence of diabetes was estimated to be 9 % among adults aged 18+ years [1]. WHO insisted with an alarm that Diabetes is the 7th leading cause of death in the world. In 2012, an estimated 1.5 million deaths were directly caused by diabetes. Total deaths due to diabetes are projected to rise by more than 50 % in the next 10 years.

Moreover, the International Diabetes Federation said that nearly 52 % of Indians are not aware that they are suffering from high blood sugar. More than 62 million diabetic individuals are currently diagnosed with the disease. It is predicted that, by 2030 diabetes mellitus may affect up to 79.4 million individuals in India. A nation-wide study, conducted by the Indian Council of Medical Research`s INDIAB (India Diabetes) has confirmed that the one out of 10 people in Tamil Nadu are affected by diabetes, and every two persons with age group of 25 are in the pre-diabetic stage. It is stated that 14.8 per cent of urban population and 11 per cent of rural population of Tamil Nadu are affected by diabetes. Madras Diabetes Research Foundation suggested that about 42 lakh individuals have diabetes and 30 lakh people are in pre-diabetes stage. At least, 1,000 people avail treatment for diabetes out of the 12,000 outpatients who visit Rajiv Gandhi Government General Hospital (RGGGH), a leading Government hospital in Chennai in 2013.

2. Review of the Literature

Raghupathi et al. [2] presented a review of big data analytics in healthcare. He pointed out the promises and potentials of big data in healthcare and outlined an architectural framework and methodology for applying big data in healthcare. He elaborated the advantages of using big data analytics in healthcare and offered the available platforms and tools for applying big data analytics in healthcare. Aiswarya Iyer et al. [3] used Decision Tree and Naïve Bayes algorithms for the prediction of diabetes in pregnant women. Tenfold cross validation is used to prepare training and test data and the J48 algorithm is employed on the dataset using WEKA on the Pima Indians Diabetes Database of

National Institute of Diabetes and Digestive and Kidney Diseases. The authors concluded that both algorithms are efficient for the diagnosis of diabetic and Naïve Bayes technique resulted in least error rate.

A.A. Aljumah et al. [4] suggested a predictive analysis of diabetic treatment using a regression based data mining technique. Oracle Data Miner (ODM) tool was employed for predicting diabetics and support vector machine algorithm was applied for experimental analysis on Datasets of Non Communicable Diseases (NCD) risk factors in Saudi Arabia. Mohammed et al. [5] presented a review of existing applications of the Map Reduce programming framework and its implementation platform Hadoop in clinical big data and related medical health informatics.

N.M. Saravana Kumar et al. [6] presented Predictive Analysis System Architecture with various stages of data mining. Prediction was carried out in Hadoop / Map Reduce environment. Predictive Pattern matching system helped in comparing the analyzed threshold value with the obtained value. Saumya et al. [7] applied analytical techniques to reduce the hospital readmission of diabetic patients. They proposed a methodology using Hive as its preprocessing tool and RHadoop as its analysis and predictive modeling tool. Classification was performed by Logistic Regression, SVM, KNN and Decision Tree methods and miss-classification error rates were calculated.

D. Peter Augustine [8] presented a concept paper on analyzing the data flowing from health monitoring devices. He also presented the present status of healthcare in India. He explained the application of Hadoop's map reduce in healthcare data and presented an interface HIPI (Hadoop Image Processing Interface) in Hadoop environment. Sadhana et al. [9] analyzed the Pima Indians Diabetes Database of National Institute of Diabetes and Digestive and Kidney Diseases data set using a proposed architecture which comprised of Hive and R. The raw data (CSV file) was given as input to the Hive for analyzing and partitioning. The output file then passed through R system for statistical computing.

MuniKumar N et al. [10] has pointed out the massive shortage of proper healthcare facilities and addressed his concern to provide greater access to primary healthcare services in rural India. He also presented the ability of Big Data analytics in processing huge volumes of data in real-time situations to turn the dream of Swachh Bharath (Clean & Healthy India) into reality. AditiBansal et al. [11] proposed an architecture consisting of Dynamic Hadoop Slot Allocation (DHSA) which uses the slot based resource model. He presented two more alternatives for DHSA namely, Pool Independent DHSA (PIDHSA) and Pool dependent DHSA (PDDHSA). He concluded that DHSA focused on the maximum utilization of slots by allocating map (or reduce) slots to map and reduce tasks dynamically. K.Sharmila et al. [12] presented a survey paper on the advancement in the field of data mining and its latest adoption in Hadoop platform and Bigdata algorithms used, and the open challenges in the Indian medicinal data set.

3. Objectives

This project aims to propose novel predictive model to predict diabetes mellitus using the clinical and e-diabetic Big Data. The objectives of the proposed work are formulated as below:

- To create an e-diabetic portal
- To build data warehouse using cloud computing technology
- To apply Big Data analytics to derive patterns
- To predict diabetes using the generated pattern

4. Methodology:

In this modern era, human beings encounter different health issues. Most of the health issues are due to the food habits of the individuals. In this project work, a predictive approach is proposed to pre-treat Diabetic Mellitus. The proposed approach has three phases namely data collection, data storage and analytics. This approach plays an important role in predicting diabetes and pre-treating diabetic patients. The phases in the proposed approach for diabetic prediction are presented in Fig. 1.

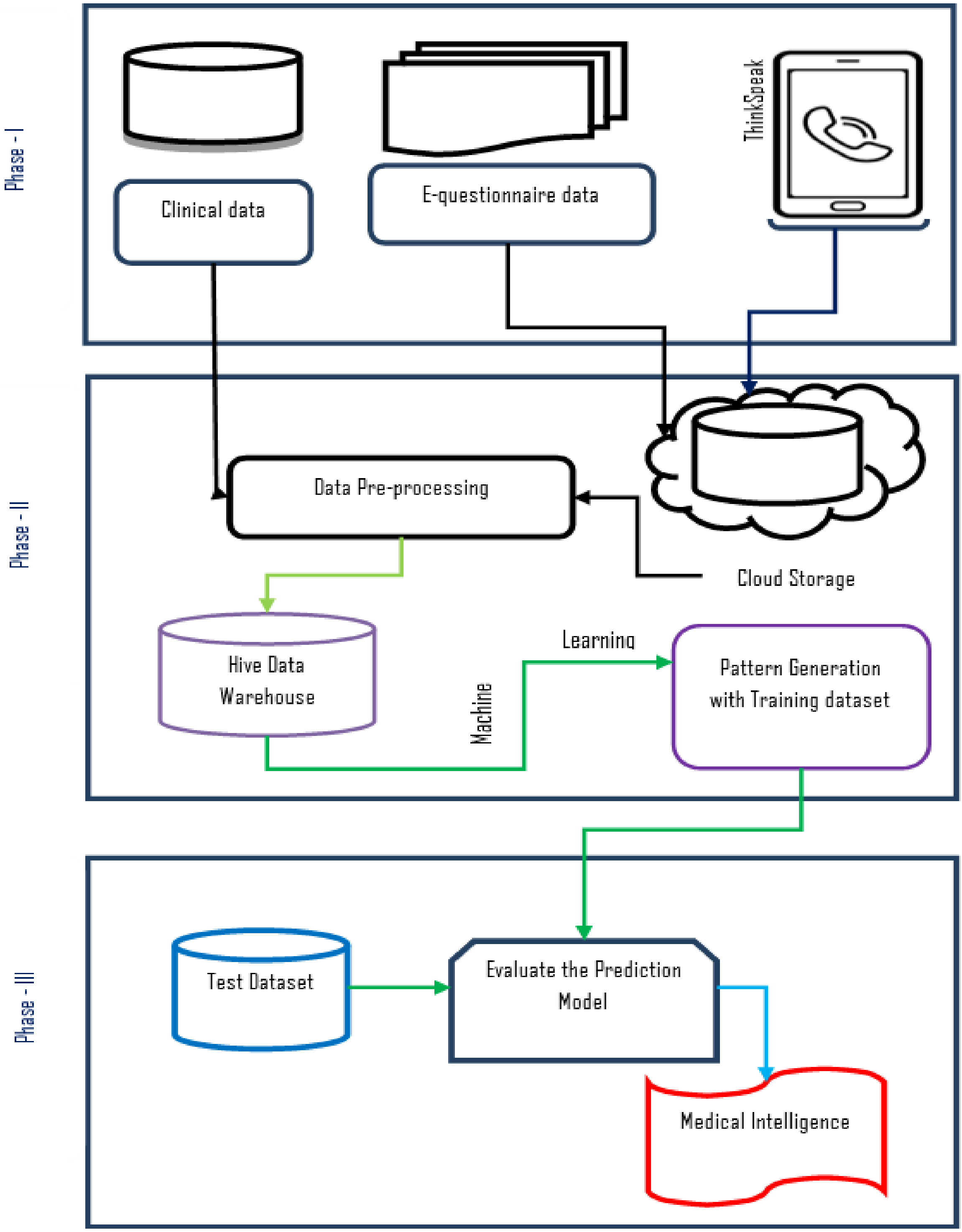


Fig 1. Methodological Diagram for Predicting Diabetes Mellitus

In the first phase, data collection is done through IoT devices and other sources. The collected data are cleaned using pre-processing techniques.

Phase two deals with data storage. The pre-processed data are stored in warehouses. Since data is collected from various sources and are huge in size to store, cloud storage is used. The records stored in the cloud were analyzed to establish association between the factors such as BP, BMI, Air Pollution level etc., with Diabetic Mellitus.

The third phase of the proposed approach deals with Predictive Analytics where the decisions are taken based on association rules with respect to diet pattern, physical fitness, current medicine intake etc.

5. Work Completed:

5.1 Sample Selection for Data Collection

Tamil Nadu is one of the states in India. For the purpose of data collection, it is divided into four regions Chennai City Region viz Central Region, Western Region, and Southern Region for data collection. Each region has more than 4 districts. The Central Region includes 10 districts (Cuddalore, Thanjavur, Perambalur, Tiruchirappalli, Ariyalur, Karur, Nagapattinam, Thiruvarur, Pudukottai, and Karaikal). The western region consists of 6 districts which are Coimbatore, Erode, Nammakal, Salem, Dharmapuri and Nilgiris. Dindigul, Madurai, Theni, Sivaganga, Virudunagar, Ramanathapuram, Tirunelveli, Thoothukudi and Kanyakumari are included in the southern region. Finally, Chennai, Thiruvalluvar, Kancheepuram, Vellore and Tiruvannamalai districts are classified under the Chennai city region.

5.2 Phase I: Data Collection (1 – 6 Months)

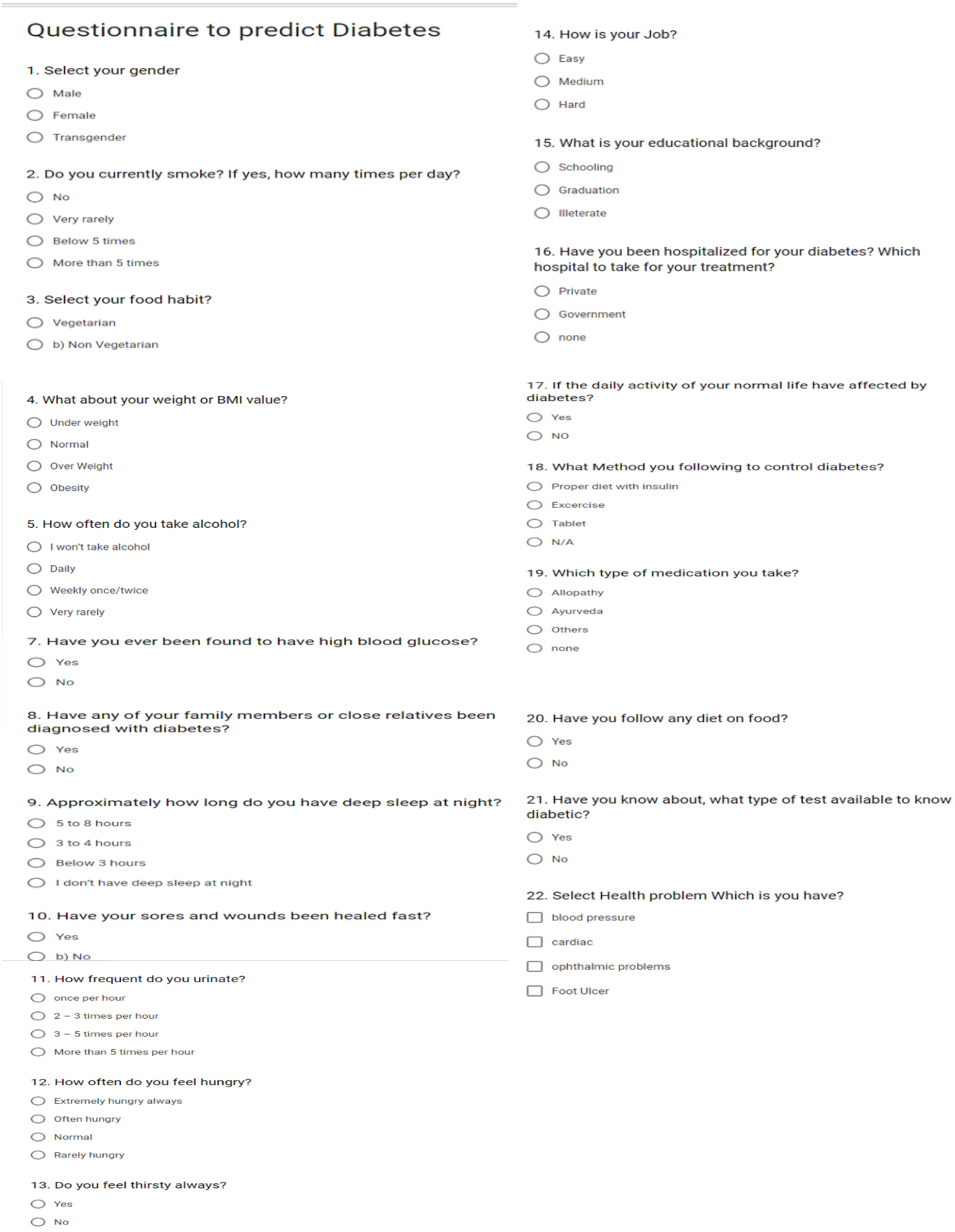
(a) Collection of clinical data from Medical Laboratories and storing it in raw Data Store (1–3 Months)

The data set selection is one of the important processes in data mining. For this, most relevant data is chosen from a particular domain to attain values. The derived values should be more flexible and informative in that domain. In this study, the data collection is done through questionnaire, sensors and from the clinical experiments.

The data are collected in person using pre-tested questionnaire, medical devices and sensors. The survey included participants above 18 years of age who belong to Tamilnadu.

(b) Collection of e-questionnaire data from users through online survey (1–3 Months)

A standardized questionnaire, including items of common risk factors of diabetes, was sent to the participants to obtain information on demographic characteristics, family diabetes history, and lifestyle risk factors. It contains a set of questions under three categories. The sample screenshots of the e-questionnaire are presented in fig. 2.



Questionnaire to predict Diabetes

1. Select your gender

○ Male
○ Female
○ Transgender

2. Do you currently smoke? If yes, how many times per day?

○ No
○ Very rarely
○ Below 5 times
○ More than 5 times

3. Select your food habit?

○ Vegetarian
○ b) Non Vegetarian

4. What about your weight or BMI value?

○ Under weight
○ Normal
○ Over Weight
○ Obesity

5. How often do you take alcohol?

○ I won't take alcohol
○ Daily
○ Weekly once/twice
○ Very rarely

7. Have you ever been found to have high blood glucose?

○ Yes
○ No

8. Have any of your family members or close relatives been diagnosed with diabetes?

○ Yes
○ No

9. Approximately how long do you have deep sleep at night?

○ 5 to 8 hours
○ 3 to 4 hours
○ Below 3 hours
○ I don't have deep sleep at night

10. Have your sores and wounds been healed fast?

○ Yes
○ b) No

11. How frequent do you urinate?

○ once per hour
○ 2 – 3 times per hour
○ 3 – 5 times per hour
○ More than 5 times per hour

12. How often do you feel hungry?

○ Extremely hungry always
○ Often hungry
○ Normal
○ Rarely hungry

13. Do you feel thirsty always?

○ Yes
○ No

14. How is your Job?

○ Easy
○ Medium
○ Hard

15. What is your educational background?

○ Schooling
○ Graduation
○ Illeterate

16. Have you been hospitalized for your diabetes? Which hospital to take for your treatment?

○ Private
○ Government
○ none

17. If the daily activity of your normal life have affected by diabetes?

○ Yes
○ NO

18. What Method you following to control diabetes?

○ Proper diet with insulin
○ Excercise
○ Tablet
○ N/A

19. Which type of medication you take?

○ Allopathy
○ Ayurveda
○ Others
○ none

20. Have you follow any diet on food?

○ Yes
○ No

21. Have you know about, what type of test available to know diabetic?

○ Yes
○ No

22. Select Health problem Which is you have?

□ blood pressure
□ cardiac
□ ophthalmic problems
□ Foot Ulcer

Fig 2. e-Questionnaire

Category A (Demographic): contains **personal information** including Name, Age, Sex, and Address.

Category B: contains details about **family details** that includes Diabetic history of parents, financial status, and educational status

Category C: includes the **physical data** like Height, weight, smoking habit, consumption of Alcohol, work type and physical activity.

Anthropometric measurements were taken from participants. Body Mass Index (BMI) was calculated as weight in kilograms divided by the square of height in meters (kg/m^2). A BMI ≥25 was defined as overweight. In category C, the data related to the following physical activities were collected: cigarette smoking at least 500 cigarettes in one's life, Alcohol consumption of at least 100 g per week for 1 year or longer and participation in moderate or vigorous physical activity for 30 minutes or more per day for at least 3 days in a week.

c) Collection of IoT data from users through ThingSpeak IoT cloud (4–6 Months)

IoT data are accumulated from the devices that are connected to the Internet. MQ-3, MQ-7 and Honeywell HPm Particle Sensor (represented in fig 3.) are used to collect air quality data. These sensors are connected to ThingSpeak cloud storage through NodeMCU board. ThingSpeak enables sensors, instruments, and websites to send data to the cloud and it is stored in a private channel. The platform is based on a MSP430 16-bit CPU running at 3.9 MHz. It provides a CC2420 radio chip, 48 kB of program flash and 10 kB of RAM. One Tmote Sky implements the 6LoWPAN border router connected to a computer running on Windows OS. The 802.15.4 radio is configured to be channel 15, which underlies WiFi interference. NodeMCU ESP8266 is a microcontroller with WiFi functionality that is used as bridge between sensors and cloud. The readings of the sensors are stored in this microcontroller temporarily and sent to cloud. In this data collection process, ThingSpeak web API is used for storing data into the cloud. They provide an IoT analytics platform service that allows aggregating, visualizing and analyzing live data streams in the cloud.

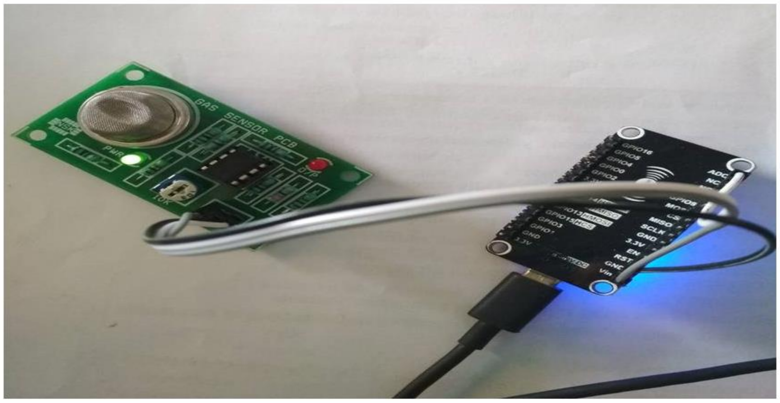


Fig 3.Honeywell Particle Sensor

The data related to blood pressure (diastolic and systolic) and blood sugar along with their screening data, the Body Mass Index (BMI), dietary history, physical activity, Air pollution level(Pm2.5& Pm10) are collected using this system. The selected population is also included the blood tests for collection of Random Blood Sugar level. The blood pressure is screened using Arm Bp digital monitor. The data relating dietary history and physical activity were collected through the questionnaire which was given in Fig. 2

Phase II: Data Storage and Analytics (7–18 Months)

(a) Creation of Experimental set-up with cloud storage (7–9 Months)

Cloud storage is a cloud computing model in which data is stored on remote servers accessed from the internet. It is maintained, operated, managed by a cloud storage service provider. In this proposed approach, Data is collected and stored in ThingSpeak which is a cloud service provider (experimental setup is explained in section 5.2 (c)).

(b) Pre-Processing and Diabetic Data Warehouse creation (10 – 12 Months)

Pre-processing of large volume of clinical data becomes essential. Data pre-processing is an important step in the data mining process. Observing the data which has not been carefully examined for such issues can produce misleading outcomes. If there is an ample amount of incorrect and redundant information or noisy and unreliable data, then knowledge discovery becomes challenging. Data pre-processing includes various steps such as cleaning, normalization, transformation, feature extraction and selection. The outcome of data pre-processing is the complete data set with reduced attributes. The major drawback with clinical data set is the existence of redundant records. These redundant records cause the learning algorithm to be biased. So, eliminating redundant records is essential to enhance the detection accuracy. During pre-processing, the raw data is supplied as input and several suitable data pre-processing methods are applied thereby decreasing the invalid instances in the dataset. In our diabetes dataset, data transformation and data validation are considered as the important pre-processing techniques for eliminating impure and invalid data.

Evaluation of collected data sets

After doing storage at cloud, data are analyzed using statistical approaches at regional level. Region level analysis is performed by classifying the diabetic patients into different categories according to the age, sex, income level, work type, treatment being taken, type of diabetes, type of medication, drugs consumed by the people, other health problems like blood pressure, cardiac and ophthalmic problems and foot ulcer. As per the survey results, prevalence of diabetes in central region of Tamil Nadu is **16.7%.** The percentage of diabetes affected persons based on their gender is presented in fig. 4.

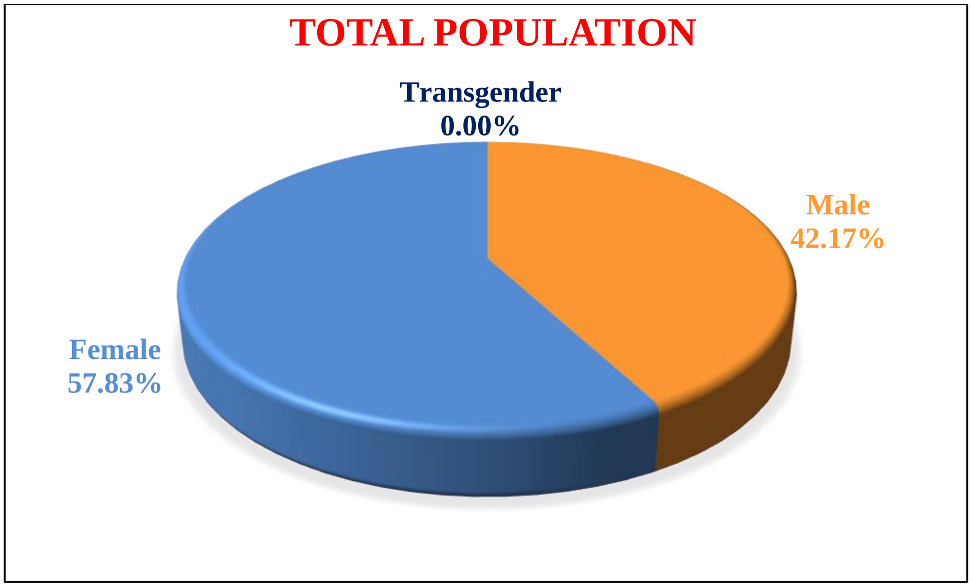


Fig 4. Percentage of Diabetes affected Persons

From the collected data, it is found only 4.5% of the populations is aware of the availability of HBA1C test and 12.2% of them are not aware of their carbohydrate level.

Based upon the income level and other factors like time, food habits etc. **82.7%** of people are facing hard to follow suggested food habits in their daily life. It is also found that 5.1% of those who are confirmed to be suffering from diabetes mellitus do not follow the suggested diet chart.

The above are the works that have been completed so far. The overall summary of the report has been presented in the table 1.

Table 1. Summary of the Report

Phases	Work to be done	Duration	Remarks
Phase I	Collection of clinical trials from the medical laboratories and storing it in raw Data Store	1-3 months	Completed
	Development of e-diabetic portal and IoT based system for collecting data from users	4-6 months	
Phase II	Creation of experimental set up with cloud storage	7-9 months	
	Preprocessing of collected data and building diabetic Data Warehouse	10-12 months	
	Hadoop Cluster setup	13-15 months	In progress
	Pattern generation with training data set	16-18 months	
Phase II	Implementation of analytical techniques with test data sets and generation of medical intelligence	19-21 months	
	Evaluations of project and presentation of the report	22-24 months	

6. Conclusion and Further Work to be done:

A novel approach for predicting the diabetic is proposed. It has three well defined layers with defined functionalities. Sample data collection is done from the selected cities in Tamil Nadu through various methods such as questionnaire, interviews and surveys. IoT devices were used to collect related to Air pollution, Blood pressure etc. The prevalence of diabetics for Tamil Nadu is identified by dividing the cities into four regions for data collection namely Central Region, western region, Southern Region and Chennai city region. Further data analysis is yet to be carried out.

7. Papers Presented and Published:

Based on the work done on the proposed research project, following papers are presented and published.

- Paper Published

1. L. Arockiam, A. Dalvin Vinoth kumar, "A Model to Predict and Pre-Treat Diabetes Mellitus", International Journal of Engineering and Techniques, ISSN: 2395-1303, pp. 1 – 6 **(Impact Factor: 3.546)**

- Papers Communicated

1. L. Arockiam et.al, "Prevalence of Type-II Diabetics Association with PM 2.5 and PM 10 in Central Region of Tamil Nadu, India", Indian Journal of Medical Research, ISSN 0971-5916, Paper Id: IJMR_1724_18, 2018, pp 1-12. **(Communicated) (Web of Science)**

2. L. Arockiam et.al, "Performance Analysis of Classification Algorithms for Diabetic Prediction using Pima- Indian Dataset", International Journal of Emerging Technologies and Innovative Research, ISSN: 2349-5162, IF: 5.87, paper Id: JETIR188432, 2018, pp 1-12. **(Communicated) (UGC Listed)**

- Paper & Poster Presented

1. L. Arockiam, A. Dalvin Vinoth kumar, "Predictive Model for Diabetes Mellitus in Tamil Nadu Using IoT Data" in Two Day International Conference on Transforming Technology held at Srimath Andavar Arts and Science College, Trichy, January 2018.

2. L. Arockiam, A. Dalvin Vinoth kumar, "A Model to predict and pre-treat Diabetes Mellitus" in International Conference on Advances in Computers Science and Technology held at St. Francis de Sales College, Bengaluru, January 2018.

3. L. Arockiam, A. Dalvin Vinoth kumar, "Swatchh Bharat - IoT Based Diabetic Predication", Young Scientist conclave at Indian International Science Festival (IISF-2017) conducted by Ministry of Earth Sciences, Govt. of India, October 13-16 2018.

Annexure – IV

UNIVERSITY GRANTS COMMISSION
SOUTH EASTERN REGIONAL OFFICE (SERO)
A.P.S.F.C. Building (4th Floor), 5-9-194, P.B. No. 152, Chirag-Ali-Lane
HYDERABAD– 500 001.

Utilization Certificate

Certified that the partial grant of **Rs. 1,50,000** (Rupees **One lakh fifty thousand** only) received from the University Grants Commission under the scheme of support for Minor Research Project entitled **"Developing Predictive Models to Forecast and Pre - Treat Diabetes Mellitus in Tamil Nadu using Clinical Big Data"** vide UGC letter no. **F.MRP-6517/16 (SERO-UGC)** dated **June 2017** has been fully utilized for the purpose for which it was sanctioned and in accordance with the terms and conditions laid down by the University Grants Commission.

(Dr. L. Arockiam)

SIGNATURE OF THE PRINCIPAL INVESTIGATOR

Dr. L. AROCKIAM
MCA.,M.Tech.,MBA.,CSM.,BLIS.,M.Phil.,Ph.D.,
Associate Professor in Computer Science
St.Joseph's College (Autonomous)
Tiruchirappalli- 620 002,Tamilnadu,India

(Rev. Dr. M. Arockiasamy Xavier SJ)

PRINCIPAL

PRINCIPAL
St. JOSEPH'S COLLEGE
(AUTONOMOUS)
TIRUCHIRAPPALLI 620 002

STATUTORY AUDITOR

ROY JOHN THOMAS, B.Com.,F.C.A.,
CHARTERED ACCOUNTANT
M.No. 200 / 25188

ROY JOHN THOMAS, B.Com.,F.C.A., * CHARTERED ACCOUNTANT *
D-4, BISHOP COMPLEX,
12A, CONVENT ROAD,
MELAPUDUR,
TRICHY - 620 001.

AMODEL TO PREDICT AND PRE-TREAT DIABETES MELLITUS

Arockiam L[1], Dalvin Vinoth Kumar[2]
1,2Department of Computer Science, St. Joseph's College (Autonomous), Thiruchirapalli, Tamil Nadu, India

Abstract:

Diabetes Mellitus (DM) is one of the non-communicable diseases and it causes major health problems. According to a study, there will be 552 million diabetic patients by 2030 all over the world. The Things embedded with sensors that are connected to the internet is referred as Internet of Things (IoT). The collection, storage and analysis of data from IoT devices facilitate effective monitoring diabetic patients. In this paper, a model for prediction of diabetes is proposed. This prediction model consists of layer of sensors for data collection, layer for storage and layer for analytics. The diabetic data collection may include the data from other sources such as clinical experiments and questionnaire. The collected data are cleaned using pre-processing techniques. In the storage layer, the preprocessed data are stored in the warehouses. The predictive analytics is performed using statistical, data mining and machine learning algorithms in the analytical layer. This model provides an approach to predict the diabetic mellitus

Keywords—**Diabetes Mellitus, Predictive model, Diabetes predication, Clinical data analytics, Diabetics in India.**

I. INTRODUCTION

Diabetes is referred as diabetic mellitus in which blood sugar levels are too high. It is defined as a clinical syndrome characterized by hyperglycaemia, due to inadequacy of insulin in the human body. High levels of blood glucose can damage the blood vessels in kidney, heart, eyes and entire nervous system. Lack of awareness about diabetes can lead to these complications. According to WHO, India is the residence of the highest number of diabetics with the population of 79.4 million by 2030 [1]. The International Diabetes Federation said that nearly 52 % of Indians not aware that they are suffering from high blood sugar [2]. In particular, Madras Diabetes Research Foundation suggested that about 42 lakh individuals have diabetes and 30 lakh people are in pre-diabetes.

II. TYPES OF DIABETES MELLITUS

There are three types of diabetes. They are Type-1 diabetes, Type-2 diabetes and Gestational diabetes [3]. The presence of diabetics is identified using the following factors long-term blood sugar (HbA1C), fasting blood sugar, fotal triglycerides, Family history of high blood sugar, Waist measurement, Height and Waist-to-hip ratio[4].

i. **Type-1 Diabetes:** This type of diabetes is also called as insulin dependent diabetes. It will start from childhood. It is immune mediated and idiopathic forms of b cell dysfunction, which lead to absolute insulin deficiency [5]. This is also an auto-immune mediated disease process which gives rise to absolute deficiency of insulin and therefore total dependency upon insulin for survival. It increases the risk of heart disease and stroke. The symptoms are very thirsty, urinating frequently, rapid weight loss, feeling very hungry, feeling extreme weakness and fatigue, Nausea, vomiting and irritability. The treatments of type -1 diabetes are injections of insulin, oral medications or dietary modifications, physically activity, regular check-up of blood sugar levels, controlling blood pressure and monitoring cholesterol levels [6].

ii. **Type-2 Diabetes:** It is also called as non-insulin dependent diabetes and adult onset diabetes. It is the most common form of diabetes. People may be affected by type-2diabetes at any stage, even during childhood [7]. Being overweight and inactive increases thechances of developing type-2 diabetes. It may originate from insulin resistance andrelative insulin deficiency. It can be controlled with weightmanagement, nutrition and exercise. The symptoms are very thirsty, urinating frequently,rapid weight loss, feeling very hungry, feeling extreme weakness, fatigue, nausea,vomiting, irritability, blurred vision, excessive itching, skin infections, sores that healslowly and dry and itchy skin [8]. Treatments such as using diabetes medicines, insulininjections, healthy food choices, exercise, Self Monitoring of Blood Glucose (SMBG),controlling blood pressure and monitoring cholesterol levels are some measures to control Diabetes Mellitus [9].

iii.**Gestational Diabetes:** According to the National Institutes of Health, the reported rate of gestational diabetes is between 2% to 10% of pregnancies [10]. Gestational diabetes usually resolves itself after pregnancy. It is caused by the hormones of pregnancy or a shortage of insulin. It causes risks to the life of baby which include abnormal weight gain before birth, breathing problems at birth, and higher obesity and diabetes risk later in life.

III. Problem Definition

In this modern era, human beings encounter different health issues. Most of the health issues are due to the food habits of the individuals. Based on the questionnaire, clinical data and sensor data, a predictive model is proposed to prevent the Diabetic Mellitus.

IV. The Proposed IDPM Model

A Diabetes Based Prediction Model plays animportant role in predicting diabetes and pre treatingdiabetic patients. It consists of three layers namelystorage layer and analytics and action layer. The layers in the proposed model for diabetic prediction are presented in Fig. 1.

The device layer consists of body sensors or wearables such as inflammation sensor, insulin pump, glucose sensor and environment sensor. These sensors are implanted in the body. These sensors are monitoring the health status of the human body.

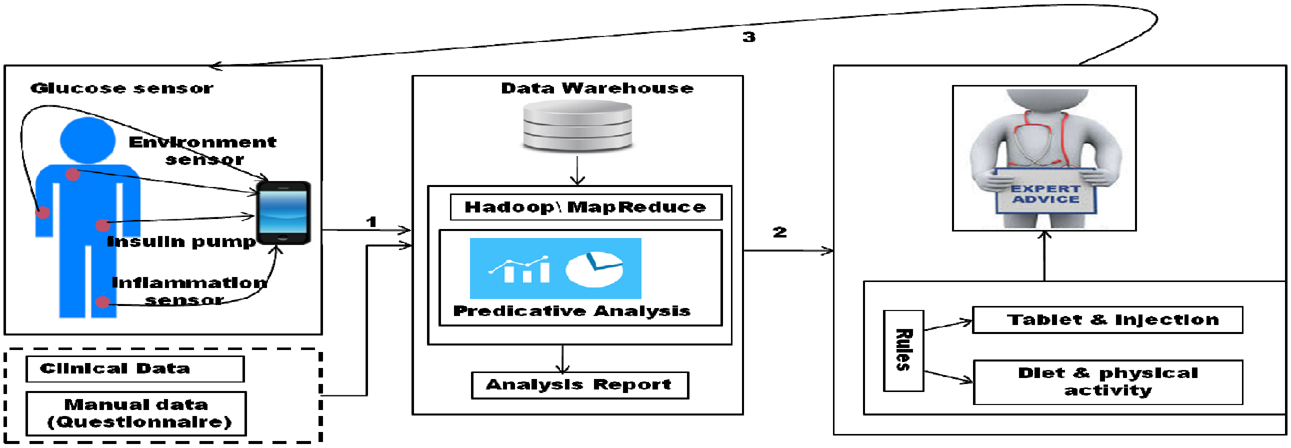
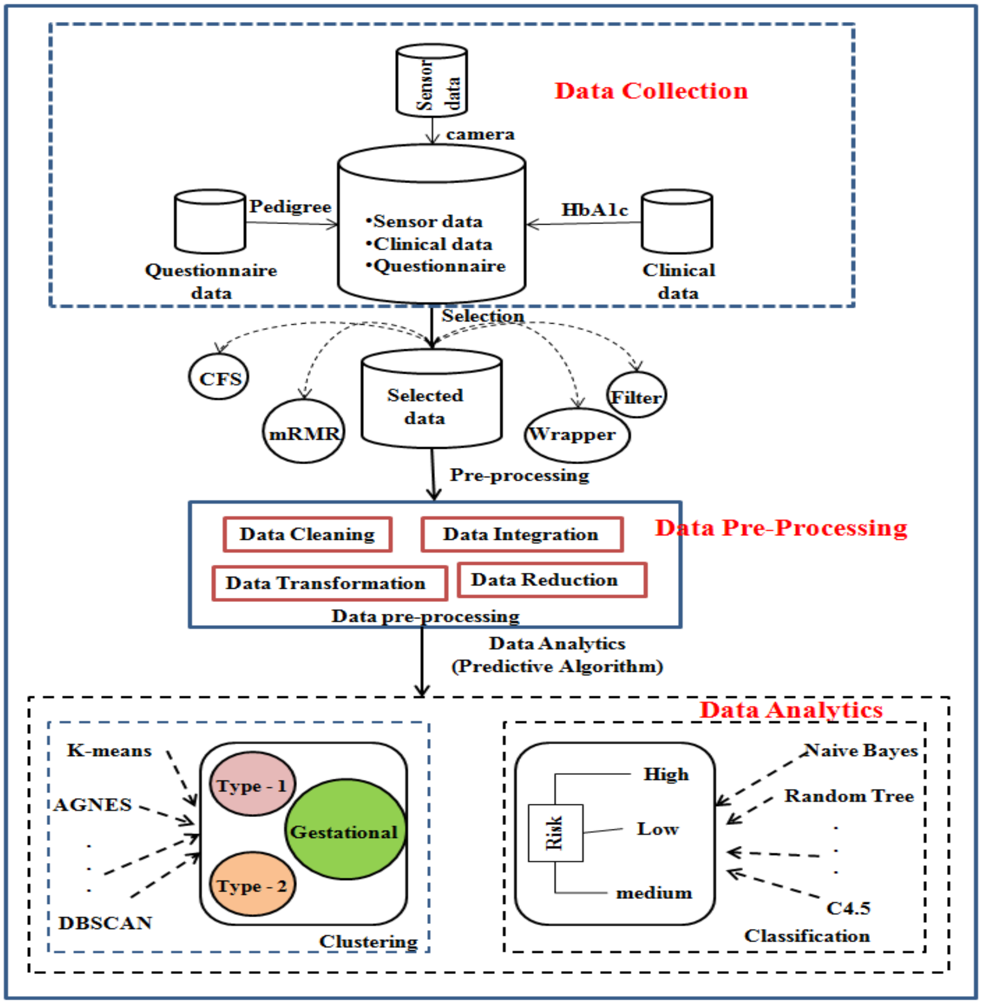


Fig. 1. The proposed IDPM Model

Each sensor has its own capability to sense the data. The data gathered from these body sensors are electronic health record vitals, lab results and medical history of patients from hospitals or clinics. The second layer is cloud storage where the electronic records are stored securely. Cloud computing can be used for: data processing, data analysis and predictive analysis using statistical and data mining techniques and tools. Distributed Data analysis is performed by Hadoop/MapReduce.The predicative analytic engine sends report to doctors for consultation. The decision making at the cloud level is based on some rules such as diet pattern, physical fitness, current medicine intake etc. This layer is called as analytics and action layer. The rules for taking actions are set by physicians and medical experts for various health related issues. The rule based consultation will also consider previous health records and medical actions already taken. Here, the text pre-processing techniques are used. The doctor checks the analysis report of the patient. The doctor sends the treatment details from the prediction report to the insulin pump actuator as shown in Fig. 2.



A. Data collection

Data collection is one of the most important stages of a research. Data collection is very demanding job which needs thorough planning, hard work, endurance, resolve and more to be able to complete the task successfully. Data Collection has two critical components. They are information gathering and decision making. Data collection is divided into two types. They are qualitative and quantitative data. Qualitative data are mostly non-numerical in nature. This means data will collect in the form of sentences or words. Quantitative data is numerical in nature and can be mathematically calculated.. In this proposed model, three types of data involved namely sensor data, clinical data and questionnaire data. The blood glucose level, body temperature, sleep time etc are collected from sensors. The clinical data like HbA1C test data are collected from clinical data. The family blood sugar history, number of time got pregnant etc., are collected from Questionnaire. Some of the data collected from sensors are qualitative and some are quantitative in nature.

B. Pre-processing

In real-time, it is very tedious to process massive amount of medical datasets containing information of individual patient health records so as to identify the disease pattern and to find out the causal association between them for planning curative actions. As a result, pre-processing of large volume of clinical data becomes essential. Data pre-processing is an important step in the data mining process. Data collection methods are largely loosely controlled, resulting in out-of-range values, missing values, etc. Observing the data which has not been carefully examined for such issues can produce misleading outcomes. If there is ample amount of incorrect and redundant information or noisy and unreliable data, then knowledge discovery becomes challenging.

Fig. 2. The Diabetes Prediction

Data pre-processing includes various steps such as cleaning, normalization, transformation, feature extraction and selection, etc. The outcome of data pre-processing is the complete data set with reduced attributes. The major drawback with clinical data set is the existence of redundant records. These redundant records cause the learning algorithm to be biased. So, eliminating redundant records is essential to enhance the detection accuracy. During pre-processing, the raw data is supplied as input and several suitable data pre-processing methods are applied thereby decreasing the invalid instances in the dataset.Data transformation and data validation are two important pre-processing techniques. They are explained below.

i. Data Transformation

In data transformation, the data are transformed or consolidated into forms appropriate for mining. Data transformation can involve Normalization, Smoothing, Aggregation and Generalization of the data.

a. *Normalization:* It is the process of converting the given values into a smallest range such as -1.0 to 1.0 or 0.0 to 1.0
b. *Smoothing*: Smoothing refers to removal of the noise from data. Smoothing techniques include binning, regression and clustering.
c. *Aggregation*: The Process of gathering information and expressing a summary form for the purpose of statistical analysis.
d. *Generalization*: Generalization is the process where low level or primitive data are replaced by higher level concepts through the use of concept hierarchies.

ii. Data validation

Data validation is defined as the assessment of all the collected data for entirety and reasonableness, and the elimination of error values. This step changes the raw data into validated data.. Data validation may be simple or complex depending on the way it is performed. Data validation can be updated either automatically or manually. The data validation helps to control the invalid data being entered into the system.

C. Predictive Analytics

The predictive analysis is carried out using classification and clustering algorithms. The clustering algorithms like K-means, DBSCAN etc. are used to cluster the pre processed data. The population is clustered based on urban/ village, male / female, educated / un-educated, diabetic / non diabetic etc. The clustered population is classified into DM high, low and medium using the classification algorithms like Bayesian network, J48, random tree etc. Some of the classification algorithms are explained below.

Bayesian network is statistical model which represents a set of variables and conditions. The relationship between the variables is carried out using Directed Acyclic Graph (DAG). In Bayesian network the nodes in DAG represent variables and edges between the nodes represent conditional dependency. Diabetic prediction using Bayesian network is influenced by the parameters. The Bayesian network for diabetic prediction with three variables (Blurred Vision (BV), Hunger and Fatigue (HF) and Family History (FH) of high blood sugar) implies Diabetes Mellitus (DM) with conditional dependency as shown in the fig. 3. The conditional dependency between the variables has two possibilities true and false respectively. The probability (pr) of occurrence of diabetics with respect to FS, BV and FV is given in equation 1.

$$Pr\ (FS\ BV\ FH) = Pr\left[\frac{FH}{(BVHF)}\right] * Pr\left[\frac{BV}{(HF)}\right] * Pr\ [HF]\ (1)$$

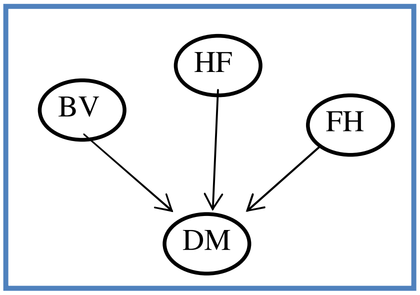


Fig. 3. Bayesian Network for Diabetes predication

The logistic regression is a statistical model used in various fields including machine learning. The decision is made by the influence of dependent variable (DV). In diabetic prediction, logistic regression utilizes general characteristics like age, Walking steps per day, Body Mass Index (BMI), Blood sugar level etc. Table 1 gives information about a set of patients (p1, p2, p3, and p4) walking steps per day and Blood sugar status. Logistic regression is suitable for the data in the table 1. The reason is, dependent variable Blood sugar status value is 1 or 0 represented to reduced or not reduced.The relationship between the steps walked per day and blood sugar status as shown in Fig. 4.

Table 1 Diabetes predication variable

Patient	Walking steps/day	Blood sugar Status
P1	4400	1
P2	3300	1
P3	2200	0
P4	0000	0

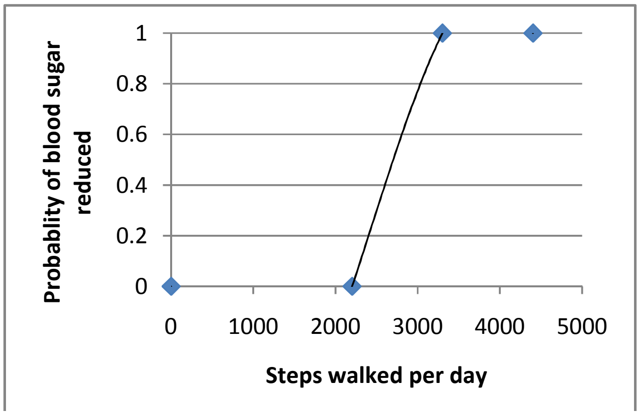


Fig. 4 Probability of blood sugar reduced versus number of steps walked.

V. Conclusion

Diabetes Mellitus is a chronic non communicable disease which has impact on human life span. A lot of datais collected from diabetic patients using IoT devices, clinical experiments and questionnaire, etc. The doctors can find value and make decisions when analysing the data accumulated from these sources. Early prediction of the deficiency will help the doctors to decide the treatment methods. Hence, a model for prediction of diabetes is proposed enabling pre-treatmentof the patients. .

Acknowledgment

This research work is financially supported by University Grants Commission, Government of India, under the Minor Research Project scheme. Ref. No.: F MRF-6517/16 (SER)/UGC).

REFERENCES

1. NDTV Food Desk, Updated: November 14, 2017 13:06 IST, available at: https://www.ndtv.com/food/world-diabetes-day-2017-number-of-diabetics-to-double-in-india-by-2023-1775180

2. Olson, Brooke. "Applying medical anthropology:Developing diabetes education and prevention programs in American Indian cultures", American Indian Culture and Research Journal, Vol:23, No:3, 1999, pp: 185-203.

3. World Health Organization. "Definition, diagnosis and classification of diabetes mellitus and its complications: report of a WHO consultation. Part 1, Diagnosis and classification of diabetes mellitus.", 1999, pp. 17-21.

4. American Diabetes Association. "Diagnosis and classification of diabetes mellitus", Vol:37, No:1 , 2014, pp : 81-90.

5. Gavin III, James R, "Report of the expert committee on the diagnosis and classification of diabetes mellitus." Diabetes care, Vol: 20, No:7, 1997, pp :963-967.

6. Diabetes Control and Complications Trial. "Intensive diabetes treatment and cardiovascular disease in patients with type 1 diabetes." The New England journal of medicine, Vol: 353, No: 25, 2005, pp: 26-43.

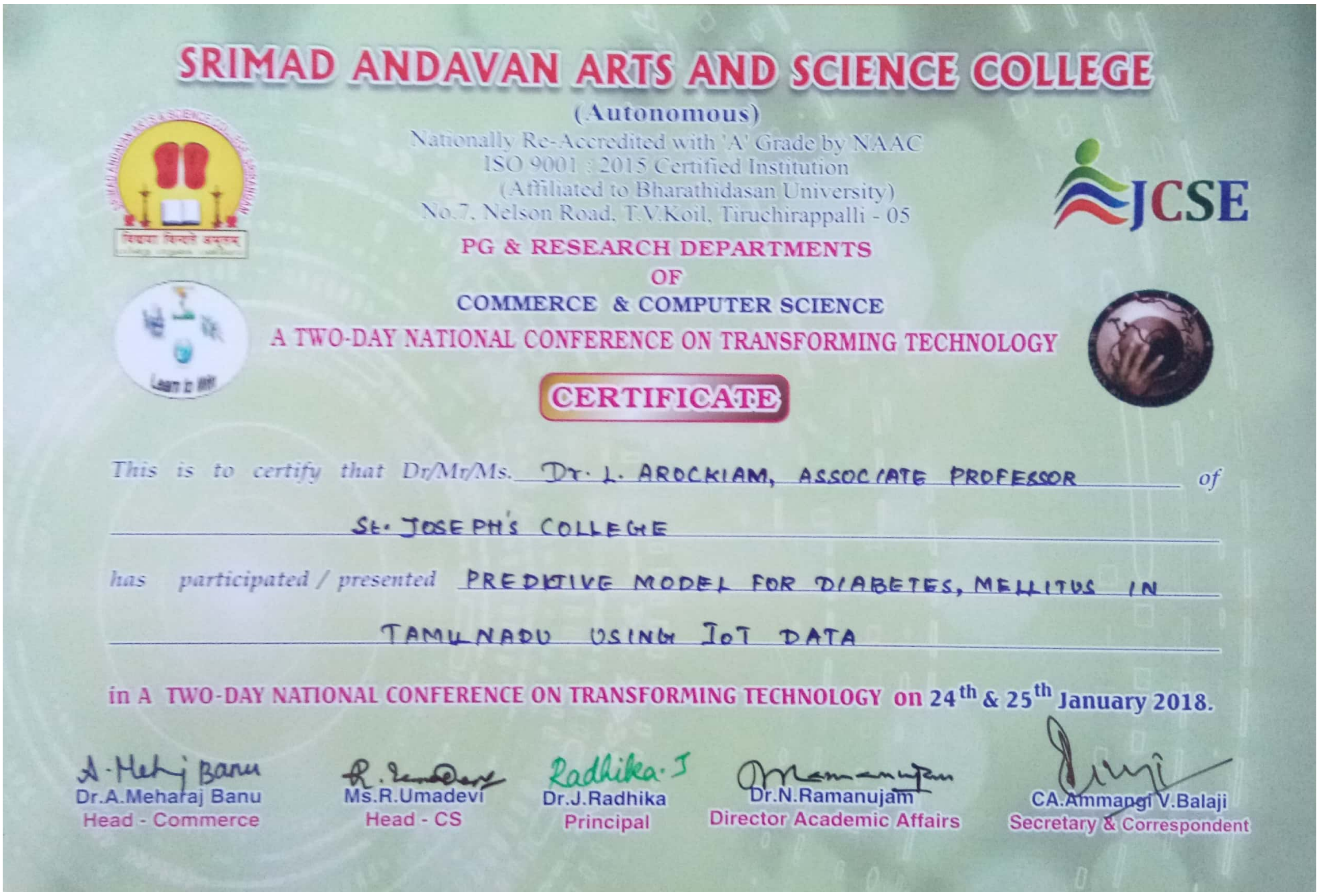
7. Alberti, G., Zimmet, P., Shaw, J., Bloomgarden, Z., Kaufman, F. Silink, M. Type 2 diabetes in the young: the evolving epidemic. Diabetes care,Vol:27,No:7,2004, pp: 1798-1811.

8. Recognising Type-2 Diabetes', Feb 2016 Available : https://www.healthline.com/ health/type-2 -diabetes /recognizing-symptoms, [Accesed : 15-jan-2018].

9. Czupryniak, Leszek, "Self-monitoring of blood glucose in diabetes: from evidence to clinical reality in Central and Eastern Europe—recommendations from the international Central-Eastern European expert group.",Diabetes technology & therapeutics , Vol:16, No:.7, 2014, pp: 460-475.

10. Alatab, S., Fakhrzadeh, H., Sharifi, F., Mirarefin, M., Badamchizadeh, Z., Ghaderpanahi, M. Larijani, B.,." Correlation of serum homocysteine and previous history of gestational diabetes mellitus.",Journal of Diabetes and Metabolic Disorders, Vol : 12, No :1,2004, pp: 34-36.

11. Gillman, M. W., Rifas-Shiman, S., Berkey, C. S., Field, A. E., &Colditz, G. A. "Maternal gestational diabetes, birth weight, and adolescent obesity".Pediatrics, Vol:111, No: 3, 2003, pp:221-226.



SRIMAD ANDAVAN ARTS AND SCIENCE COLLEGE

(Autonomous)

Nationally Re-Accredited with 'A' Grade by NAAC
ISO 9001 : 2015 Certified Institution
(Affiliated to Bharathidasan University)
No.7, Nelson Road, T.V.Koil, Tiruchirappalli - 05

JCSE

PG & RESEARCH DEPARTMENTS
OF
COMMERCE & COMPUTER SCIENCE

A TWO-DAY NATIONAL CONFERENCE ON TRANSFORMING TECHNOLOGY

CERTIFICATE

This is to certify that Dr/Mr/Ms. Dr. L. AROCKIAM, ASSOCIATE PROFESSOR *of* St. JOSEPH'S COLLEGE *has participated / presented* PREDICTIVE MODEL FOR DIABETES, MELLITUS IN TAMILNADU USING IoT DATA

in A TWO-DAY NATIONAL CONFERENCE ON TRANSFORMING TECHNOLOGY on 24th & 25th January 2018.

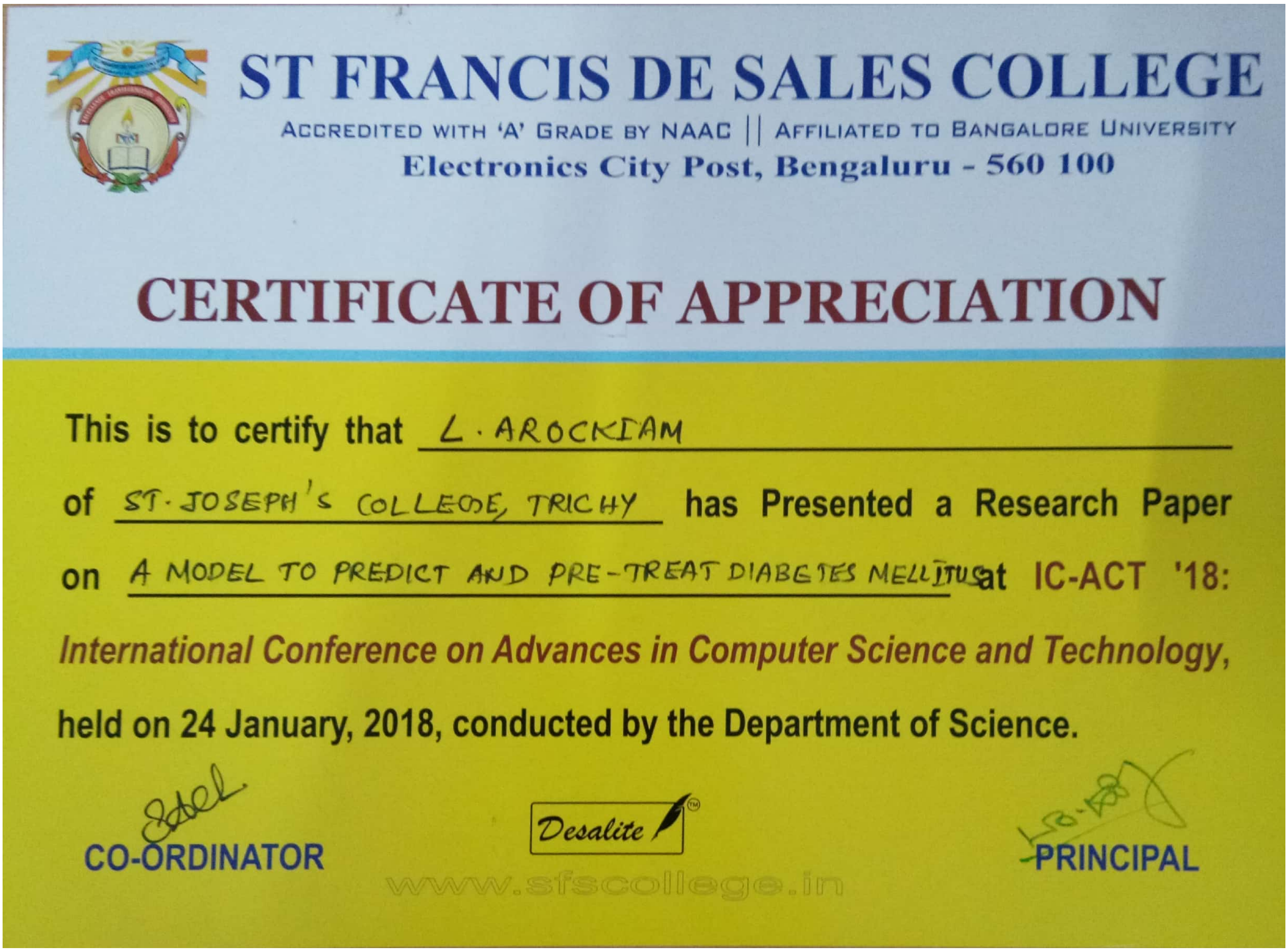
A. Mehraj Banu
Dr.A.Meharaj Banu
Head - Commerce

R. Umadevi
Ms.R.Umadevi
Head - CS

Radhika. J
Dr.J.Radhika
Principal

Ramanujam
Dr.N.Ramanujam
Director Academic Affairs

CA.Ammangi V.Balaji
Secretary & Correspondent



ST FRANCIS DE SALES COLLEGE

ACCREDITED WITH 'A' GRADE BY NAAC || AFFILIATED TO BANGALORE UNIVERSITY

Electronics City Post, Bengaluru - 560 100

CERTIFICATE OF APPRECIATION

This is to certify that L. AROCKIAM

of ST. JOSEPH'S COLLEGE, TRICHY has Presented a Research Paper

on A MODEL TO PREDICT AND PRE-TREAT DIABETES MELLITUS at IC-ACT '18:

International Conference on Advances in Computer Science and Technology,

held on 24 January, 2018, conducted by the Department of Science.

CO-ORDINATOR

Desalite™

www.sfscollege.in

PRINCIPAL

SCIENCE FOR NEW INDIA

INDIA
INTERNATIONAL
SCIENCE
FESTIVAL
2017

INDIA INTERNATIONAL SCIENCE FESTIVAL

भारतीय अंतर्राष्ट्रीय विज्ञान महोत्सव 2017

Venue: Anna University, NIOT, CSIR-CLRI, CSIR-SERC, IIT Madras

Certificate

Congratulations for being awarded TOP Poster Presenter in IISF 2017

This is to certify that ....***Mr A. Dalvin Vinoth Kumar Aron & Dr. L. Arockiam***.........................

from ....*St.Joseph's College, Tiruchirappalli*...

presented a poster titled ......*Swatchh Bharat* ***- IoT based Diabetic predication***...

....

under the theme ....*Swatchh Bharat*.................................during Young Scientists' conclave at IISF 2017 held at Chennai on October 13 - 16, 2017.

Dr. M Rajeevan
Secretary, MoES

Vijay Bhatkar
Dr. Vijay Bhatkar
President, Vijnana Bharati

सत्यमेव जयते
Ministry of Science & Technology

सत्यमेव जयते
Ministry of Earth Sciences

Government of Tamil Nadu

vibha
विज्ञान भारती
Vijnana Bharati

NATIONAL INSTITUTE OF OCEAN TECHNOLOGY
NIOT
CHENNAI
NIOT

75 Years of
CSIR