2019 IEEE Student Conference on Research and Development (SCOReD)
October 15-17, 2019, Seri Iskandar, Perak, Malaysia

Evaluation of predisposing factors of Diabetes Mellitus post Gestational Diabetes Mellitus using Machine Learning Techniques

Devi R Krishnan, Gayathri P Menakath, Anagha Radhakrishnan, Yarrangangu Himavarshini, Aparna A, Kaveri Mukundan, Rahul Krishnan Pathinarupothi, Bithin Alangot, Sirisha Mahankali, Chakravarthy Maddipati

***Abstract*—Diabetes Mellitus (DM) is one of the major global health challenges of the 21st century. It is a chronic disease leading to multiple complications bearing a lot of social, physical and financial impact on individuals and society. Gestational diabetes mellitus (GDM) is a type of DM that is developed in a few pregnant women although it usually reverts back to normalcy after delivery. However, it is well established that the risk in developing DM at a later stage in their lives increases with GDM. Very few works done in this area explore the possibility of using prognostic Machine Learning algorithms to predict occurrence of DM after GDM. In this paper, we conduct a methodical review of current practices, and then analyze GDM data from our University hospital to identify predisposing factors that could be used as inputs to different ML techniques.**

***Index Terms*—Type II Diabetes Mellitus, Gestational Diabetes Mellitus, Machine Learning**

I. Introduction

In diabetes mellitus, blood glucose metabolism is affected due to impaired insulin production by the pancreas along with increased cellular resistance. There are 2 types of Diabetes - Type I and Type II, the latter being more prevalent. Gestational Diabetes Mellitus (GDM) is seen in pregnant women and is known to be a significant predisposing factor for development of DM after delivery. According to the report by Center for Disease Control and Prevention (CDC) [1], 4.6-9.2% of pregnant women may develop GDM. 10 percent of these are found to develop Type II DM just after the delivery. The remaining women have a 35-60% chance of being diagnosed with Type II diabetes within 10-20 years. There is a high risk that the baby may also develop Type II diabetes later in life. Identifying women with greater risk of developing diabetes helps to introduce several measures to prevent or delay the onset of diabetes.

Medical literature recommends blood glucose level, Body Mass Index, age, OGTT, OGCT test results as some of the potential predictors for the development of DM after GDM. However, we notice from the literature and current clinical practice that it is rather difficult to predict the occurrence of DM after GDM with a high degree of surety. The research in the area of predicting DM post GDM using techniques of machine learning and AI is in its infancy.

In a collaborative work with physicians at our University hospital, Amrita Institute of Medical Sciences (AIMS), we conducted a 10-year follow up clinical study to identify predisposing factors that could be used as inputs to different ML techniques.

D. R. Krishnan, G. P. Menakath, A. Radhakrishnan, Y. Himavarshini, A. Avanaly, K. Mukundan are with the Dept. of Computer Science and Engineering, Amrita School of Engineering, Amritapuri, India. R. K. Pathinarupothi is with Amrita Center for Wireless Networks & Applications (AmritaWNA), Amritapuri Campus, Amrita School of Engineering, Amrita Vishwa Vidyapeetham, India. B. Alangot is with Amrita Center for Cyber Security Systems and Networks, Amrita School of Engineering, Amritapuri, India. S. Mahankali, C. Maddipati are with the Dept. of Internal Medicine, Kochi Campus, Amrita Institute of Medical Sciences, Amrita Vishwa Vidyapeetham, India.

II. Related Work

Lee *et al.* [2] determined the longstanding possibility of developing Type II diabetes for women who had GDM during pregnancy. The study assessed which of the maternal postpartum, antepartum and neonatal factors are determinants for the progression of Type II diabetes in future. It was found that the development risk of Type II diabetes increases with time for women with GDM and without GDM. However, it was seen that the patients with GDM have a higher risk of 9.6 times than the patients without GDM. Asian background women have the probability of developing diabetes two and one tenth time greater than those of non-Asian background.

Cheung *et al.* [3] established the association between fasting and postprandial hyperglycemia during pregnancy with GDM which resulted in subsequent progression of Type II DM. A sample size of 102 women with up to 8 years post their pregnancy was considered in their study. Preprandial and bedtime insulin requirements in pregnancy were used as markers for the degree of postprandial and fasting hyperglycemia respectively. It was concluded that the necessity for insulin at bedtime, which reflects persistent fasting hyperglycemia during GDM, is highly predictive for the subsequent progression of DM. Women who required intermediate-acting insulin at bedtime were 6.2 times more prone to progression of diabetes than others.

Another study was done by *Cho et al.* [4] on determining predictive factors for the progression of DM in women diagnosed with GDM during pregnancy. This study put forth that the predictor for the development of diabetes is the use of plasma homocysteine level. The follow-up examinations included two hour 75-g oral glucose tolerance tests (OGTTs), lipid profiles, homocysteine levels, anthropometric measurements, history taking, diet, and style of living of 177 patients. It was noted that fasting insulin-to-glucose ratios, glucose and

978-1-7281-2613-5/19/$31.00 ©2019 IEEE

homocysteine ranges were significantly higher in the group of diabetic people.

Carr *et al.* [5] explored the predictive power of results of 3-hour oral glucose tolerance test (OGTT) or 1-hour oral glucose challenge test (OGCT) lower than the GDM level in developing DM. It was found that those women having moderately high levels of glucose, though less than the threshold for GDM, were at a higher risk for DM. Those women with OGCT results in the range of 5.4 and 7.3 mmol/l were observed to have a 1.7- to 2-fold higher risk. Women with an OGTT, containing a single abnormal value, had an increased risk of subsequent diabetes by two folds compared to those with no abnormal values. In a meta-analysis and systematic review by Rayanagoudar *et al.* [6] for quantifying the individual risk of progression of type 2 DM for women having GDM, they state that although there is evidence that diet, lifestyle interventions, and treatment with drugs can be effective in reducing the risk, only less than a fifth of mothers with GDM undergo postpartum glucose screening. Some of the other related work in the clinical domain on management of gestational diabetes mellitus include Bhasy *et al.* [7].

As seen above, medical literature recommends blood glucose level, Body Mass Index, age, OGTT, OGCT test results as some of the potential predictors for the development of DM after GDM. However, we notice from the literature and current clinical practice that it is rather difficult to predict the occurrence of DM after GDM with a high degree of surety.

Machine learning techniques have been extensively used in disease prediction [8]–[10], though many of these works are still in the early research stage. Specifically, predicting DM post GDM using techniques of machine learning is in its infancy. Lin et al. [11] built Artificial immune recognition system (AIRS) model which is a type of supervised learning algorithm, for prediction of subsequent development of Type II diabetes in women who previously had GDM. Their model achieved an overall recall of 62.8%, while their accuracy was 60% for prediction of diabetes mellitus in GDM affected women with the patient sample size of 152 GDM patients.

Although the use of ML for predicting DM after GDM has not been explored much there are extensive works on the use of ML for prediction of DM in other domains. One among them by Barakat et al. [12] proposed a hybrid model using a sample size of 4682 persons of age 20 and above. K-means clustering algorithm, an unsupervised algorithm, was used sub-sampling to ensure the selection of representative training set and Support Vector Machine (supervised learning algorithm), did the classification which is followed by a rule-based explanation component for justification of the classification by SVM. The dataset collected includes Family history of diabetes, Sex (male/female), Body mass index (BMI) in terms of height and weight, Hip circumference, Waist circumference (Waist), Systolic blood pressure, Cholesterol, Diastolic blood pressure (BPDIAS), Fasting blood sugar (FBS), 2 hours post-glucose load from Oral Glucose Tolerance Test (OGTT). This model could achieve a specificity of 94%, accuracy of 94%, and 93% sensitivity in predicting DM. There are similar works using various ML techniques as described by Kumar et. al. [13] where C4.5, random forest (RF), the Bayesian Net classification and Multilayer Perceptron techniques were proposed for classification of diabetes on Pima Indians diabetes dataset. Dataset consisted of values of Diastolic Blood Pressure, Plasma glucose, Serum-Insulin, Triceps Skin Fold Thickness, Diabetes Pedigree Function, Age, BMI. The study experiments with the classification of data using hybrid and individual models for classification of data. Hybrid models of C4.5+RF is used which gives an accuracy of 79.31%. The mlp+bayesnet hybrid model could achieve an accuracy of 81.89% which is higher than its individual models and the other hybrid model.

SVM was used for classification of persons with diabetes and prediabetes from the dataset from National Health and Nutrition Examination Survey (NHANES) by Yu et al. [14] Dataset consisted of health history, demographic, and behavioral information from people in-home interviews. Results of SVM with several kernel functions were recorded and the best performing kernal was the RBF kernel that Classified diabetes and prediabetes whereas linear kernel performed best with the classification of prediabetes and no diabetes in patients.

We observe that there is a strong clinical requirement for prediction of DM after GDM. However, the state of the art ML techniques used for this purpose has achieved only very moderate success. We set out to push the frontiers of research in this specific area to augment decision support in current clinical practice.

III. Predisposing Factors in GDM Patients

Initially, we set out to find out what are the major clinically relevant predisposing factors in GDM patients that leads to higher chances of Type II diabetes after delivery. For this we consulted with doctors at our University hospital, the Amrita Institute of Medical Sciences. We began by collecting data of GDM patients who delivered at our hospital and then followed up by checking if they have developed Type II diabetes within a time window of 10-years post-delivery.

We obtained ten-year data from 200 women detected with GDM during pregnancy. The data collected from the hospital medical records included pre-delivery parameters obtained from in-hospital blood tests, age, weight, and other health issues. We followed up with patients using a telephonic survey to know their current Type II diabetes status. Out of 200 patients, only 90 responded back.

The pre-delivery parameters were initially scrutinized by the collaborating physicians, and then we used data analytics to further select the predisposing factors.

A. Feature Selection

From the 32 features that were selected by the physicians, only 15 features have values for more than 75 patient records. Among the 90 patients, the 77 patients records with more than 10 feature values are considered for our study. In this, 17 patients were diagnosed with Type II diabetes after pregnancy, while 60 of them had normal blood glucose counts. The group

TABLE I: Predisposing factors for prediction of DM after GDM

Features	Variables	Units	Normal Range
F1	White Blood cells (WBC)	K/uL	4.4 - 11.3
F2	Neutrophils (NEU)	%	37.0 - 80.0
F3	Lymphocytes (LYM)	%	10.0 - 50.0
F4	Monocytes (MONO)	%	0.0 - 12.0
F5	Eosinophil (EOS)	%	0.0 - 7.0
F6	Basophil (BASO)	%	0.0 - 2.5
F7	Red Blood Cells (RBC) Count	M/uL	4.04 - 6.13
F8	Hemoglobin (Hgb)	g/dl	12.2 - 18.1
F9	Hematocrit (HCT)	%	37.7 - 53.7
F10	Mean Corpuscular Value (MCV)	fL	80.0 - 97.0
F11	Mean Corpuscular Hemoglobin (MCH)	pg	27.0 - 31.2
F12	Mean Corpuscular Hemoglobin Concentration (MCHC)	g/dl	31.8 - 35.4
F13	Red blood cells distribution width (RDW)	%	11.6 - 14.8
F14	PLT Count	K/uL	150.0 - 450.0
F15	Mean Platelet Volume (MPV)	fL	6.8 - 10.0

of 15 features that are considered for our research is listed in Table I.

In order to further confirm the selection made regarding the best features from the initial set of variables, we used a data analytics parameter called mutual information (MI). It is a measure of dependence between the selected feature and the corresponding classification label. An MI value of 0 shows that the feature is completely independent to the corresponding classification result, and 1 corresponds to the feature that shows complete dependence on the result hence indicating a reduction in uncertainty of the corresponding classification label.

In mathematical terms, MI of two discrete random variables X and Y can be expressed as:

$$I(X;Y) = \sum_{y \epsilon Y} \sum_{x \epsilon X} p(x,y) \log \frac{x}{y}. \quad (1)$$

It is observed from the results of MI analysis that features F1 to F7, F10, F14, F15 gave an MI score above 0.4, while only five features F8, F9, F11, F12 and F13 showed slightly below 0.4, although it was higher than 0.2. Hence, the MI analysis reinforced the selection criteria, and we use the 15 features for building the subsequent prediction models.

IV. Building Machine Learning Models

A. Data Preprocessing

The performance of any machine learning algorithm significantly depends upon the organization and distribution of data.

We employed the following steps for data processing:

1) Formatting: The data of each patient was extracted from OXPS file format (from hospital information system) into a CSV file which was suitable for our machine learning program.
2) Data imputation: There were patients for whom few of the feature values were missing. The missing numerical values were substituted with respective median of the features.

The resulting 15 feature dataset of 77 patient was vectorized into 15-elements feature vector for each patient and labeled as P (Diabetes after GDM) or N (no diabetes after GDM). These vectors were then used as inputs to different machine learning models. For training and testing we used stratified k-fold cross-validation. Using stratified k-fold cross-validation, the dataset is split into k equal folds with same proportions of class labels. One of these folds is made the testing data, the (k−1) folds forms the training data. This validation occurs iteratively with three different random states considering each fold as testing data and the rest as training data. We used the following ML techniques with stratified 5-fold cross-validation.

B. Support Vector Machine (SVM)

This supervised machine learning technique is used widely for classification problems, where the classifier constructs a hyperplane that best divides the data points into different classes. The dimension of the hyperplane will depend on the number of features of our dataset.

C. Naive Bayes

Naive Bayes uses a probabilistic approach and it is usually employed when the data is high and the attributes are independent of each other and hence the name naive. The classifier calculates the probabilities for every factor and the outcome with the highest probability is selected. We used Gaussian model for our dataset.

D. Other Models

We used other classification models including decision trees and random forests. Decision trees are constructed by splitting the dataset based on different conditions and classify the examples by sorting them down to the leaf node where it provides the result of the classification. Random forest generates a forest with a number of decision trees. Apart from these, algorithms such as AdaBoost with DT, Adaboost with SVM was also used for prediction.

AdaBoost is a boosting algorithm developed to improve the performance of machine learning algorithms. It is best used with decision trees on binary classification problems. Figure 1 depicts the architecture of our approach.

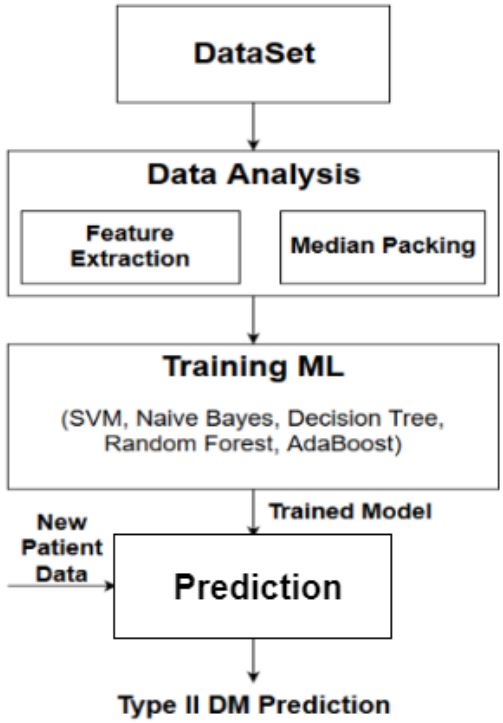

Fig. 1: Pipeline architecture of Machine Learning prediction model for DM after GDM.

TABLE II: Sensitivity, specificity, accuracy and F1-score formulae

Accuracy	(TN + TP) / (TN + TP + FN + FP)
Sensitivity	TP / (TP + FN)
Specificity	TN / (TN + FP)
F1-score	2* ((precision*recall) / (precision+recall))

V. Evaluation Results

The machine learning models are evaluated using the standard measures of accuracy, sensitivity, specificity (see Table II). Accuracy is the ratio of the sum of true positives and true negatives to the total cases examined. Sensitivity is the probability of patients with diabetes to be identified as having diabetes, whereas specificity measures the proportion of patients without diabetes to be diagnosed as not having diabetes. The above measures are calculated using true positive (TP), true negative (TN), false positive (FP) and false negative (FN).

The results from our study are summarized in Table III. All the models achieved a sensitivity score of above 0.8, which meant that the tool correctly identified most of the GDM patients at risk of developing DM. The specificity score varied from a minimum of 0.12 to a maximum of 0.23, which shows that the models predicted false positive cases too. All the models demonstrated $F1-score$ in the range of $0.67-0.97$. The $F1-score$ shows the performance of each model taking both false positives and false negatives into account.

TABLE III: Performance of different machine learning models in predicting DM after GDM.

Model	Sensitivity	Specificity	Accuracy	F1-score
Random Forest	0.94	0.12	0.75	0.7
Naive Bayes	0.93	0.14	0.75	0.7
SVM	0.89	0.12	0.72	0.67
Decision Tree	0.82	0.14	0.64	0.97
AdaBoost	0.88	0.23	0.74	0.72

VI. Discussion

The experimental results obtained in predicting DM after GDM show that ML classifiers like Random Forest, Gaussian Naive Bayes and SVM models performed better than Decision Tree. It is seen that the average accuracy of the ML models ranged from 0.66 for Decision Trees (with a standard deviation of 0.109) to 0.75 for both Random Forests and Gaussian Naive Bayes (with a standard deviation of around 0.05). The constraint of our research is a biased dataset with 17 patients diagnosed with Type II diabetes (less than 25% of the dataset size) and 15 features data. Therefore, our experimental results show that predicting DM after GDM using ML requires much more clinical data than presently collected as well as better understanding about the underlying physiological processes.

VII. Conclusion

In this study, we tested various ML classification models on 15 feature dataset of 77 patients. In clinical practice, predicting the risk of DM after GDM is a challenging task. In our research, we make use of prognostic machine learning(ML) techniques that inputs 15 predisposing factors and outputs the risk of DM. The results from our study demonstrated that Random Forest and Gaussian Naive Bayes classifies the data comparatively better than the other models.

References

[1] C. L. DeSisto, S. Y. Kim, and A. J. Sharma, "Peer reviewed: Prevalence estimates of gestational diabetes mellitus in the united states, pregnancy risk assessment monitoring system (prams), 2007–2010," *Preventing chronic disease*, vol. 11, 2014.

[2] A. J. Lee, R. J. Hiscock, P. Wein, S. P. Walker, and M. Permezel, "Gestational diabetes mellitus: clinical predictors and long-term risk of developing type 2 diabetes: a retrospective cohort study using survival analysis," *Diabetes care*, vol. 30, no. 4, pp. 878–883, 2007.

[3] N. W. Cheung and D. Helmink, "Gestational diabetes: the significance of persistent fasting hyperglycemia for the subsequent development of diabetes mellitus," *Journal of Diabetes and its Complications*, vol. 20, no. 1, pp. 21–25, 2006.

[4] N. H. Cho, S. Lim, H. C. Jang, H. K. Park, and B. E. Metzger, "Elevated homocysteine as a risk factor for the development of diabetes in women with a previous history of gestational diabetes mellitus: a 4-year prospective study," *Diabetes care*, vol. 28, no. 11, pp. 2750–2755, 2005.

[5] D. B. Carr, K. M. Newton, K. M. Utzschneider, J. Tong, F. Gerchman, S. E. Kahn, and S. R. Heckbert, "Modestly elevated glucose levels during pregnancy are associated with a higher risk of future diabetes among women without gestational diabetes mellitus," *Diabetes Care*, vol. 31, no. 5, pp. 1037–1039, 2008.

[6] G. Rayanagoudar, A. A. Hashi, J. Zamora, K. S. Khan, G. A. Hitman, and S. Thangaratinam, "Quantification of the type 2 diabetes risk in women with gestational diabetes: a systematic review and meta-analysis of 95,750 women," 2016.

[7] B. Bhasy, L. Varghese, A. Krishnan, and L. Viswanath, "Perceived risk and abilities for health practices related to prevention of type ii diabetes mellitus among postnatal mothers with gestational diabetes mellitus," *Asian Journal of Pharmaceutical and Clinical Research*, vol. 11, no. 3, pp. 354–356, 2018.

[8] R. Ani, G. Sasi, U. R. Sankar, and O. S. Deepa, "Decision support system for diagnosis and prediction of chronic renal failure using random subspace classification," in *2016 International Conference on Advances in Computing, Communications and Informatics, ICACCI 2016*, 2016, pp. 1287–1292.

[9] R. K. Pathinarupothi, J. Dhara Prathap, E. S. Rangan, A. E. Gopalakrishnan, R. Vinaykumar, and K. P. Soman, "Single sensor techniques for sleep apnea diagnosis using deep learning," in *Proceedings - 2017 IEEE International Conference on Healthcare Informatics, ICHI 2017*, 2017, pp. 524–529, cited By :3.

[10] R. K. Pathinarupothi, P. Durga, and E. S. Rangan, "Iot-based smart edge for global health: Remote monitoring with severity detection and alerts transmission," *IEEE Internet of Things Journal*, vol. 6, no. 2, pp. 2449–2462, 2019.

[11] H.-C. Lin, C.-T. Su, and P.-C. Wang, "An application of artificial immune recognition system for prediction of diabetes following gestational diabetes," *Journal of medical systems*, vol. 35, no. 3, pp. 283–289, 2011.

[12] N. Barakat, A. P. Bradley, and M. N. H. Barakat, "Intelligible support vector machines for diagnosis of diabetes mellitus," *IEEE transactions on information technology in biomedicine*, vol. 14, no. 4, pp. 1114–1120, 2010.

[13] A. Kumar Dewangan and P. Agrawal, "Classification of diabetes mellitus using machine learning techniques," *Int. J. Eng. Appl. Sci*, vol. 2, no. 5, pp. 145–148, 2015.

[14] W. Yu, T. Liu, R. Valdez, M. Gwinn, and M. J. Khoury, "Application of support vector machine modeling for prediction of common diseases: the case of diabetes and pre-diabetes," *BMC medical informatics and decision making*, vol. 10, no. 1, p. 16, 2010.