International Journal of
INTELLIGENT SYSTEMS AND APPLICATIONS IN ENGINEERING

ISSN:2147-6799 www.ijisae.org Original Research Paper

Gestational Diabetes Detection Using Machine Learning Algorithm: Research Challenges of Big Data and Data Mining

K. Arun Kumar[*1], R Rajalakshmi[2], Shashikala H K[3], Prof (Dr) Maithli Ganjoo[4], Dr Aman Vats[5], Dr. Rajneesh Tyagi [6]

Submitted: 14/08/2022 **Accepted:** 20/11/2022

Abstract: The prevalence of gestating moms from various countries and ethnic groups worldwide who have gestational diabetes mellitus (GDM), a disorder characterised by abnormally high blood glucose levels, has rapidly increased. This research propose novel technique in gestational diabetes detection using machine learning technique in big data with data mining analytics. Here the input has been collected as data of pregnant women for diabetes prediction. This data has been processed for dimensionality reduction and normalization. Then it has been segmented and feature fused using attention mechanism based weighted convolutional neural networks. The experimental analysis has been carried out in terms of accuracy, precision, recall, F-1 score and AUC. Proposed technique attained accuracy of 96%, precision of 92%, recall of 85% and F_1 score of 89%, AUC of 71%.

Keywords: *Gestational diabetes Mellitus, machine learning technique, big data, data mining analytics, convolutional neural networks*.

1. Introduction

The disease known as GDM is distinct from type 1 and type 2 diabetes. It is referred to as gestational when it happens during pregnancy. Many mothers with diabetes will experience remission after giving birth to their child [1]. Up to 25% of pregnancies are affected by gestational diabetes. Pregnancy hormones start to have an impact on the regular procedures and actions involved in producing insulin around the 20th week of gestation. GDM is diagnosed with blood testing. The two largest indicators of diabetes are blood sugar and a woman's weight [2]. Another cause of the condition is high blood pressure. Different ML methods, including LR, RF, XGBoost, SVM, and ANN studies, have recently been used by researchers to predict GDM. These machine learning algorithms can categorise risk factors, determine how features are correlated, build risk prediction models, and forecast the development of disease [3].

[1]*Assistant Professor, Computer Science and Engineering, Sree Vidyanikethan Engineering College, Tirupathi Arunkumar.K@Vidyanikethan.Edu*
[2] *Department Of Ece , Panimalar Engineering College, Chennai , India Rajeeramanathan@Gmail.Com*
[3] *Assistant Professor, Department Of Computer Science And Engineering Jain(Deemed-To-Be University), Bangalore, India Hk.Shashikala@Jainuniversity.Ac.In*
[4] *Prof And Dean, Faculty Of Media Studies And Humanities, Manav Rachna International Institute Of Research And Studies, Faridabad Mganjoooffice@Gmail.Com,*
[5] *Professor And Hod, Department Of Journalism And Mass Communication, Faculty Of Media Studies And Humanities, Manav Rachna International Institute Of Research And Studies, Faridabad, Haryana, Amanvats@Live.Com; Vatsaman@Gmail.Com,*
[6] *Department Of Agriculture, Sanskriti University, Mathura, Uttar Pradesh, India, Dean.Soa@Sanskriti.Edu.In*

2. Related Works

A significant amount of research is conducted in nowadays to predict diabetes using machine learning techniques. LDA, QDA, NB, GPC, SVM, ANN, AB, LR, DT and RF are some of the dimensionality reduction and cross-validation methods that were proposed by the authors in [4]. To enhance performance of ML model, extensive experiments were also conducted to reject outliers and fill in missing values by estimating the mean and the median. These efforts yielded the largest area under the curve of 0.930 achievable. In [5], authors used decision trees, SVMs, and Nave Bayes classifiers to compute diabetes as accurately as possible. After testing all three classifiers, it was determined that Nave Bayes was the best of the three, with an AUC of 0.819. In the modelling process, [6] took into account the effect of number or accessibility of elements incorporated on clinical usefulness of methods. In addition, [7] are only researchers to have taken explainability into account while designing their models. [8] created meta-learning algorithms for identifying the presence of diabetes. Diabetes can be predicted using the CART, Adaboost, Logiboost, and grading learning algorithms [9]. According to [10] gestational diabetes (GDM) is a condition that often manifests itself during the second to third trimester of pregnancy.

3. System Model

This section discuss novel technique in gestational diabetes detection using machine learning technique in big data with data mining analytics. Here the input has been collected as data of pregnant women for diabetes prediction. This data has been processed for dimensionality reduction and normalization. Then it has been segmented and feature fused using attention mechanism based weighted convolutional neural networks. The overall proposed architecture is shown in figure-1.

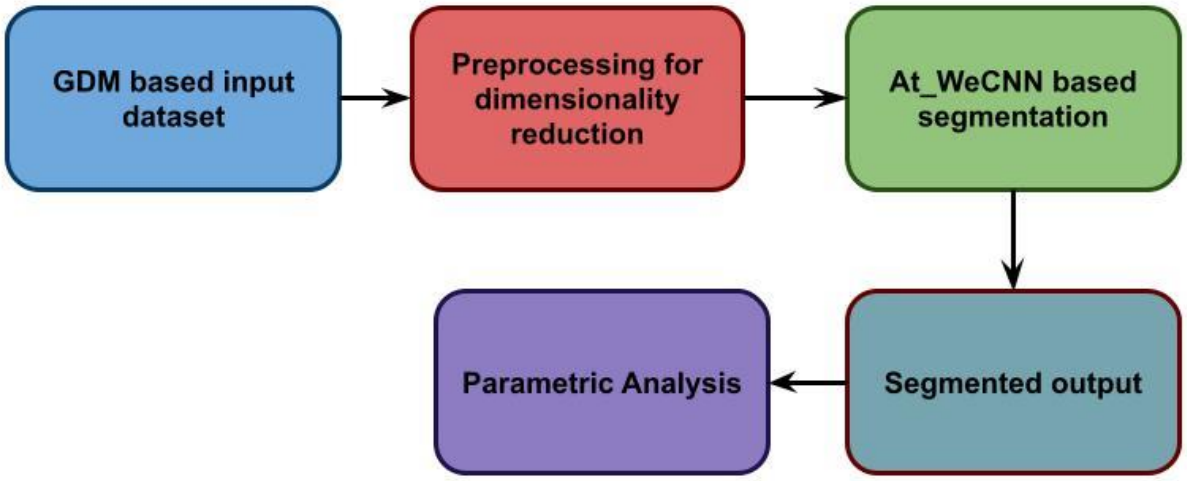

Figure 1. Overall proposed architecture

Attention mechanism based weighted convolutional neural networks (Att_WeCNN):

Here, the attention module's job is to draw attention to Fcon's key characteristics. Squeeze-and-Excitation (SE) attention is the most well-liked of all the attention modules now in use. In particular, as demonstrated in Figure 2, SE attention first aggregates the input feature map into a channel descriptor using a 2D global average pooling. The relevance of every channel in input feature map is then determined using the channel descriptor that is then supplied to two fully connected layers. In order to generate two separate feature maps in response to the loss of positional information, the coordinate attention module pools the input feature map along the horizontal and vertical axes using two pooling layers with kernel sizes of (1, W) and (H), respectively. Then, using 1 × 1 convolution layers, these two feature maps are converted into two attention maps. In order to accentuate the features of targets, the input feature map is multiplied with both attention maps.

At same time as obtaining global features and semantic associations of text, weighted local features are also acquired. When compared to single features, the multi-feature representation can gather both local and global contextual semantic information, weight the local features to reduce secondary feature representation, and improve the diversity of feature representation to improve acquisition of text data. Finally, fully connected layers in combination with Softmax are used to produce the text classification results.

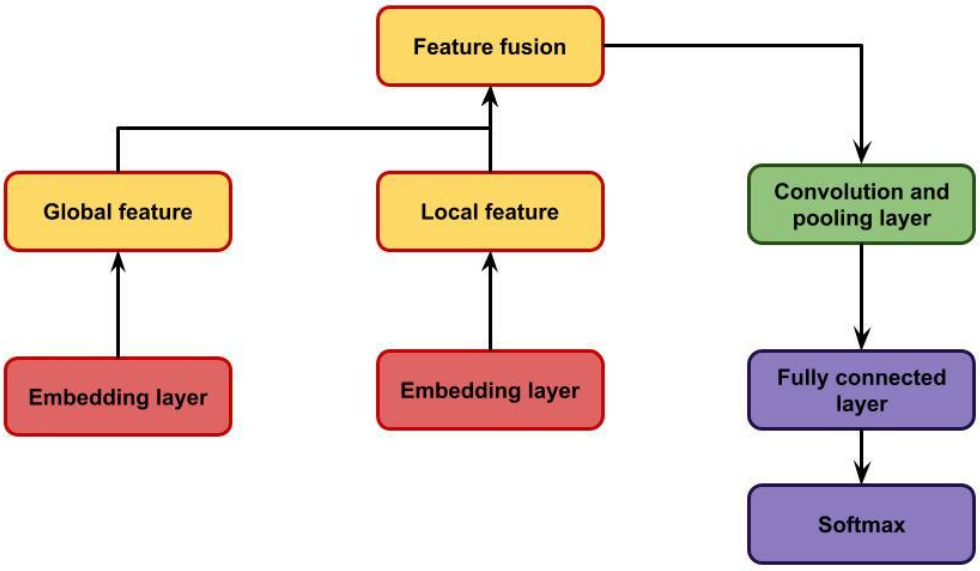

Figure-2 architecture of Att_WeCNN

In a subsequent series of row representation form, the convolution layer applies convolution operation to the word vector it receives from embedding layer. Let's assume that w is a weight matrix of dimension $w \in R^{h*m}$ and that h words are chosen at a time to execute the convolution operation as eq. (1):

$C_i = f(X_{i+h-1*w+b_i})$ (1)

The feature map created with h words repeatedly is called $C_i \in R^{n-h+1}$ the non-linear Relu function is called f, and the bias term is called bi. Then, we apply max-pooling operation to the convolutionally generated features map, which divides features map in half by choosing features with the highest activation values, as eq. (2):

$p_i = maxC_i$ (2)

Here,$p_i \in R^{n-h+1/2}$ is new feature map. We employ three convolutional layers with filter sizes of 3, 4, and 5 to extract various level features from convolutional layer. Following max-pooling procedure, we fuse features from various levels of convolution layers to obtain final multilevel feature fusion output

from CNN. The outcome is a probability distribution across all categories that looks like this by eq. (3):

$$Y_i = \frac{\exp(h_i)}{\sum_{j=1}^{c} \exp(h_i)} \quad (3)$$

the probability distribution of the I th class's overall classes is results Yi. We used crossentropy as a loss function to calculate the difference between the input data's diabetes and non-diabetes distributions as eq. (4):

$$loss = \sum_{s=T} \sum_{i=1}^{\hat{s}} Y_i^t(C) \log(Y_i(C)) \quad (4)$$

where s 0 is GDM classes, T is training corpus text, Y t (C) is element matching to true emotion of phrase, Y (C) iselement corresponding to predicted GDM of data. With pre-trained word vectors from GloVe as our starting point, we then train and update model's parameters using stochastic gradient descendent approach.

Dataset description:The following characteristics are part of the dataset (1-8 attributes as input and last attribute as target variable) how many times you've been pregnant, your plasma sugar level at a 2-hour oral glucose tolerance test, Body mass index, Triceps skin fold thickness (mm), 2-hour serum insulin (mu U/ml), Diabetes pedigree function, and age (years). Whether a patient's test results are positive or negative will determine the class.There are 768 cases available in PIDD altogether. 192 patients had readings for skin fold thickness, 5 patients had readings for glucose, 11 patients had readings for body mass index, 28 others had data for diastolic blood pressure, and 140 patients had readings for serum insulin.There were 392 cases left with no missing values after these cases were deleted (130 tested positive cases and 262 tested negative).

4. Experimental Analysis

Table-1 Comparative analysis between proposed and existing technique

Parameters	LDA	GDM	GDD_MLA
Accuracy	91	93	96
Precision	85	89	92
Recall	79	83	85
F1_Score	81	85	89
AUC	65	68	71

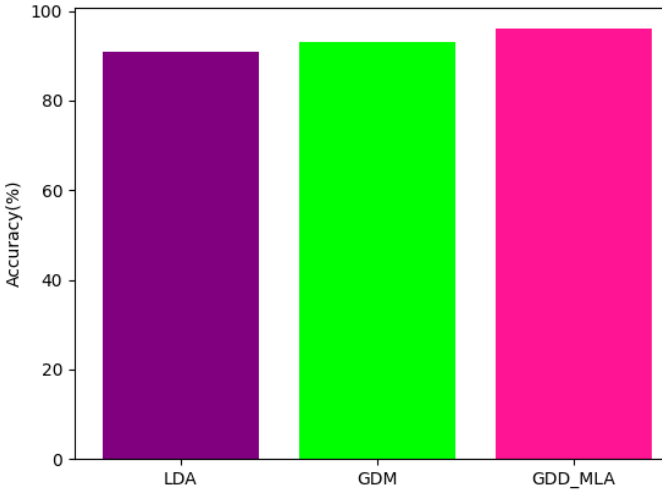

Figure-3 Comparison of accuracy

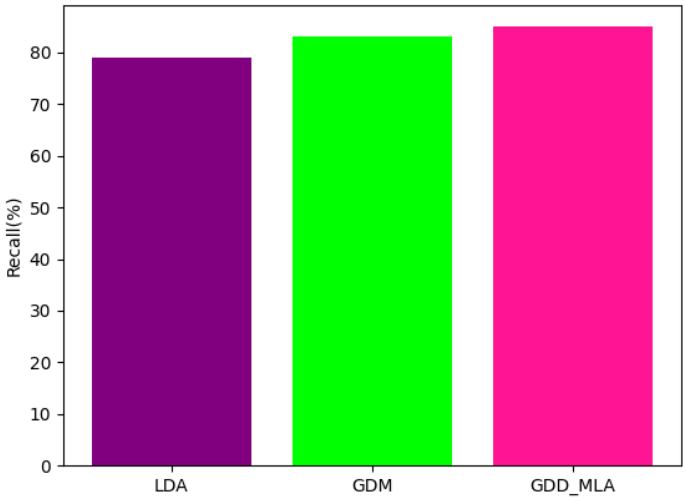

Figure-4 Comparison of Recall

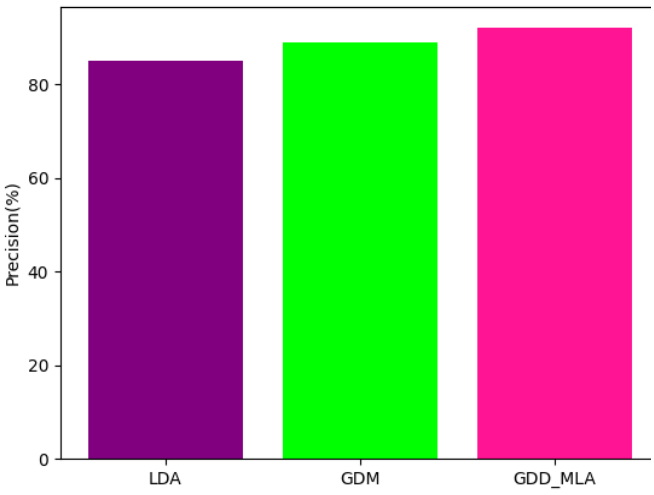

Figure- 4 Comparison of precision

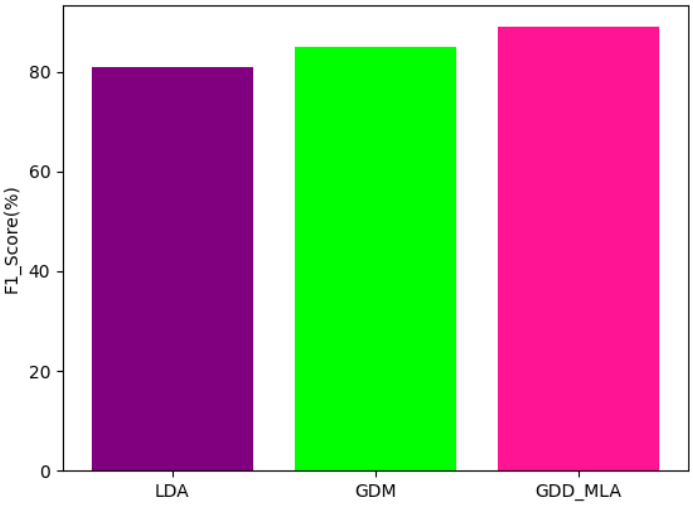

Figure-5 Comparison of F-1 score

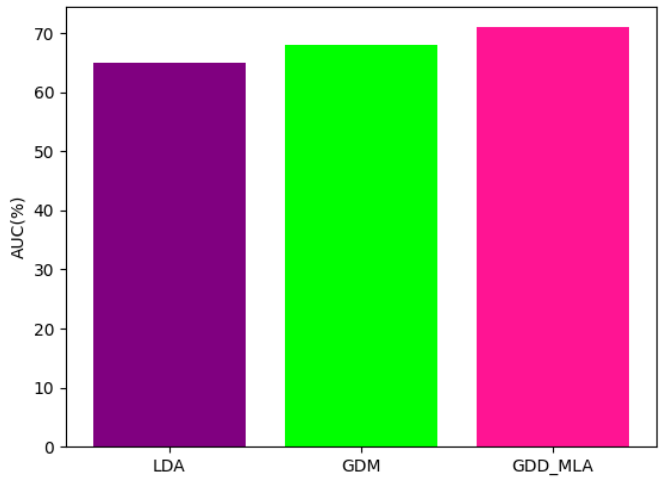

Figure-6 Comparison of AUC

From above figure 3-6 the comparative analysis between proposed and existing technique is shown in terms of accuracy, precision, recall, F_1 score and AUC. Here proposed technique attained accuracy of 96%, precision of 92%, recall of 85% and F_1 score of 89%, AUC of 71%. One specification for assessing classification methods is accuracy. Official description of accuracy is as follows: Total number of accurate guesses is equal to total number of accurate guesses. By dividing number of accurate predictions by overall sample size, we may determine accuracy. Our model was 44 percent accurate on this multiclass problem, according to the outcome. One indicator of the model performance is precision, or the quality of a successful prediction. Total number of accurate positive predictions is divided by total number of real positives to determine precision. Recall literally refers to how many right hits were also discovered, or how many genuine positives were remembered. Precision is the percentage of returning hits that were true positive, or correct hits. The recall is determined by comparing the proportion of correctly labelled Positive samples to all Positive samples. Model ability to make distinctions Recall is used to measure positive samples. As more positive samples are identified, the recall rises. The F1 score is computed using harmonic mean of recall as well as precision. Recall that harmonic mean serves as a replacement for the arithmetic mean, which is utilised more frequently. It frequently helps when figuring up an average rate. We figure out the F1 score mean precision and recall.

5. Conclusion

This research propose novel technique in gestational diabetes detection using machine learning technique in big data with data mining analytics. Here the input data has processed for dimensionality reduction segmented with feature fused using attention mechanism based weighted convolutional neural networks. The medical examiner's ID number did not match medical examination record in source data. This portion of record was eliminated to guarantee accuracy of the information. In order to maintain the quality of the data, records with a lot of default values are also eliminated. The metabolite candidate with the highest score is preferred for the assignment of mass spectral characteristics.

References

[1] Zhang, Z., Yang, L., Han, W., Wu, Y., Zhang, L., Gao, C., ... & Wu, H. (2022). Machine learning prediction models for gestational Diabetes mellitus: Meta-analysis. Journal of medical Internet research, 24(3), e26634.

[2] Wu, Y. T., Zhang, C. J., Mol, B. W., Kawai, A., Li, C., Chen, L., ... & Huang, H. F. (2021). Early prediction of gestational diabetes mellitus in the Chinese population via advanced machine learning. The Journal of Clinical Endocrinology & Metabolism, 106(3), e1191-e1205.

[3] Xiong, Y., Lin, L., Chen, Y., Salerno, S., Li, Y., Zeng, X., & Li, H. (2022). Prediction of gestational diabetes mellitus in the first 19 weeks of pregnancy using machine learning techniques. The Journal of Maternal-Fetal & Neonatal Medicine, 35(13), 2457-2463.

[4] Du, Y., Rafferty, A. R., McAuliffe, F. M., Wei, L., & Mooney, C. (2022). An explainable machine learning-based clinical decision support system for prediction of gestational diabetes mellitus. Scientific Reports, 12(1), 1-14.

[5] Liu, Y., Wang, Z., & Zhao, L. (2021). Identification of diagnostic cytosine-phosphate-guanine biomarkers in patients with gestational diabetes mellitus via epigenome-wide association study and machine learning. Gynecological Endocrinology, 37(9), 857-862.

[6] Araya, J., Rodriguez, A., Lagos-SanMartin, K., Mennickent, D., Gutiérrez-Vega, S., Ortega-Contreras, B., ... & Guzmán-Gutiérrez, E. (2021). Maternal thyroid profile in first and second trimester of pregnancy is correlated with gestational diabetes mellitus through machine learning. Placenta, 103, 82-85.

[7] Kumar, M., Chen, L., Tan, K., Ang, L. T., Ho, C., Wong, G., ... & Karnani, N. (2022). Population-centric risk prediction modeling for gestational diabetes mellitus: A machine learning approach. Diabetes Research and Clinical Practice, 185, 109237.

[8] Liu, Y., Geng, H., Duan, B., Yang, X., Ma, A., & Ding, X. (2021). Identification of diagnostic CpG signatures in patients with gestational diabetes mellitus via epigenome-wide association study integrated with machine learning. BioMed research international, 2021.

[9] Wang, J., Lv, B., Chen, X., Pan, Y., Chen, K., Zhang, Y., ... & Liu, Y. (2021). An early model to predict the risk of gestational diabetes mellitus in the absence of blood examination indexes: application in primary health care centres. BMC Pregnancy and Childbirth, 21(1), 1-8.

[10] Eleftheriades, M., Chatzakis, C., Papachatzopoulou, E., Papadopoulos, V., Lambrinoudaki, I., Dinas, K., ... & Sotiriadis, A. (2021). Prediction of insulin treatment in women with gestational diabetes mellitus. Nutrition & Diabetes, 11(1), 1-5.

.