Using Visceral Adipose Tissue Measurements to build Classification Models for Gestational Diabetes Mellitus

Aditi Deokar[1]
[1]Boston University Academy

Abstract

In gestational diabetes mellitus (GDM), temporary glucose intolerance develops during pregnancy, and can cause a wide variety of adverse maternal and fetal outcomes, from morbidity to future obesity and diabetes in the mother or the child. The incidence of GDM is increasing, and early detection methods for GDM will help patients receive treatment sooner to avert some of these outcomes. This study used a variety of machine learning techniques to predict the risk of GDM in a cohort of 100 patients of varying gestational ages with clinical information. Notably, we used visceral adipose tissue (VAT) measurements as a risk factor, which have not to our knowledge previously been used in a machine learning predictive model for GDM. The classification techniques used included logistic regression, support vector machine, k-nearest neighbors, decision tree and multi-layer perceptron, as well as the ensemble techniques random forest, gradient boosting, and XGBoost. Of the machine learning models tested, gradient boosting performed the best, with a cross-validation recall of 71.4% and AUC-ROC score of 0.864. The results of this model outperformed existing literature. VAT was the most important feature for the gradient boosting algorithm, indicating its importance in gestational diabetes prediction. Future work could investigate other risk factors for gestational diabetes, such as diet, socioeconomic status, environmental factors, and lifestyle choices, which might add value to future models, and also analyze the changes in VAT and other predictive factors over different gestational ages to develop a model that accounts for these changes.

Introduction

Gestational diabetes mellitus (GDM) is a disease in which temporary glucose intolerance develops during pregnancy (Buchanan & Xiang, 2005). The incidence of GDM is increasing in several ethnic groups, including non-Hispanic whites, Hispanics, African Americans, and Asians (Dabelea et al., 2005). This is most likely due to the increase in obesity throughout the world, as fast food consumption is a risk factor for GDM (Dominguez et al., 2014). GDM is associated with an increased risk of maternal and fetal morbidity, as well as more specific complications including preterm delivery, need for cesarean section, infants being large for their gestational age, and future obesity and diabetes in the mother and the child (Buchanan & Xiang, 2005; Xiong et al., 2001).

GDM is currently diagnosed using an oral glucose tolerance test (OGTT) between 24 and 28 weeks of pregnancy (Sukumaran et al., 2014), but because this time period is after fetal and placental development, earlier detection and treatment may better alleviate the detrimental outcomes associated with GDM (Ye et al., 2020). Machine learning can be used to determine from clinical factors which patients are at higher risk for GDM and should be tested and monitored early on in pregnancy. This is a cost-effective alternative to testing all pregnant women earlier, because GDM manifests in mid to late pregnancy for most women, so earlier testing would require multiple rounds of testing in all women (Ye et al., 2020).

Ye et al. (2020) compared the performance of several other machine learning models with that of logistic regression in predicting the risk of GSM from 104 clinical factors. They found that none of the machine learning models outperformed logistic regression in their area under the receiver

operating characteristic (AUC-ROC) curve. Gradient boosting did best, with a 0.709 AUC-ROC score as compared to the 0.7351 AUC-ROC score of logistic regression.
Risk factors for GDM include high blood pressure (Hedderson & Ferrara, 2008), first fasting glucose (Corrado et al., 2012), age > 35 years, and obesity (Xiong et al., 2001). Although first fasting glucose is indicative of glucose intolerance, it is a predictive, not diagnostic factor for GDM because there is no clear cut-off value (Corrado et al., 2012). This study aimed to use information on these common risk factors, along with visceral adipose tissue measurements, which are predictive of GDM (Martin et al., 2009) but have not to our knowledge been used in machine learning models, to predict the risk of gestational diabetes in pregnant women.

Methods

We used the database "Visceral adipose tissue measurements during pregnancy" (Rocha et al., 2020) from PhysioNet (Goldberger et al., 2000), which contains clinical information on 133 mothers with and without gestational diabetes mellitus (GDM). The clinical information included maternal age, existence of previous diabetes mellitus (DM), blood pressure, visceral adipose tissue (VAT) measurement in the periumbilical region, gestational age at time of inclusion, number of pregnancies, level of first fasting glucose, and pregestational body mass index (BMI). The following pregnancy outcomes were also included: gestational age at birth, type of delivery (vaginal or caesarean section), child birth weight and the diagnosis of GDM.

Patients who had missing data were excluded, leaving 101 patients in the database. As there was only one mother with previous DM, she was also excluded, and the existence of previous DM was not used in the models. Of the remaining 100 patients, 13 had GDM.

These 100 patients were then divided into training (80%) and testing (20%) sets. The training set contained 10 patients with GDM, and the testing set contained 3. All of the data was then normalized using Python scikit-learn library's StandardScaler, which subtracted the mean and divided by the variance for each of the clinical features. Because the data was imbalanced, data for the GDM patients was upsampled using the Python imblearn library (this increased the weight given to the GDM patient data so that both classes became more comparable during model training while using the true distribution during validation) before hyperparameters of each of the machine learning models (described below) were tuned using scikit-learn's GridSearchCV. GridSearchCV performs a cross-validated grid-search across all hyperparameter combinations given and determines the best score. A series of machine learning models from the Python scikit-learn library and XGBoost from the Python XGBoost library were then trained with these hyperparameters and cross-validated with 7 folds. The number of folds was chosen as 7 to have a reasonable number of observations in both the training and the test folds. These models are described in detail below.

Logistic Regression: Logistic regression is a binary classification model which creates a linear weighted combination of the input features and then compresses the result into a probability between 0 and 1 using a sigmoid function. This probability is then turned into a binary output depending on the threshold value for the probability. The weights on the linear combination are determined by the solver function.

Support Vector Machine: Support vector machines are linear classifiers that aim to maximize the margin between the two classes. In the case of two dimensions, this means that the support vector machine aims to find the line which can divide the two classes with the greatest margin, and for higher dimensions the same theory is applied. For nonlinear boundaries, the kernel trick can be applied, where the data is transformed to a higher dimension (by squaring, cubing, or some other function) in which a linear boundary exists.

K-Nearest Neighbors: K-Nearest Neighbors is a classification model which calculates the Euclidean distance between each new test point and all of the training points using all of the features, and classifies the point as the category that the majority of the *k* closest points (where *k* is specified by the programmer) are in.

Decision Tree: The decision tree is a relatively simple classification model. It begins with all of the data at the root node. The data is then divided into binary groups (nodes) based on the feature that best separates the data into the classes. This partitioning continues until either all of the data is classified into nodes that contain only one category or the specified maximum depth is reached. The tree can then be used to classify new data based on the features used in the tree.

Random Forest: Random forest is a decision tree-based model that employs the ensemble technique bagging. In random forest, a subset of the data is selected, and a decision tree is created from that subset. This process is then repeated (with resampling, so each time any data point could be selected) for the desired number of trees. When predicting, a data point is classified independently by each tree and the final classification is that of the majority of the trees.

Gradient Boosting: Gradient boosting is another decision tree-based model, but employs the ensemble technique boosting rather than bagging. In gradient boosting, a decision tree is first fit to the data. Then, misclassified observations are assigned more weight, and another tree is fit. This process repeats for the desired number of iterations.

XGBoost: XGBoost is an optimized implementation of gradient boosting which is more flexible and computationally efficient than regular gradient boosting.

Multi-Layer Perceptron: The multi-layer perceptron is a type of neural network, specifically a feedforward artificial neural network. A neural network consists of an input layer (the input features), the output layer (the classification), and hidden layers. Hidden layers are each connected to their respective previous layer and next layer by weights, which are determined by nonlinear functions that are optimized to best predict the classifications. Multi-layer perceptrons are feedforward neural networks, meaning that the weights are adjusted as more training data is provided to the model.

The f1 score, accuracy, precision, recall, and AUC-ROC score were used to determine the quality of the cross-validated models. These metrics are all commonly used in evaluating machine learning models. The accuracy is the total proportion of data points that were classified correctly. For an imbalanced dataset such as this one, the accuracy can be deceptively high, for example if all data points are classified as negative. The precision is the proportion of data points that were classified as positive that were actually positive, and the recall is the proportion of actual positives that were classified as positive. Precision and recall are more informative for this study. In the cross-validated grid search with GridSearchCV that determined the best hyperparameters, recall was optimized because for medical diagnosis, recall is much more informative than the other metrics included, as false positives are preferred than false negatives. The f1 score is a weighted harmonic mean of precision and recall which gives precision and recall equal weights, and the AUC-ROC score is the area under the receiver operating characteristic curve, which is a measure of the ability of the model to separate the classes. An AUC-ROC of 0.5 indicates no separability, an AUC-ROC of 1.0 indicates perfect separability, and an AUC-ROC of 0 indicates that the model separates the classes in the opposite direction. After obtaining these metrics, we also recorded the feature importances for the decision tree, gradient boosting, XGBoost and random forest models.

Results and Discussion

Of the several machine learning models we used, the gradient boosting, random forest, and support vector machine algorithms performed best in recall (71.4%). XGBoost performed best in accuracy (80.9%), and gradient boosting performed best in f1 score (48.6%), precision (44.0%), and AUC-ROC score (0.864). Thus, gradient boosting was overall the best machine learning model (Table 1). The hyperparameters for all of our models are listed in Table 2.

Table 1 – Model evaluation metrics for cross-validation of machine learning models predicting gestational diabetes on scaled and upsampled clinical patient data (best scores in test set bolded)

	Logistic Regression	Support Vector Machine	K-Nearest Neighbors	Decision Tree	Random Forest	Gradient Boosting	XGBoost	Multi-Layer Perceptron
Test F1 Score	0.302 ± 0.217	0.378 ± 0.086	0.314 ± 0.148	0.343 ± 0.256	0.386 ± 0.202	**0.486 ± 0.239**	0.460 ± 0.224	0.179 ± 0.238
Test Accuracy	0.770 ± 0.078	0.699 ± 0.108	0.669 ± 0.078	0.750 ± 0.114	0.690 ± 0.156	0.800 ± 0.132	**0.809 ± 0.123**	0.731 ± 0.102
Test Precision	0.231 ± 0.179	0.279 ± 0.107	0.219 ± 0.109	0.326 ± 0.346	0.286 ± 0.173	**0.440 ± 0.307**	0.439 ± 0.307	0.119 ± 0.159
Test Recall	0.500 ± 0.408	**0.714 ± 0.267**	0.643 ± 0.378	0.571 ± 0.450	**0.714 ± 0.393**	**0.714 ± 0.393**	0.643 ± 0.378	0.357 ± 0.476
Test AUC-ROC score	0.753 ± 0.173	0.811 ± 0.161	0.670 ± 0.129	0.726 ± 0.146	0.685 ± 0.219	**0.864 ± 0.096**	0.862 ± 0.044	0.765 ± 0.184
Train F1 Score	0.541 ± 0.037	0.407 ± 0.044	1.000 ± 0.000	0.555 ± 0.064	0.384 ± 0.032	0.637 ± 0.034	0.639 ± 0.060	0.773 ± 0.065
Train Accuracy	0.822 ± 0.022	0.723 ± 0.028	1.000 ± 0.000	0.812 ± 0.046	0.700 ± 0.047	0.857 ± 0.024	0.872 ± 0.025	0.922 ± 0.027
Train Precision	0.408 ± 0.033	0.282 ± 0.031	1.000 ± 0.000	0.409 ± 0.063	0.264 ± 0.027	0.478 ± 0.036	0.506 ± 0.052	0.634 ± 0.089
Train Recall	0.807 ± 0.064	0.729 ± 0.094	1.000 ± 0.000	0.896 ± 0.152	0.719 ± 0.131	0.961 ± 0.049	0.871 ± 0.089	1.000 ± 0.000
Train AUC-ROC score	0.899 ± 0.021	0.784 ± 0.025	1.000 ± 0.000	0.880 ± 0.058	0.734 ± 0.038	0.947 ± 0.008	0.944 ± 0.014	0.987 ± 0.005

Ye et al. had also found that gradient boosting was the best machine learning model, as measured by AUC-ROC score (0.709), but their logistic regression model was their best predictor (0.7351). However, we found that our gradient boosting model (0.864) outperformed both our logistic regression model (0.753) and Ye et al.'s logistic regression model in AUC-ROC score. Our gradient boosting model also slightly outperformed Artzi et al.'s gradient boosting model (0.85 AUC-ROC score), created based on electronic health records.

A graph of the feature importances for each of the clinical features in the decision tree, gradient boosting, XGBoost and random forest models is shown in Figure 1. Central armellini fat (VAT) measurement is the most important feature for the decision tree, gradient boosting and random forest models. Pregestational body mass index is the most important feature for the XGBoost model, with VAT a close second. As VAT was a feature that had previously not, to our

knowledge, been used in predictive machine learning models for GDM, it is notable that VAT was the most important feature for the decision-tree related models overall. Our inclusion of VAT may have been the reason that our gradient boosted trees outperformed Ye et al.'s logistic regression and gradient boosted trees and Artzi et al.'s gradient boosted trees (Artzi et al., 2020; Ye et al., 2020).

Table 2 – Hyperparameters selected for machine learning models by GridSearchCV

Machine Learning Model	Hyperparameters Selected by GridSearchCV
Logistic Regression	penalty = l2, solver = newton-cg, max_iter = 50, class_weight = None
Support Vector Machine	kernel = sigmoid
K-Nearest Neighbors	algorithm = auto, n_neighbors = 7, weights = distance
Decision Tree	max_depth = 2, n_estimators = 5, splitter = random
Random Forest	max_depth = 2, n_estimators = 50, criterion = gini, max_features = auto
Gradient Boosting	max_depth = 2, n_estimators = 5, criterion = mae
XGBoost	max_depth = 2, subsample = 0.25, n_estimators = 50, colsample_bytree = 0.25, learning_rate = 0.05
Multi-Layer Perceptron	hidden_layer_sizes = (20,) , max_iter = 500

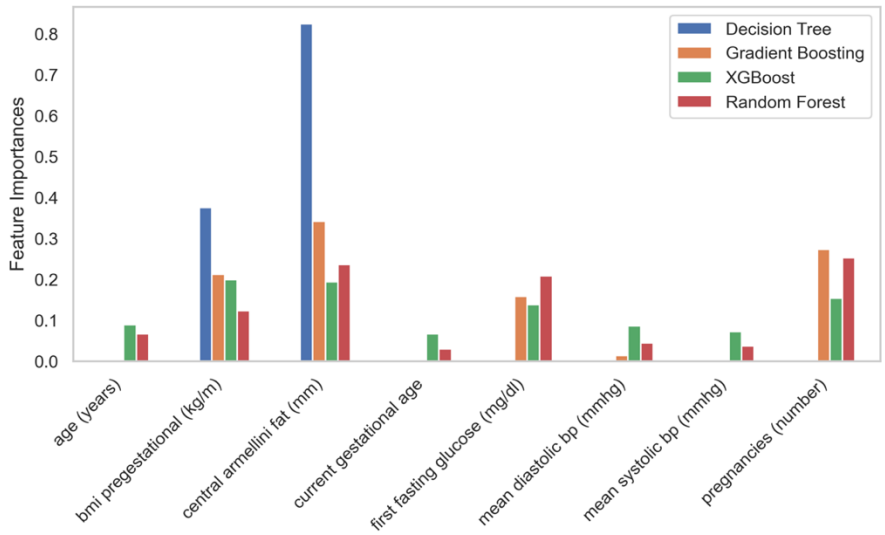

Figure 1 – Feature importances for decision-tree based models (decision tree, gradient boosting, XGBoost, and random forest).

Shortcomings of this study included the limited number of patients in total, particularly GDM patients, and that all of the patients in this study came from a single institution in Brazil. Further validation on a larger amount of data and data from other institutions is necessary to eliminate any bias present in our dataset. Also, the clinical features used in this study were some of the most common risk factors for GDM, but other features such as diet, socioeconomic status, environmental factors, and lifestyle choices are also risk factors for GDM and may add value to future machine learning predictive models (Carroll et al., 2018).

Additionally, data from patients in this study included gestational ages from 6 weeks to 32 weeks. This variation in gestational age likely affected the predictive capabilities of visceral

adipose tissue measurements and first fasting glucose levels. Visceral adipose tissue measurements increase during pregnancy for all women, not just those with GDM (Gunderson et al., 2008), and fasting glucose levels are also affected by the gestational age, dropping during the first trimester, remaining constant in the second trimester, and again dropping in the third trimester (Angueira et al., 2015). While our models did use current gestational age as a feature, future work on a larger dataset might use it more explicitly, perhaps by developing separate models for use at different gestational ages.

References

Angueira, A. R., Ludvik, A. E., Reddy, T. E., Wicksteed, B., Lowe, W. L., & Layden, B. T. (2015). New insights into gestational glucose metabolism: Lessons learned from 21st century approaches. *Diabetes*, *64*(2), 327–334. https://doi.org/10.2337/db14-0877

Artzi, N. S., Shilo, S., Hadar, E., Rossman, H., Barbash-Hazan, S., Ben-Haroush, A., Balicer, R. D., Feldman, B., Wiznitzer, A., & Segal, E. (2020). Prediction of gestational diabetes based on nationwide electronic health records. *Nature Medicine*, *26*(1), 71–76. https://doi.org/10.1038/s41591-019-0724-8

Buchanan, T. a, & Xiang, A. H. (2005). Gestational diabetes mellitus. *The Journal of Clinical Investigation*, *115*(3), 485–491. https://doi.org/10.1172/JCI200524531.The

Carroll, X., Liang, X., Zhang, W., Zhang, W., Liu, G., Turner, N., & Leeper-Woodford, S. (2018). Socioeconomic, environmental and lifestyle factors associated with gestational diabetes mellitus: A matched case-control study in Beijing, China. *Scientific Reports*, *8*(1), 1–10. https://doi.org/10.1038/s41598-018-26412-6

Corrado, F., D'Anna, R., Cannata, M. L., Interdonato, M. L., Pintaudi, B., & Di Benedetto, A. (2012). Correspondence between first-trimester fasting glycaemia, and oral glucose tolerance test in gestational diabetes diagnosis. *Diabetes and Metabolism*, *38*(5), 458–461. https://doi.org/10.1016/j.diabet.2012.03.006

Dabelea, D., Snell-Bergeon, J. K., Hartsfield, C. L., Bischoff, K. J., Hamman, R. F., & McDuffie, R. S. (2005). Increasing prevalence of gestational diabetes mellitus (GDM) over time and by birth cohort: Kaiser Permanente of Colorado GDM screening program. *Diabetes Care*, *28*(3), 579–584. https://doi.org/10.2337/diacare.28.3.579

Dominguez, L. J., Martínez-González, M. A., Basterra-Gortari, F. J., Gea, A., Barbagallo, M., & Bes-Rastrollo, M. (2014). Fast food consumption and gestational diabetes incidence in the SUN project. *PLoS ONE*, *9*(9), 1–7. https://doi.org/10.1371/journal.pone.0106627

Goldberger, A., Amaral, L., Glass, L., Hausdorff, J., Ivanov, P. C., Mark, R., Mietus, J. E., Moody, G. B., Peng, C. K., & Stanley, H. E. (2000). *PhysioBank, PhysioToolkit, and PhysioNet: Components of a new research resource for complex physiologic signals. 101*(23), e215–e220.

Gunderson, E. P., Sternfeld, B., Wellons, M. F., Whitmer, R. A., Chiang, V., Quesenberry Jr, C. P., Lewis, C. E., & Sidney, S. (2008). Childbearing may increase visceral adipose tissue independent of overall increase in body fat. *Obesity*, *16*(5), 1078–1084. https://doi.org/10.1038/oby.2008.40

Hedderson, M. M., & Ferrara, A. (2008). High blood pressure before and during early pregnancy is associated with an increased risk of gestational diabetes mellitus. *Diabetes Care*, *31*(12), 2362–2367. https://doi.org/10.2337/dc08-1193

Martin, A. M., Berger, H., Nisenbaum, R., Lausman, A. Y., MacGarvie, S., Crerar, C., & Ray, J. G. (2009). Abdominal visceral adiposity in the first trimester predicts glucose intolerance in later pregnancy. *Diabetes Care*, *32*(7), 1308–1310. https://doi.org/10.2337/dc09-0290

[dataset] Rocha, A. d. S., von Diemen, L., Kretzer, D., Matos, S., Rombaldi Bernardi, J., & Magalhães, J. A. (2020). *Visceral adipose tissue measurements during pregnancy* (1.0.0). PhysioNet. https://doi.org/10.13026/p729-7p53

Sukumaran, S., Madhuvrata, P., Bustani, R., Song, S., & Farrell, T. A. (2014). Screening, diagnosis and management of gestational diabetes mellitus: A national survey. *Obstetric Medicine*, *7*(3), 111–115. https://doi.org/10.1177/1753495X14536891

Xiong, X., Saunders, L. D., Wang, F. L., & Demianczuk, N. N. (2001). Gestational diabetes mellitus: Prevalence, risk factors, maternal and infant outcomes. *International Journal of Gynecology and Obstetrics*, *75*(3), 221–228. https://doi.org/10.1016/S0020-7292(01)00496-9

Ye, Y., Xiong, Y., Zhou, Q., Wu, J., Li, X., & Xiao, X. (2020). Comparison of Machine Learning Methods and Conventional Logistic Regressions for Predicting Gestational Diabetes Using Routine Clinical Data: A Retrospective Cohort Study. *Journal of Diabetes Research*, *2020*, 1–10. https://doi.org/10.1155/2020/4168340