The global incidence of diabetes was estimated at 422 million in the year 2014 and its prevalence among the adult population has seen an ion deaths worldwide were estimated to be attributed to diabetes. In addition, compared to a non-diabetic individual, a diabetic patient is at a greater risk of developing cardiovascular diseases, visual impairment and limb amputations. Due to the substantial socio-economic burdens that are associated with diabetes, its early detection, intervention and prevention has become a worldwide top-level health concern. There is even experimental evidence that the development of diabetes can be delayed or even prevented, provided an individual undergoes a lifestyle change that includes managing diet, incorporating exercise, and adhering to a pharmacological treatment [REFs here]. Therefore, early identification of high risk individuals is essential for targeted prevention strategies.

While clinical studies diagnosing accurately the state of diabetes have been growing in numbers for the last two decades, the number of studies predicting individuals at risk of developing diabetes are limited but has seen an increased amount of research interest in the last decade [9]. However, the clinical significance of such predictions largely depend on the type and quality of data collected. There are studies collecting sociodemographic characteristics such as age, ethnicity, body mass index (BMI) and genealogical information through conducting population surveys, and then assign a probability to individuals of having diabetes [10], [11]. Moreover, such self-assessment techniques can often be misleading and cannot be relied upon. On the other hand, the outcomes of the diabetes related studies that involve physiological data such as blood samples collected in a laboratory environment provide an accurate clinical insight. Currently, the World Health Organization (WHO) and the American Diabetes Association (AMA) define the impaired glucose tolerance (IGT) as an indicator of emerging diabetes, a pre-diabetic state of hyperglycemia [needs REF]. Oral glucose tolerance test (OGTT) measures plasma glucose and insulin concentration at baseline and during 2 hours interval after standardized glucose intake [needs REF]. OGTT is recommended by the WHO to diagnose IGT [REF] as …. However, only 50% of subjects with diagnosed IGT developed diabetes within 10 years [3], [4]. In addition, long-term population studies have also shown that around 50% of diabetic patients did not exhibit IGT at any time prior to the diagnosis [5]. Furthermore, there are studies indicating that plasma glucose concentrations, measures during OGTT, correlate to the future diabetes risk [6]–[8].

In contrast to traditional scientific approaches using population based statistics, machine learning (ML) methods develop models that are trained using ample data. ML has currently been proposed as an accurate method for diabetes screening. Barakat et al use socio-demographic information, and point of care measurement from blood and urine to develop diabetes risk models [13]. This approach uses ensemble approach based upon a combination of SVM and random forest. Heikes et al [5] use logistic regression and classification tree analysis to develop a tool to calculate the probability of an individual to have either undiagnosed diabetes or prediabetes. The labels (normo-glycemic, undiagnosed diabetes, prediabetes) were deduced form point measurements at baseline, and the variables were extracted form socio-demographic information, and point of care measurement from blood: cholesterol, plasma glucose and insulin. Han et al employed SVM to develop a decision making algorithm for diagnosis of diabetes [12]. This approach uses socio-demographic information and glucose levels at baseline and 2h thereafter during an OGTT.

On top of an accurate diagnosis of diabetes, there are also attempts to predict diabetes in a healthy population. Abdul-Ghani et al [REF16] developed a model for the prediction of the diabetes risk on basis of a multivariate logistic model and 1-h plasma glucose concentration measured during OGTT. They fine-tuned the 1-h plasma glucose concentration with the best prediction potential. Furthermore, Abdul-Ghani et al [14] developed also a diabetes prediction model a regression model using fasting glucose, age, family history, cholesterol, blood pressure and waist circumference. Although the accuracy is low, the sensitivity, this model demonstrated a high specificity. Furthermore, to assess the risk of developing diabetes in healthy population, Abdul-Ghani et al [REF10] use OGTT measurements (plasma glucose and insulin based) with the labels assessed at follow-up. They used statistical testing and ROC curve to evaluate the quality of the prediction. They reached a ROC of 0.8 when combining several variables. Stern et al [REF15] identified the individuals with high risk of developing diabetes by developing a multiple logistic regression model. As variables they used medical history, age, sex; ethnicity; fasting and 2-hour glucose levels, systolic and diastolic blood pressures, total, LDL, and HDL cholesterol levels, triglyceride level, body mass index, and parental or sibling history of diabetes. The ROC varied between 77.5% when using solely 2h plasma glucose, and 85.9% when using a full model including 2h glucose.

We hypothesized that features extracted from OGTT data will be able to predict future onset of diabetes accuratelly. In this paper, we develop a diabetes prediction model by identifying the features, computed from the OGTT data, that strongly correlate to future diabetes. Furthermore, we develop a support vector machine prediction model using these features. For this purpose, we use the OGTT data generated from the population-based, epidemiological study, the San Antonio Heart Study (SAHS) [16], [17].