# 

# Abstract

### Diabetes is a fast-growing worldwide health problem affecting millions. The rise in the number of diabetics requires a working detection system that would be effective and timely. This report tackles the issue of early predictions of diabetes based on healthcare indicators. The aim is to create and compare traditional (Analytical regression; Random Forest, SVM) and deep learning models (LSTM, GRU) to diagnose correctly. We used methods like data preprocessing, SMOTE for class balancing, and several metrics of evaluation for analysis of the performance. Our research adds to the existing understanding by demonstrating the effectiveness of different models, and by showing that GRU outperforms. This work helps healthcare systems in adopting accurate diagnostic procedures.

# Introduction

Diabetes is a long-term health condition in which there are high levels of blood glucose. It is usually asymptomatic in the beginning, which makes early detection extremely necessary. Diabetes according to the World Health Organization (WHO) is one of the leading causes of death worldwide causing millions of deaths annually.

Lifestyle changes and the ageing population are responsible for increasing the incidence of diabetes. Many of the patients go undiagnosed until shocks emerge. Therefore, timely and correct prediction mechanisms are crucial. In the traditional screening methods, although effective, may not be present or affordable in the low resource settings.

This research focuses on investigating how big data analytics and machine learning/deep learning approaches can be applied to building a powerful and available diabetes detection system for increasing preventive care and early treatment.

## 1.2 Contribution of the Work Connected with Methodology

* Random Over Sampling (ROS) for balancing data to control class imbalance.
* Training and evaluation of three machine learning classifiers.
* LSTM and GRU models development and training for improved understanding of time.
* Comparative analysis of the level of performance of the model to establish the best approach.

This report is divided into sections introduction, background, related work, methodology, results, discussion and conclusions.

# Related Work

Many studies have examined applications of machine learning in the prediction of diabetes. As an example, logistic regression has been applied to it because of its interpretability, and ensemble algorithms such as Random Forests have been applied to it due to robustness and increase in accuracy. Methods of deep learning including LSTMs have been used to model temporal dependency in longitudinal health records.

What is different in this study is the comparison of both traditional ML algorithms and RNN-based DL models (LSTM and GRU) for the same preprocessed and balanced dataset. Its pipeline differs from others in that I have managed to integrate ROS and detailed preprocessing.

# Methodology

Methodology for early prediction of diabetes included data pre-processing, data balancing with RandomOverSampler, EDA, scaling features using StandardScaler and MinMaxScaler and training various models such as Logistic Regression, Random Forest, SVM and an LSTM neural network. Accuracy, confusion matrix and a classification report were used to measure the performance of the model.

## 3.2 Data Collection

## The given dataset that was used for this study is titled diabetes\_indicators.csv and was found publicly and it has several health features including BMI, levels of physical exercise and smoking status and other lifestyle and medical indicators. These features are applicable for predicting the probability of diabetes in people. The environment was used to load the dataset using the pandas library and the dataset was the most important data source for training and testing both machine learning and deep machine learning models.

## 3.3 Data Preprocessing

In data preprocessing stage, the dataset was reviewed for any null or missing values for data completeness and quality. After being verified as clean, the feature values were standardized by changing the data type to integers allowing for uniformity of the data across the given dataset. Given this problem of class imbalance, which might have adverse effects on the model performance, the RandomOverSampler (ROS) method was used. This method takes care of the imbalance in the distribution of target classes by oversampling, minority class whence the models will be able to learn better from the data.

## 3.4 Machine Learning Models

## 3.5 Deep Learning Models

## 3.6 Evaluation Metrics

# 4. Results and Discussion

# 5. Conclusion