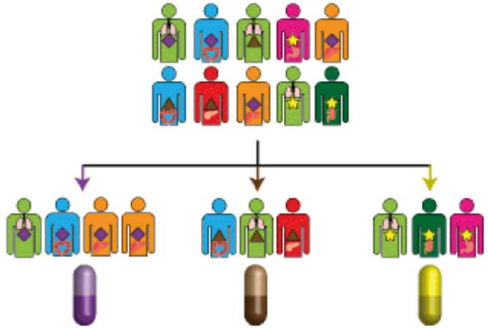
Variational Bayes methods for clinical real-world data

[Submitted on 7 Apr 2023]

Variational Bayes latent class approach for EHR-based phenotyping with large real-world dat

Brian Buckley, Adrian O'Hagan, Marie Galligan

Rationale



- The identification of specific patient groups sharing similar disease-related phenotypes is fundamental to many clinical studies

- Electronic Health Records (EHR) data has long been incorporated in phenotypic studies, which makes more complex the models to identify patient phenotypes.

Rationale

Bayesian approach: **Prior information** + **current information** at clinical trials
But!
Patient phenotyping has been **limited by the computational challenges associated with applying the Markov-Chain Monte Carlo (MCMC)** approach to large real-world data.

Computational Time

Storage Requirements

Main objective

Benchmark a VI implementation against MCMC in estimating a posterior model, predictive performance and computational performance in terms of time.

Dataset

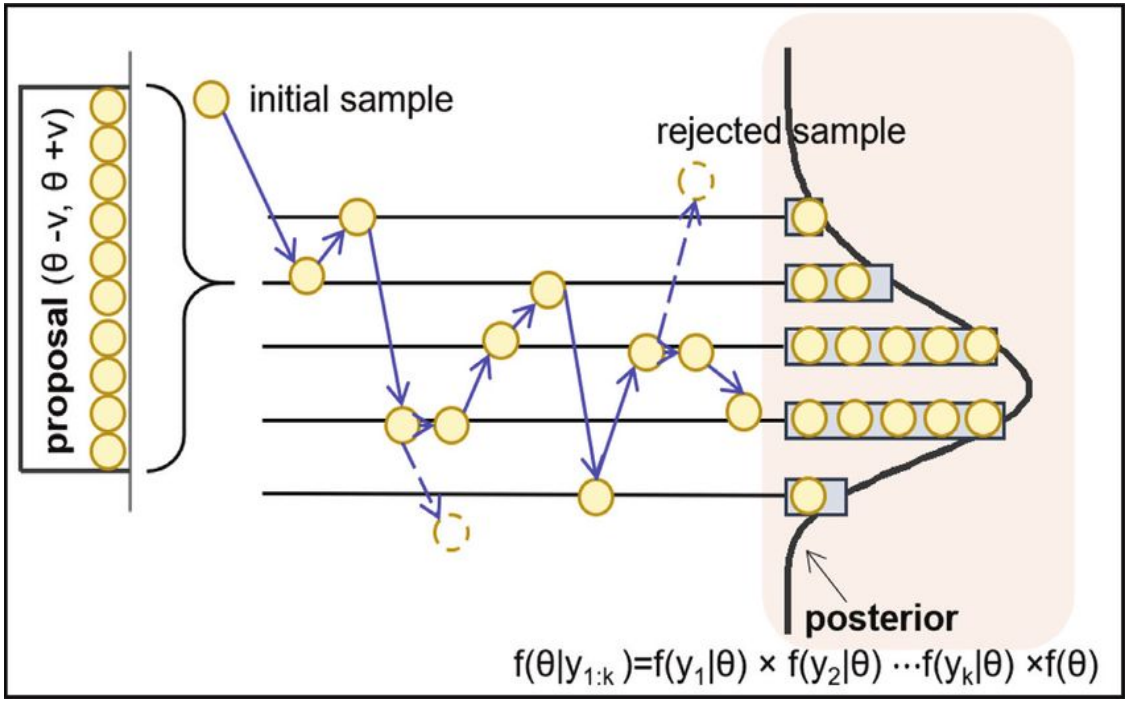
We analysed Type 2 Diabetes data. The response variable is “diagnosed type 2 diabetes”. The predictors are all continuous variables (Pregnancies, Glucose, BloodPressure, SkinThickness, Insulin, BMI, DiabetesPedigree, Age). Size: 768 rows. https://www.kaggle.com/datasets/uciml/pima-indians-diabetes-database

Predictive model

Since clinical data analyses commonly employ logistic regression, we chose this approach for comparative purposes.

MARKOV CHAIN - MONTECARLO

In order to sample from an unknown **p** distribution, using the un-normalized distribution $\tilde{\mathbf{p}}$ we create a Markov chain that has **p** as its stationary distribution and we sample from it.

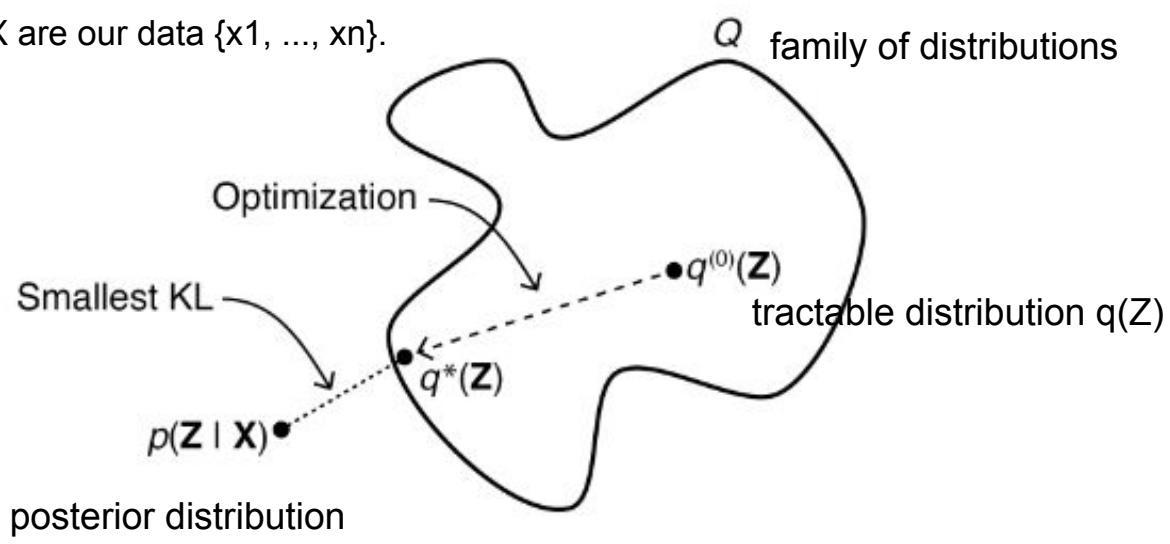


Variational Inference (VI)

Optimization process

Z are latent variables {z1, ..., zn}

X are our data {x1, ..., xn}.



This optimisation approach is often significantly more computationally efficient than MCMC in the posterior distribution estimation.

Metrics

Diagnostic Testing Accuracy: Sensitivity, Specificity, Predictive Values and Likelihood Ratios

Jacob Shreffler; Martin R. Huecker.

Author Information and Affiliations

Last Update: March 6, 2023.

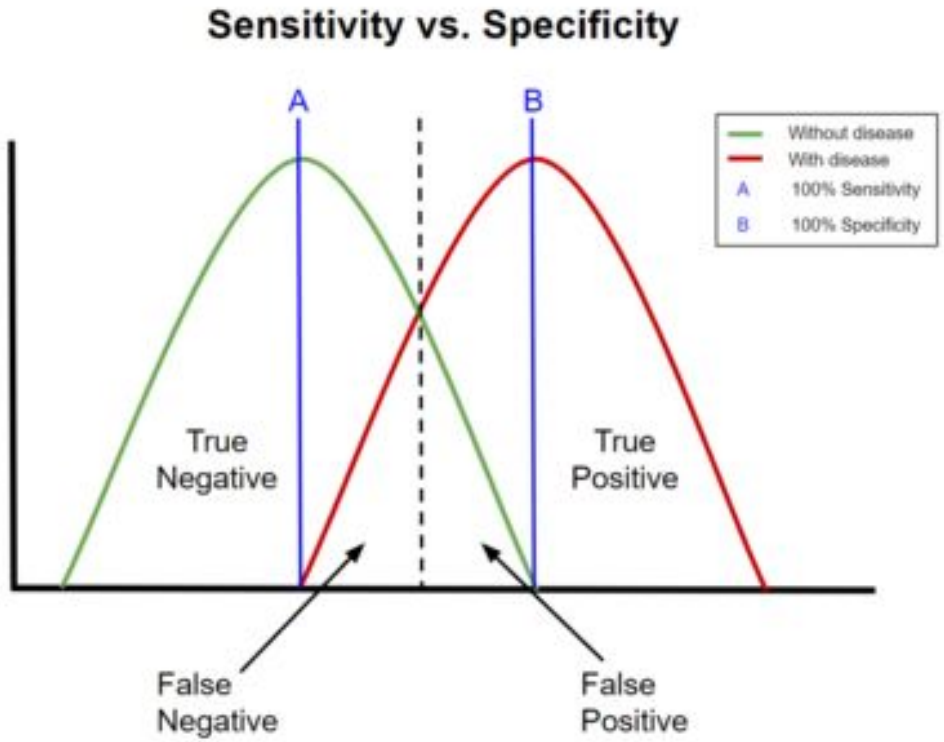
Definition/Introduction

Go to:

To make clinical decisions and guide patient care, providers must comprehend the likelihood of a patient having a disease, combining an understanding of pretest probability and diagnostic assessments.[1] Diagnostic tools are routinely utilized in healthcare settings to determine treatment methods; however, many of these tools are subject to error.

Metrics

Sensitivity = TP/(TP+FN)
Specificity = TN/(TN+FP)



Metrics

5-Folds Cross validation

MARKOV CHAIN MONTECARLO

- NUTS Kernel:

  A variation of Hamiltonian MCMC that authomatically determines the trajectory length, avoiding the need for manual hyperparameter tuning

- 2 Chains
- 200 warmup steps per chain
- 200 samples per chain

Coeficient * variable	Effective_ samples	R_hat
W1* Pregnancies	177.36	1
W2 * Glucose	600.87	1
W3 * BloodPressure	516.45	1
W4* SkinThickness	445.51	1
W5* Insulin	459.31	1
W6* BMI	303.77	1
W7* Diabetes Pedigree Function	137.09	1.01
W8 * Age	289.83	1

STOCHASTIC VARIATIONAL INFERENCE (SVI)

- SVI employs Monte Carlo estimation to estimate the gradients of the ELBO.
- SVI randomly selects a subset or mini-batch of data points to compute the gradients on the ELBO. This allows for efficient and scalable inference.
- Lastly, SVI sacrifices some accuracy compared to exact Bayesian inference. The convergence to the true posterior distribution depends on the quality of the approximation and the number of iterations in the optimization process.

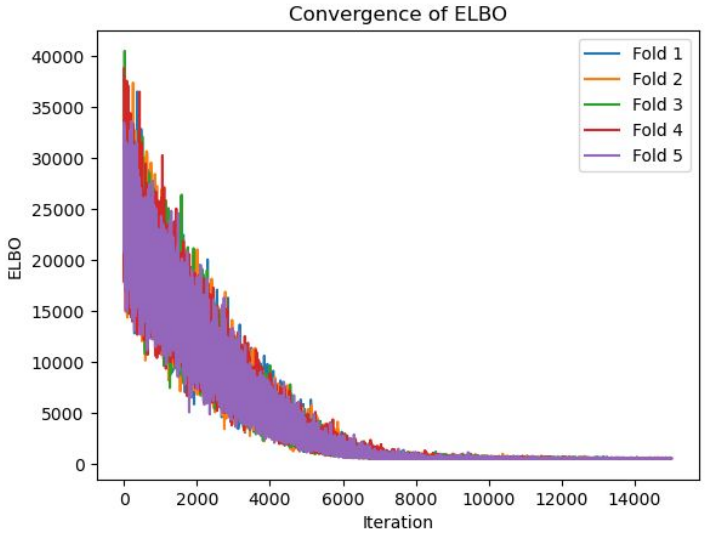
SVI implementation setting

Bayesian logistic regression model with Normal prior distributions for the model parameters. The logits are calculated based on the feature matrix X, and the observed data y is modeled using a Bernoulli likelihood. After, during SVI, a guide (variational distribution) function is used to approximate the true posterior distribution of the model parameters given the observed data.

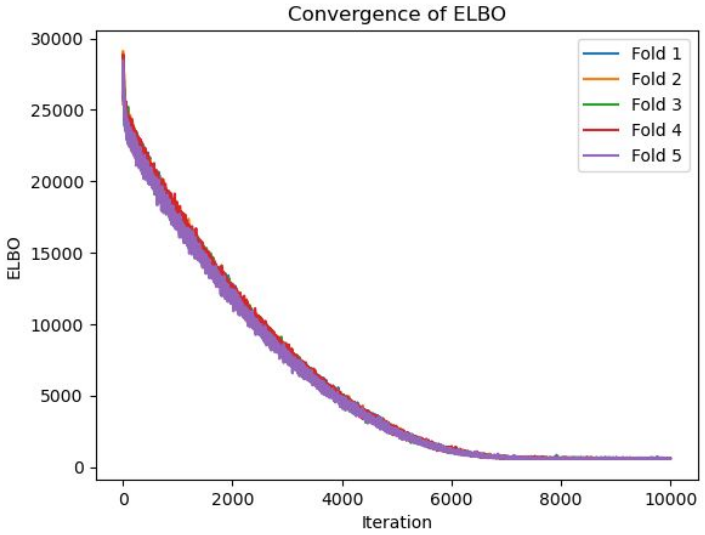
- **Optimizer: Adam** (Adaptive Moment Estimation) update the parameters (weights) during training. It is an extension of the stochastic gradient descent (SGD) optimization algorithm. Compute Gradients: In each iteration (or mini-batch), the gradients of the model's parameters with respect to the loss function are computed.
- **Learning rate: 0.01. The learning rate (ηn) controls the step size in each iteration.**
- Number of Iterations: 7000
- **Mini-batch size (determines the number of samples used for Monte Carlo estimation in the ELBO computation)**: 64

OPTIMIZATION PROCESS
Metric of convergence: ELBO

Before minibathes



After minibathes



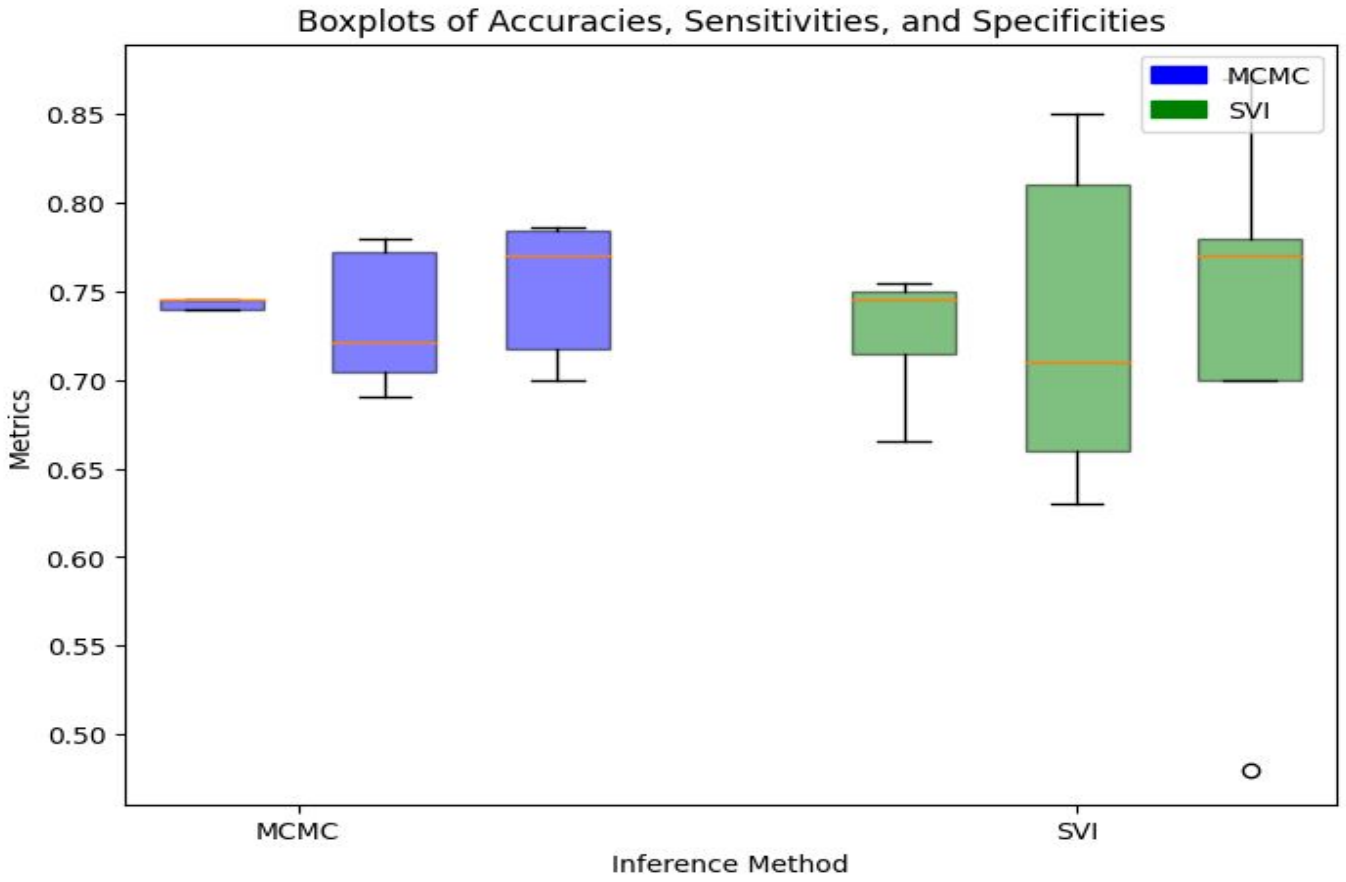
Minibatches allows to update the variational parameters more frequently, making the optimization process more efficient and reducing the variance in the parameter estimates.

COMPARISON OF RESULTS

Runtime in minutes*

MCMC: 14

SVI: 11



***Colab**

Conclusion

The Pima dataset is a valuable base to compare MCMC and VI. It is small enough for MCMC to reach convergence in a reasonable time.

We have seen that SVI needs less time than MCMC to converge and return the metric values while still having good results. However, based on the literature, we expected a more significant decrease in the runtime for SVI, we think that our SVI implementation needs for posterior optimization.

¡Many thanks!

Stanislaw Ulam (1909-1984)

S. Ulam is credited as the inventor of Monte Carlo method in 1940s, which solves mathematical problems using statistical sampling.

Andrey Markov

MARKOV CHAIN MONTECARLO

Means of weights

Coeficient * variable	Mean	Sd	Effective_ samples	R_hat
b (intersect)	-5.52	0.55	299.91	1
W1* Pregnancies	0.11	0.03	177.36	1
W2 * Glucose	0.03	0.00	600.87	1
W3 * BloodPressure	-0.03	0.01	516.45	1
W4* SkinThickness	-0.01	0.01	445.51	1
W5* Insulin	0.00	0.00	459.31	1
W6* BMI	0,08	0.02	303.77	1
W7* Diabetes Pedigree Function	0.63	0.26	137.09	1.01
W8 * Age	0.02	0.01	289.83	1

Stochastic Variational Inference

Means of weights

Coeficient * variable	Mean	Sd
b (intersect)	-6.435	1.069
W1* Pregnancies	0.124	1.014
W2 * Glucose	0.037	1.000
W3 * BloodPressure	-0.024	1.001
W4* SkinThickness	0.001	1.002
W5* Insulin	0.002	1.000
W6* BMI	0,068	1.002
W7* Diabetes Pedigree Function	0.63	1.179
W8 * Age	0.024	1.002

diagnostics

MDPI

Diagnostics (Basel). 2023 Feb; 13(4): 723.

Published online 2023 Feb 14. doi: 10.3390/diagnostics13040723

PMCID: PMC9955149

PMID: 36832207

Machine-Learning-Based Diabetes Mellitus Risk Prediction Using Multi-Layer Neural Network No-Prop Algorithm

J. Jeba Sonia,[1] Prassanna Jayachandran,[2,*] Abdul Quadir Md,[2] Senthilkumar Mohan, Resources, Visualization,[3] Arun Kumar Sivaraman, Formal analysis, Investigation, Resources,[4] and Kong Fah Tee, Methodology, Investigation[5]

Jae-Ho Han, Academic Editor

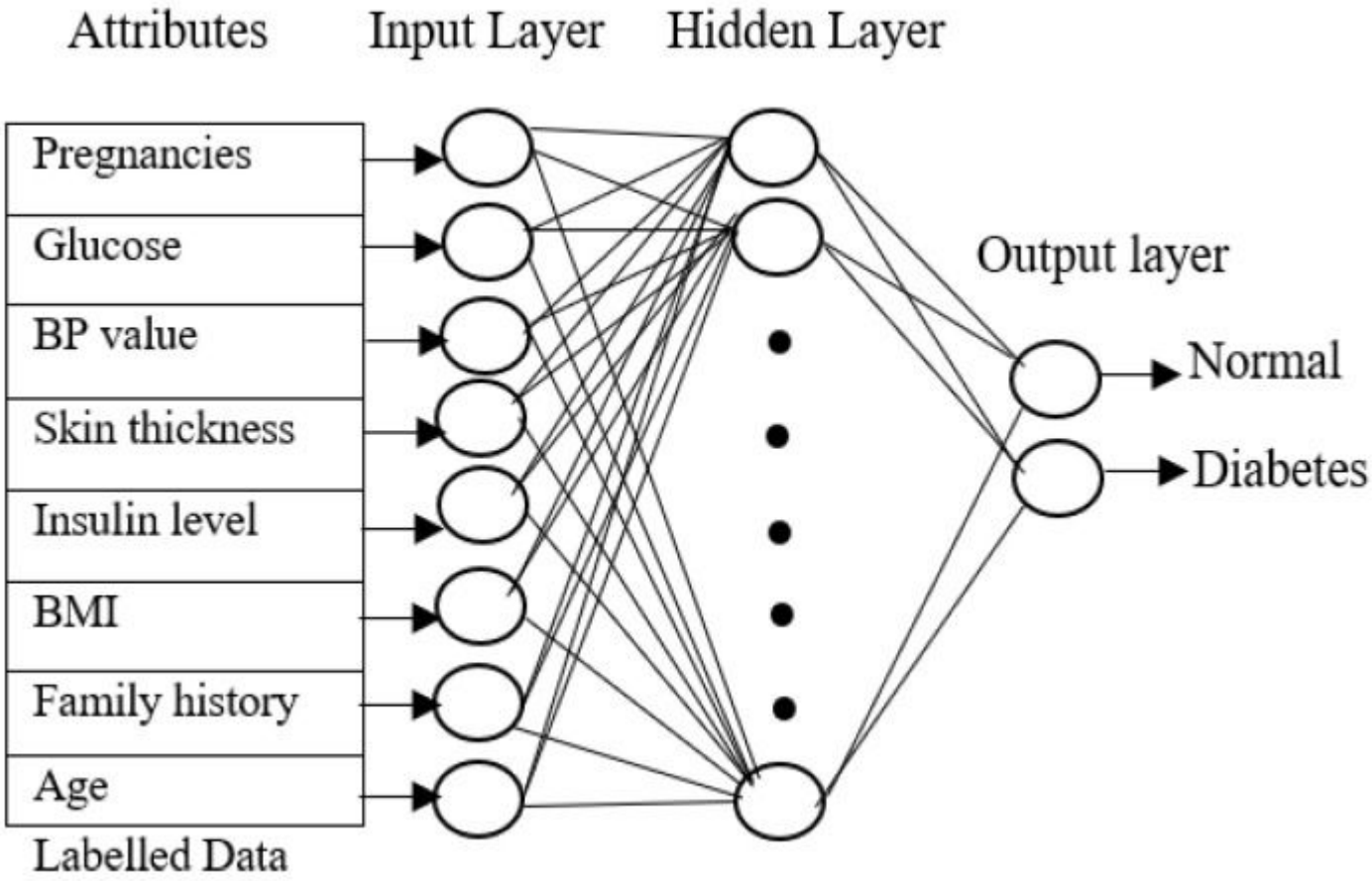

Attributes
Input Layer
Hidden Layer
Pregnancies
Glucose
BP value
Skin thickness
Insulin level
BMI
Family history
Age
Labelled Data
Output layer
Normal
Diabetes

Table 5

Classification of diabetes for dataset 2.

Parameters	Dataset 2		
	Class 2	**Class 3**	**Class 4**
Sensitivity	0.63	0.47	0.97
Specificity	0.93	0.99	0.96
Accuracy	0.89	0.97	0.98

TYPE 2 DIABETES