Patient Stratification by Drug Response using Deep Learning

J. Divyananda[1], S.McKenna[1], Prathiba [2], E. Trucco[1]

1 NIHR INSPIRED/VAMPIRE, Computer Vision and Image Processing group (SSE), University of Dundee, Dundee, UK

2 Madras Diabetes Research Foundation, Chennai, India

Introduction

NICE defines Type 2 Diabetes(T2D) as a chronic metabolic condition characterized by insulin resistance and insufficient pancreatic insulin production. It is one of the major chronic condition across the world, hence data is collected to monitor the progress of the patient over time, which can also serves various other motives like planning the healthcare system, increase the quality of patient care. Variations in drug response may be due to pharmacodynamics which implies differences in response of receptors equivalent to concentration of drugs or pharmacokinetics hints towards people receiving same dose of drug can have various concentrations in their bodily fluids . In this research we use data which has been collected over a period of time, To develop different models to address drug response prediction, we currently use GoDarts longitudinal bioresource. To study and model the data we generate synthetic data (SD), in order to develop DL models based on the same.

Results will be used to develop deep learning recurrent networks to predict response to drug and stratify patients.

Objectives

a) Longitudinal multi-step prediction

b) Predict the response of an individual to a particular treatment

d) Stratify participants by that response.

Background

T2DM conventional biomarkers are glucose-homeostasis related parameters, lipid scores, glycated haemoglobin (HbA1c) .Currently there are many models to predict progression to diabetes, pre-diabetes to T2DM etc and DL models exist to predict drug response for tumour cells, oncology, these models consider integrated genomic profiles, molecular features, phenotypes . There is a need for a model to predict the response to a drug based on the data of the particular patient.

Various Recurrent Neural Networks variations exist to deal with the particular question. Most commonly used are Long Short Term Memory(LSTM), Convolutional Recurrent Neural Network(CRNN), Bi-directional RNN's. One of the most recent discussions on drug response prediction for T2DM participants conducted, considers glycated haemoglobin, anti-diabetic drugs prescribe and also included drugs the participant was prescribed for other conditions as well were used. In this study, HbA1c is main indicator.

Synthetic Data Generation

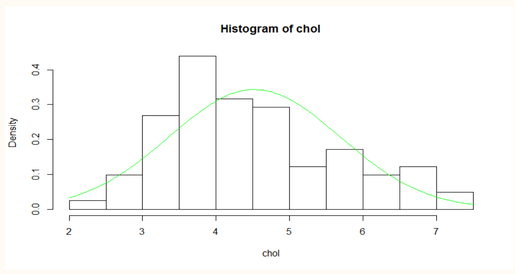
McGraw-Hill Dictionary of Scientific and Technical Terms describes synthetic data as "any production data applicable to a given situation that are not obtained by direct measurements".

In this research, we generate synthetic data (SD) for reasons being: a) to understand our data, b) To understand the advantages and limitations of the data, c) Synthetic data will enable us to work outside of safehaven and
d) SD will enable us to develop a DL model.

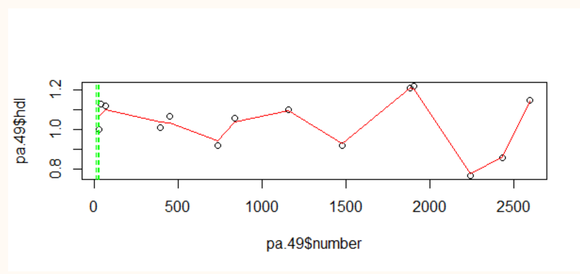
The first step is to generate synthetic cross-sectional temporal profiles building on the statistical analysis of data. The GoDarts longitudinal bioresource dataset has been utilised for the analysis. Cross-sectional data are observations made at same point of time, these observations can be of a week, month or a year . For generation of the data, begin by basic statistical analysis on the data. First, the distribution of our data, correlations between our variables. Currently for generation of SCSD currently a few features are taken into account, then move on to Principal Component Analysis (PCA) on our dataset.

Results of PCA are the main directions of variations in feature space, enabling a linear model for the data. Once equipped with the required results, we generated point distribution models for generation. The variations of our data are as shown below in Fig1 d,e. Once acquiring the variations, calculate the loading coefficients which will enable to generate a data set.

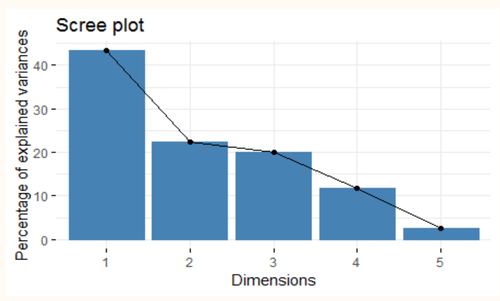
Further analysis was carried out by individual drugs, for example metformin. The cluster analysis was done on the lipids file on participants particularly on that medication.



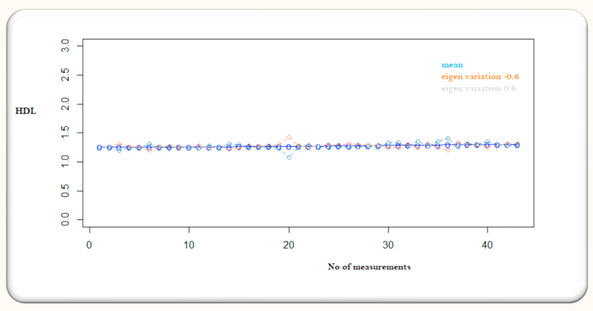
a



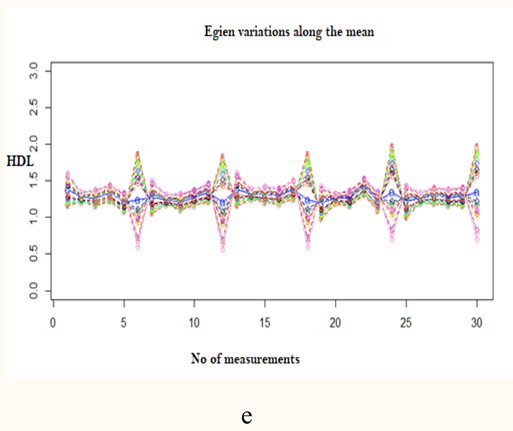
b



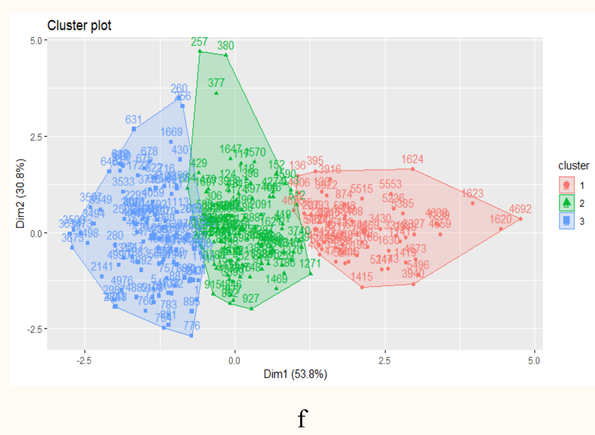
c



d



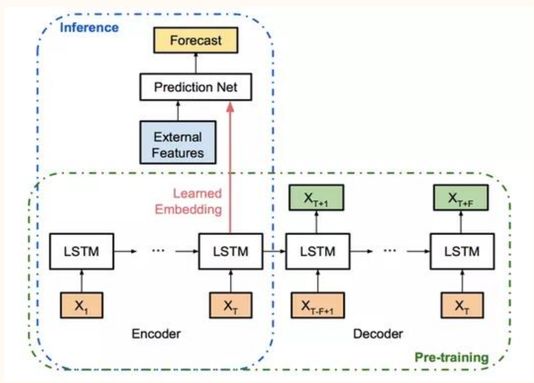
e



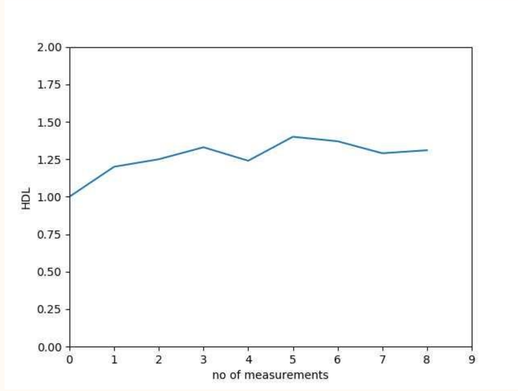
f

Fig 1 : a) Histogram of a lipid profile, b) Interpolation of data, c)Principal component analysis of data, Scree plot of the PC's, d)Eigen Variation along the mean for two values, e) Eigen Variation along the mean for all values, f) cluster analysis on lipid profiles of participants on monotherapy(Metformin)

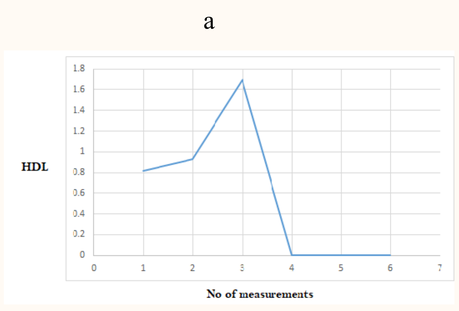
Time Series Forecasting



a



b



c

Fig 2: a) recurrent neural network with forecast (ref:http://eng.uber.com/wpcontent/uploads/2017/09/image45.png), b & c given values & predicted values based on given synthetic data

- Predict future values based on synthetic data given.
- Next step: a) Prepare data of participants with previous history.
- B) Merge the data with the lipids file & set equal number of measurements.
- C)Predict the values based on history provided.

Funding & Disclaimer

" *The research was commissioned by the National Institute for Health Research using Official Development Assistance (ODA) funding [INSPIRED 16/136/102].*

"*The views expressed are those of the author(s) and not necessarily those of the NHS, the NIHR or the Department of Health and Social Care.* "