AMRITA VISHWA VIDYAPEETHAM
DEEMED TO BE UNIVERSITY

24AIM112 - Molecular biology & basic cellular physiology
24AIM115 - Ethics, innovative research, businesses & IPR

PharmaGen AI: Personalised Drug Prediction for Type 2 Diabetes

Aayush Mohandas Nair - CB.AI.U4AIM24001
Sahana R - CB.AI.U4AIM24037
Shreya V - CB.AI.U4AIM24041
Siddharth P - CB.AI.U4AIM24043

Contents

- Introduction
- Objectives
- Computational Aspects
- Dataset
- ML model
- Testing phase
- Research paper
- Case studies
- Patents

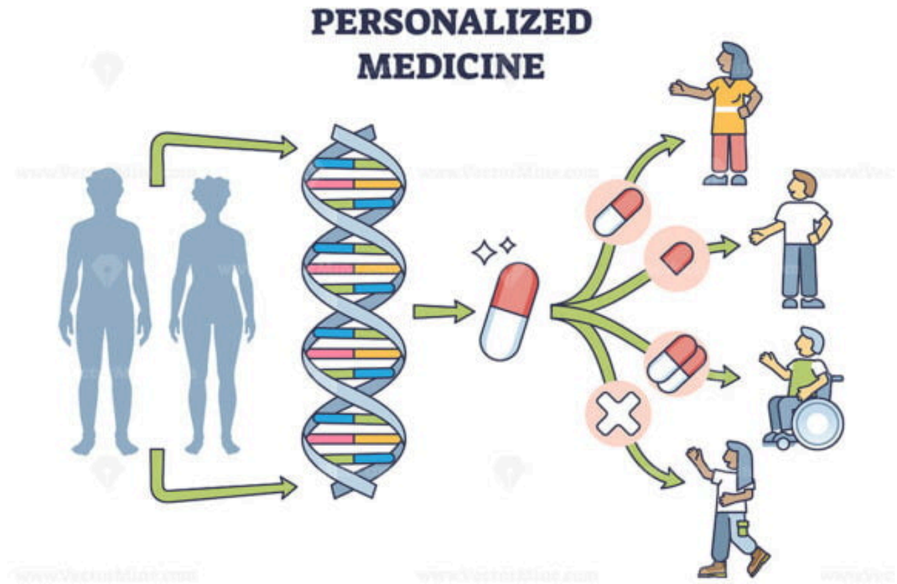
Introduction

- Type 2 Diabetes is a chronic disorder affecting millions worldwide
- 3.4 million deaths worldwide in 2024
- Patients respond differently to different drugs due to genetic variability and individual health profiles
- Need for personalised treatment methods

Objectives

1. To predict drug response: Develop an ML model (random forest) to analyze genetic profiles and clinical data to identify the most effective drug for type 2 diabetes.

2. Enhance personalized medicine: To improve treatment efficacy while minimizing adverse effects.



Computational Aspects

- Generate a synthetic dataset
- Split data into training and testing sets
- Train the decision tree model - Random forest
- Evaluate model performance - Accuracy, Confusion matrix
- Make predictions on new inputs

Dataset

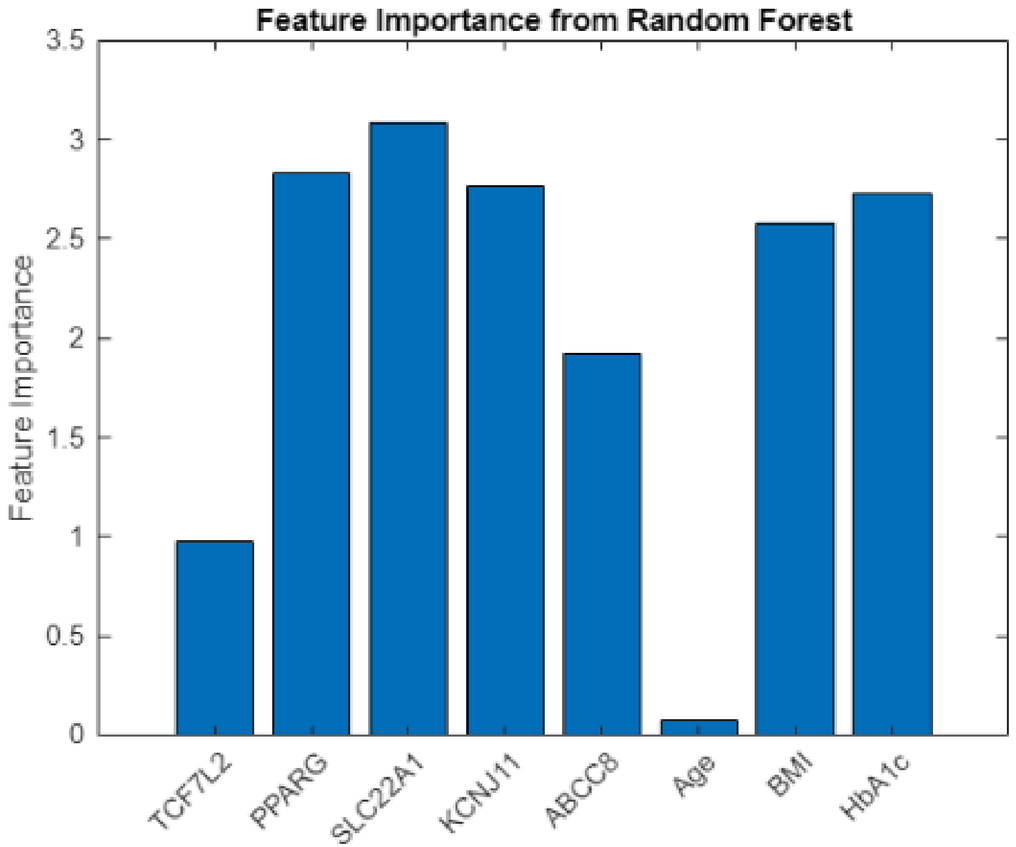
- No.of Datapoints: 500
- Features:
  a. TCF7L2
  b. PPARG
  c. SLC22A1
  d. KCNJ11
  e. ABCC8
  f. Age
  g. BMI
  h. HbA1c
- Target Variable (Drug)

TCF7L2	PPARG	SLC22A1	KCNJ11	ABCC8	Age	BMI	HbA1c	Drug
0.37	0.7	0.19	0.52	0.26	72	32.8	7.1	DPP-4 inhibitors
0.95	0.54	0.54	0.48	0.25	68	35.5	9.2	Metformin
0.73	0.31	0.87	0.03	0.91	59	23.5	10	Metformin
0.6	0.81	0.73	0.34	0.25	78	31.7	6.1	Metformin
0.16	0.68	0.81	0.38	0.27	40	30.6	9.6	Metformin
0.16	0.16	0.66	0.4	0.76	35	36.3	8.5	Thiazolidinediones
0.06	0.91	0.69	0.58	0.45	73	37.9	9.9	Metformin
0.87	0.82	0.85	0.53	0.78	52	18.3	7.9	Meglitinides
0.6	0.95	0.25	0.61	0.07	73	32.8	9.4	DPP-4 inhibitors
0.71	0.73	0.49	0.76	0.49	75	19.1	6.2	Sulfonylureas
0.02	0.61	0.22	0.81	0.03	33	30.1	7.3	Sulfonylureas
0.97	0.42	0.99	0.72	0.06	75	24.3	6.3	Sulfonylureas
0.83	0.93	0.94	0.96	0.91	52	24.7	9.9	Metformin
0.21	0.87	0.04	0.02	0.14	56	25.8	9.3	DPP-4 inhibitors
0.18	0.05	0.71	0.2	0.53	61	31.7	9.4	Thiazolidinediones

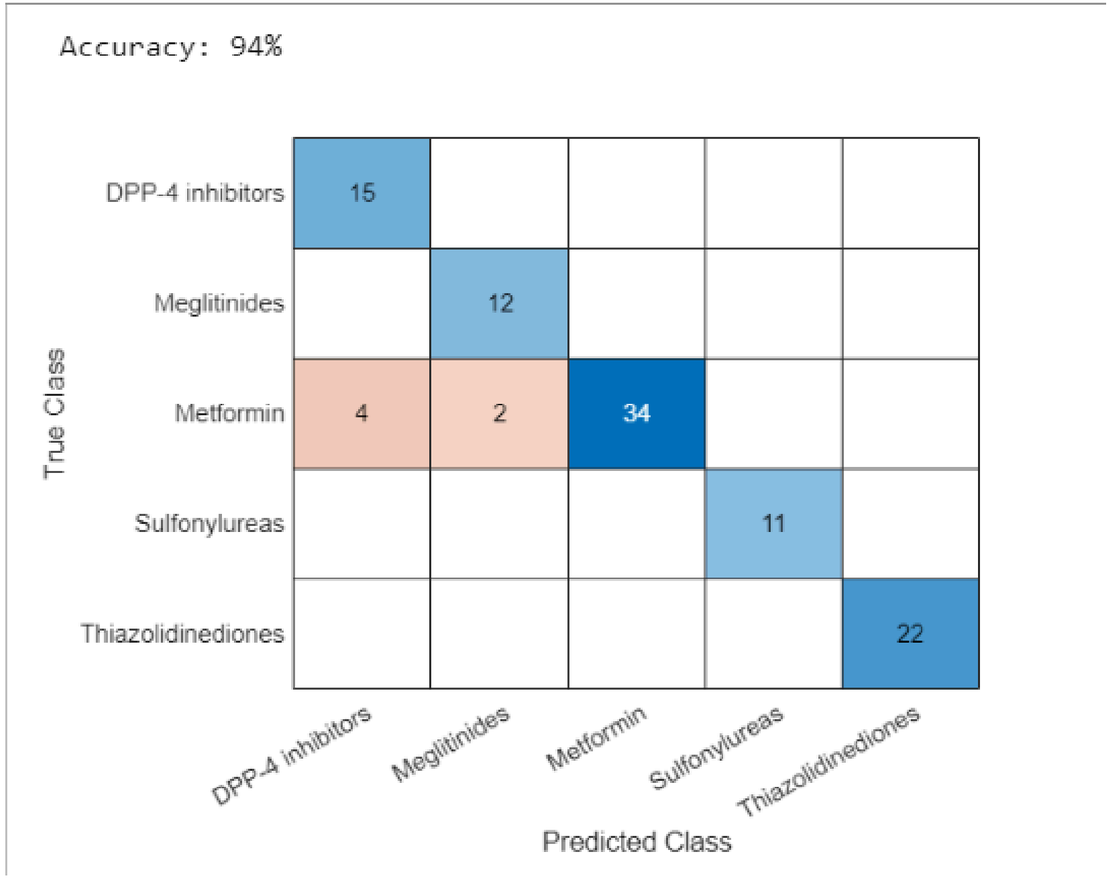
Drug assignment rules

Gene	Condition	Drug
PPARG	< 0.4 and BMI > 30	Thiazolidinediones
KCNJ11	> 0.6 and HbA1c < 8.0	Sulfonylureas
TCF7L2	> 0.6 and HbA1c > 8.0	Metformin
SLC22A1	< 0.3	DPP-4 inhibitors
ABCC8	> 0.6	Meglitinides
Fallback rule	no matches found	Metformin

RANDOM FOREST CLASSIFIER



RESULTS



Model Accuracy - 94 %

TESTING PHASE

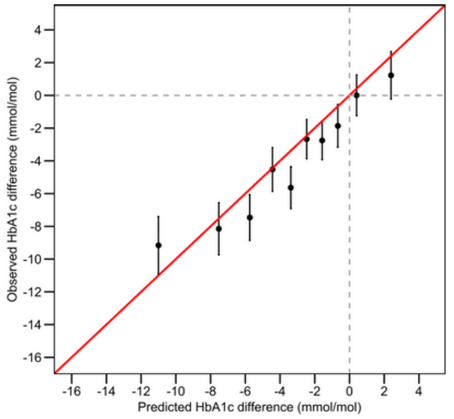
```
TCF7L2 variant score: 0.8
PPARG variant score: 0.4
KCNJ11 variant score: 0.6
SLC30A8 variant score: 0.7
ABCC8 variant score: 0.3
Age: 60
BMI: 30.2
HbA1c level: 8.1
```

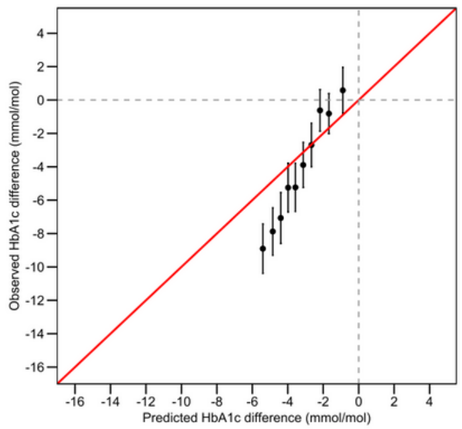
```
Predicted drug: Metformin
```

Research Paper 1 :

Comparison of causal forest and regression-based approaches to evaluate treatment effect heterogeneity:

An application for type 2 diabetes precision medicine (2023)





=> Compares two different methods
Penalised Regression | Causal Forest

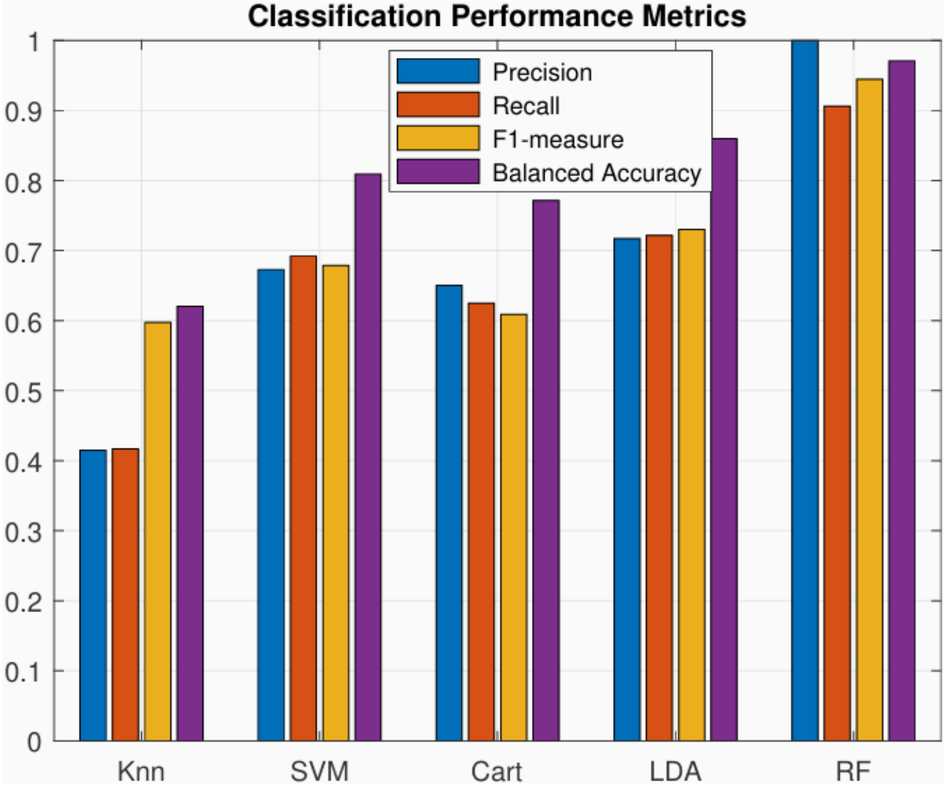
- To evaluate how individual patients with Type-2 diabetes respond to 2 different treatments.
  1. SGLT2 inhibitors
  2. DPP4 inhibitors

  [Treatment effect Heterogeneity].
- Datasets used:
  1. Clinical Trial Data (CANTATA-D) – 1428 patients
  2. Real-World Patient Data (UK, CPRD) – 18,741 Patients

Penalised Regression	Causal Forest
• Statistical method	• Decision tree-based ML model
• Factors - age, weight, kidney function, cholesterol	• Factors - age, sex, Baseline HbA1c, BMI, Blood markers, Diabetes duration.
• Advantage - Prevents model overfitting.	• Advantage - Automatically detects complex patterns.
• Disadvantage - - Manual selection of variables - Only linear relations.	• Disadvantage - - Less accuracy in predictions with real world data - class imbalance.

Research Paper 2 :

Personalized drug-response prediction model for lung cancer patients using machine learning .



- New framework → Personalized drug response prediction model for lung cancer patients using machine learning.
- Parameters used by the authors to develop the prediction model.
  1. clinical data (age, sex, smoking history etc).
  2. Geometrical properties of drug binding site.
  3. Binding free energy (how strongly the drug binds to the protein)
- Methodology
  1. computational Modeling of Mutant structures.
  2. Molecular Dynamics (MD) simulations → To study protein-drug interaction over time.
     i. Root mean square deviation (RMSD)
     ii. Binding free energy.
  3. Feature extraction and classification.
     - Geometrical features, matching rate, hydrogen bond, Euclidean distance
     - Personal features, energy features.

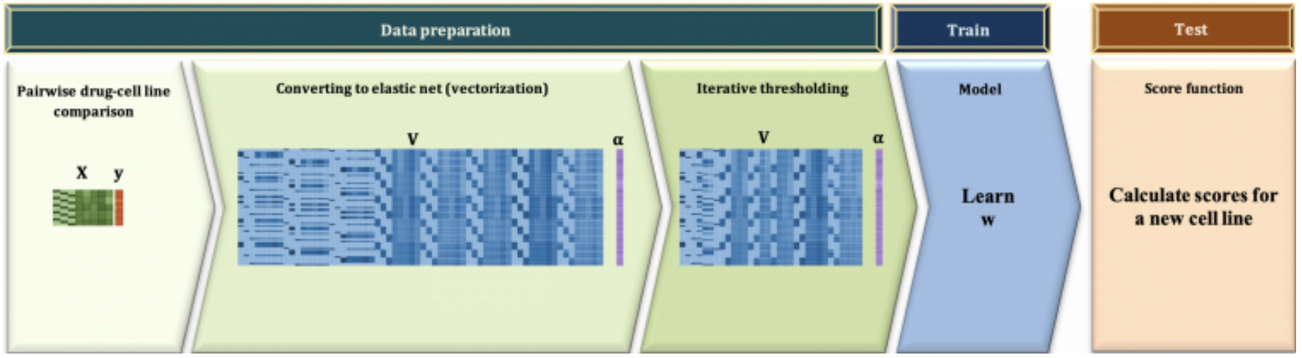
- Machine learning classification – 5 classifiers used.
  → Random forest, support vector machine (SVM), K-Nearest Neighbour (KNN), Naive Bayes, logistic regression.
- The model predicts four classes of drug response, complete response, partial response, stable disease, progressive disease.
- Results.
  Random forest classifier performs the best
  - 97.5% accuracy.
  - 100% recall
  - 95% precision
  - 96.3% F1 score.
- Conclusion
  The paper presents a remarkable model that forecasts what drug will work best for lung cancer patients using technology like machine learnin along with molecular dynamics [illegible] ometry metrics.

1280 x 862

Research Paper 3 :

They propose a novel elastic-net-based regression that utilizes gene expression features and drug sensitivity data to build a predictor

They do not predict the exact drug response for each drug, rather they develop a model that estimates the difference in sensitivity obtained by comparing two different drug-cell line pairs



→ Proposed Method: Ranking based Elastic Net Regression Model.

Main Feature / Difference : Instead of directly predicting drug responses values, it compares drug cell line pairs
- page 2 (para 7,8)

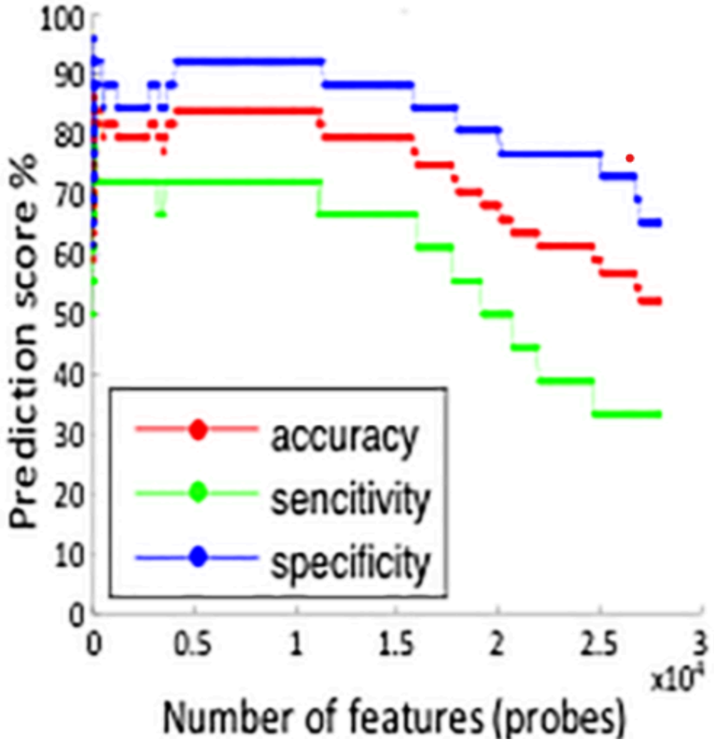
→ DATASET : They designed, evaluated, and trained their method using cell line data and drug sensitivity data from cancer cell line Encyclopedia & cancer Therapeutics Response portal
- page 3

→ PERFORMANCE METRICS - page 4. (para 1, 2·8). (para 1, 2·2).
↳ To evaluate the accuracy of their approach they used two ranking metrics namely:-
(i) AH@k ← average hit-at k.
(ii) CI^8 ← ~~Concord~~ Concordance Index
~~- page 3~~
~~(para 1, 2·2).~~

CONCLUSION
↳ Their proposed model is able to maintain its high performance even when they use high or large number of drugs and a few cell lines
- page 4
- (conclusion, point-4).

Research Paper 4 :

Open source machine-learning algorithms for the prediction of optimal cancer drug therapies



Proposed Method :

open source software platforms with Support Vector Machine (SVM) and standard Recursive feature elimination (RFE).

Focuses on predicting Response to seven drugs used for ovarian cancer. for selecting the most relevant gene.

Scoring system (by SVM) : +ve - sensitivity
-ve - resistance. [para (3,4)]

Two models built for comparison :

Model-1 : lung cancer & melanoma (trained on 18 cell lines) (78%)

Model-2 : 18 cell lines representing 9 cancer types (89%)

[para 5]

Data set used :

Gene expression data from GEO (Gene Expression Omnibus) under GSE32474

Drug sensitivity data from NCI-60 growth Inhibition Data Repository. [para 2]

Accuracy :

Prefiltered - 59.6%. Unfiltered - 89.4%.

Overall accuracy achieved > 80%.

carboplatin (75%), cisplatin (58%), peclitaxel (56%)

Conclusion :

Focuses on accuracy by Experimentation with 2 models.

Case Study-1:

Study conducted by Cheng Hu in the Shanghai Diabetes Institute to identify associations between specific gene variants and drug response

- Participants: 104 Chinese patients diagnosed with type 2 diabetes
- Treatment: Repaglinide monotherapy
- Duration: 48 weeks
- Data Analysis: Monitored changes in glycemic parameters (FPG, PPG, HbA1c) and correlated these with genetic profiles
- Identified genes: KCNJ11, ABCC8, NOS1AP, KCNQ1

Case Study-2:

- Participants: 86 patients diagnosed with type 2 diabetes
- Treatment: They took metformin for 12 months
- Analysis: Blood sugar and insulin levels were checked before taking the medicine and again after 6 months and 12 months.
- Gene : TCF7L2 rs7903146 variant (a known diabetes risk allele)
- Drug response: People with the "T" version of the gene responded better to metformin. Those with two T’s (TT) had the best response.

Patent-1:

"The present invention provides methods and systems for predicting response of a tumor to a drug using gene expression data."

Objective: To develop a system that predicts how a patient’s tumor will respond to specific drugs based on gene expression data, enabling personalized cancer treatment.

Identify gene expression features that are associated with sensitivity or resistance to one or more drugs, and use these features to predict response of a patient tumor sample to the one or more drugs."

Patent-2:

Patent US 9,195,949 B2 - Nikola Kirilov Kasabov, Auckland (NZ)
Traditional models use one global formula for all — often inaccurate for individuals.
This patent creates a custom model for each person by focusing on their closest similar cases.
It selects the most relevant features and data points to build a model tailored to that individual.
Result: More accurate, personalized predictions—especially useful in medicine (e.g., cancer treatment).
Enables better decision-making by adapting to each person’s unique profile.

Novelty

- Existing models use either clinical features or genetic profiles, this model integrates both
- It uses multiple genes as predictors (features)
- It eliminates the limitations caused by confounders
- Computationally less complex and can work efficiently with small datasets

THANK YOU