

**INFORMATICS INSTITUTE OF TECHNOLOGY**

**In Collaboration with**

**UNIVERSITY OF WESTMINSTER (UOW)**

BEng/BEng.(Hons) in Software Engineering

Final year Project 2014/2015

**Prototype Report**

For

Project Title: **Identify Inherited Diseases based on DNA (IIDDNA)**

By

Iddamalgodage Don Lahiru Manohara - 2010070

Supervised By: Mr. Achala Chathuranga Aponso

**……………………….. ….………………..**

Signature of Supervisor Signature of Student

Table of Contents

[1 Disease Prediction Application 1](#_Toc413097258)

[1.1 Application Knowledge Flow 1](#_Toc413097259)

[2 Disease Genes Prediction Application 3](#_Toc413097260)

[2.1 Process of the Application 3](#_Toc413097261)

[3 Problem Encountered 5](#_Toc413097262)

[Figure 1 Application knowledge flow 1](file:///D:\IIT\l\Final_Year_Project\Prototype_Report\PrototypeReport.docx#_Toc413097143)

[Figure 2 Prediction results 2](#_Toc413097144)

[Figure 3 Process of the genetic algorithm 3](#_Toc413097145)

[Figure 4 Fitness function code snippet 4](#_Toc413097146)

[Figure 5 Disease gene identification appliaction 4](#_Toc413097147)

# Disease Prediction Application

Disease prediction application is main feature of the IIDDNA project. The application is capable of predicting possibility of having inherited diseases on given dataset. The dataset is including disease related information. Weka dada mining tool and library was used to build the application.

The problem was considered as a Binary classification problem. The patient has inherited disease classifies as positive and native is patient has not inherited disease. In disease prediction application is considering different binary classification algorithms.

## Application Knowledge Flow

Figure Application knowledge flow

**Data Set (\*.arff, \*.csv)**

**Classification Algorithm**

**Prediction Result**

1. **Data Set**

In classification model is considering two attributes, there are geneSymbol and diseaseName. diseaseName is the class attribute, class attribute has positive and negative data related to inherited diseases. The data set prepared and preprocess according to medical perspective.

1. **Classification Algorithm**

Weka is providing many prebuilt set of algorithms. According to different prediction algorithms and compering their result SMO algorithm (Support Vector Machine) is displayed good accuracy.

1. **Prediction Result**

The prediction result is displaying in following screen shot.

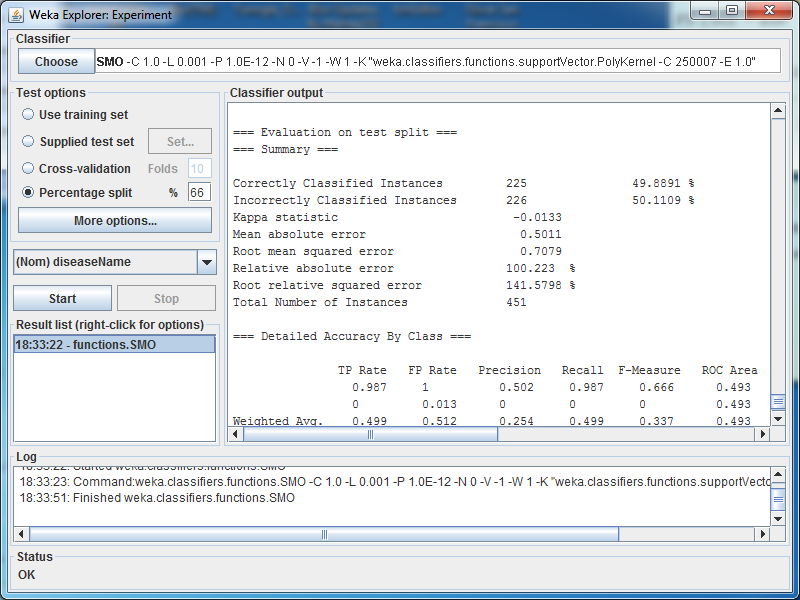


Figure Prediction results

# Disease Genes Prediction Application

Disease genes prediction application is a subprogram of the IIDDNA project. This application was build using Genetic Algorithm based software framework calls JGAP (Java Genetic Algorithm Package). Application is explaining the concept of genes association of the particular disease. In genetic inherited diseases causes many genes interaction and other conditions. The problem is difficult to find closely related set of the genes in disease.

The solution is genetic algorithm based search method to find out optimal set of genes related to particular disease.

## Process of the Application

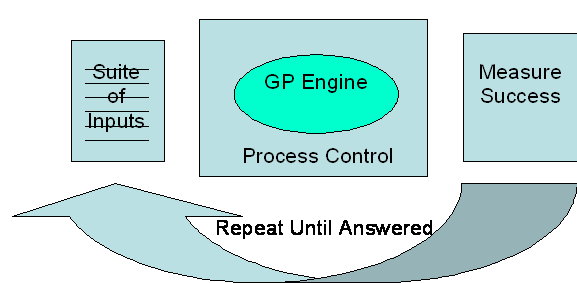


Figure Process of the genetic algorithm

1. **Suite of Inputs**

The application is getting inputs as genes and score related to particular disease. Score is displaying the how one of the gene important to the disease. The data set input from the excel file.

1. **GP Engine**

The following code is displaying the fitness function of the genetic algorithms related to disease gene identification.

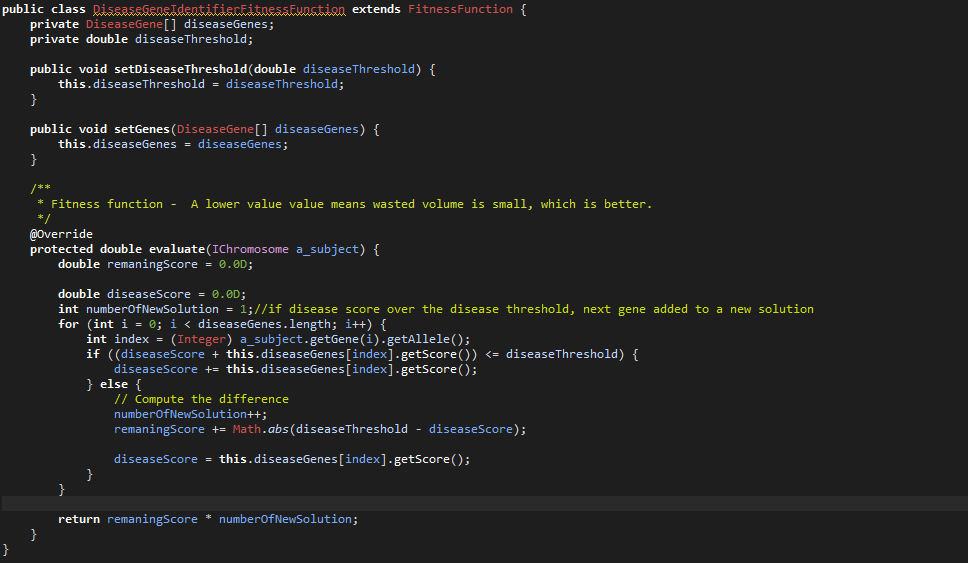


Figure Fitness function code snippet

1. **Measure Success**

Measuring success and out putting the highest fitness values included solutions. The following screen shot is displaying the final result.

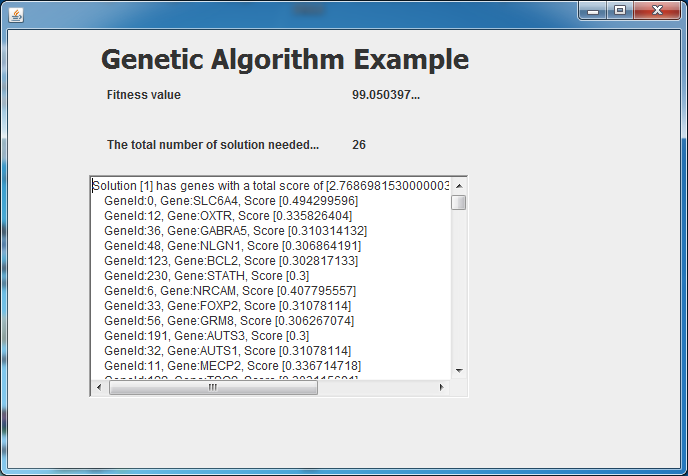


Figure 5 Disease gene identification application

# Problem Encountered

* The main problem was encountered when finding the data set for predict the disease.
* Data preprocessing and preparing was a difficult part after finding the correct data set.
* Selecting correct prediction algorithm was also difficult and some algorithms did not display the good results.
* Writing the fitness function for find out correct genes set was difficult and there was no clear documentation available for that.