INFORMATICS INSTITUTE OF TECHNOLOGY

In collaboration with

The University of Westminster, UK

**IDExplorer – Inherited Disease Prediction and Disease Genes Prioritization**

A dissertation by

Mr. I.D. Lahiru Manohara

Supervised By

Mr. Achala Chathuranga Aponso

Submitted in partial fulfillment of the requirements for the

BEng (Hons) Software Engineering degree

Department of Computing

**April 2015**

©The copyright for this project and all its associated products resides with

Informatics Institute of Technology

**Declaration**

I hereby certify that this project report and all the artifacts associated with it is my own work

And that it has not been submitted before nor is currently being submitted for any degree programme.

Full Name of Student: Iddamalgodage Don Lahiru Manohara

Registration No: 2010070

Signature: ………………………………………… Date: ……………….

Dedicating to all those who encourage me for a successful project…

**ABSTRACT**

In complexity of the genetic inherited diseases, doctors are difficult to pinpoint the correct diseases and diseases related information. Genes are playing major role of the diseases pathogenesis. There are many patterns exists on the genes related on one disease. So it is difficult to classify weather set of genes and their patterns are actually related to the particular disease. To overcome this problem, proposed solution is machine learning based prediction method to identify inherited diseases correctly. IDExplorer is supporting system for doctors and medical parsons to gain knowledge about disease and diseases related information to provide better service to the patients.

**Subject Descriptors:**

**Keywords**:

Inherited Diseases, Data Mining, Disease Gene Prioritization, Genetic Algorithm, Supervised Learning.

**ACKNOWLEDGEMENT**

**ACRONYMS AND ABBREVIATIONS**

|  |  |
| --- | --- |
| **Acronym/Abbreviation** | **Definition** |
| DNA | Deoxyribonucleic acid |
| SDK | Software Development Kit |
| SMO | Sequential minimal optimization |
| SNP | Single Nucleotide Polymorphism |
| IDE | Integrated Development Environment |
| OS | Operating System |
| PC | Personal Computer |
| HTML | Hypertext Markup Language |
| CSS | Cascading Style Sheet |
| UI | User Interface |

# INTRODUCTION

## CHAPTER OVERVIEW

This chapter intends to provide a foreword to IDExplorer project in terms of its background, problem which was attempted to solve, identified aims and objectives, activities carried out towards the completion, and terminates with an introduction on how the rest of the chapters organized into the report context.

## PROJECT BACKGROUND

“What are the chances that we will one day discover that DNA has absolutely nothing to do with inheritance? They are effectively zero.”(Harris, 2014). DNA is established in connection, and does not change throughout human’s life time. Human receive one-half of DNA from their mother and the other half from their father. Genetic mutation occurs when DNA changes, altering genetic instructions. This may result in inherited diseases. Analyzing DNA, and identify inherited diseases will show future traits of human life. Parents can be decided in their unborn children’s has a future risk of having inherited diseases.

When someone went to the doctor for some medication for disease, most probably doctor will ask from a patient regarding particular disease affected before one of their family members. This question will explain inheritance cause some of the diseases. The Table 1 shows some diseases and percentages of the inheritance involvement. According to this table inheritance percentage is high, indicates the genetic contribution is a major factor of developing particular disease. Example: - Pyloric stenosis is high probability than for peptic ulcer or congenital heart disease. “These observations have important practical implications for both research and for patient management. If the heritability is high then a search for susceptibility genes is justified.”(Nicolavich, 2010).

|  |  |
| --- | --- |
| **Diseases** | **Heritability (%)** |
| Schizophrenia | 85 |
| Asthma | 80 |
| Cleft lip / cleft palate | 76 |
| Pyloric stenosis | 75 |
| Ankylosing spondylitis | 70 |
| Club foot | 68 |
| Coronary artery diseases | 65 |
| Hypertension (essential) | 62 |
| Insulin-dependent diabetes | 60 |
| Peptic ulcer | 37 |
| Congenital heart disease | 35 |
| Insulin-nondependent diabetes | 30 |
| Hypertension (essential) | 62 |
| Insulin-dependent diabetes | 60 |

Table 1 Estimates of heritability of various diseases (Nicolavich, 2010)

Extract of meaningful information from a large experimental data set is a key element of bioinformatics. DNA pattern analysis and extract meaningful information is a technique to identify inherited diseases. Different algorithms are in data mining techniques involve finding diseases, and visualize sample data set. There are many researches are going into DNA analysis and genetic disease. GWAS (genome-wide association study) is a one of major institute support scientist to identify gene involved in human disease. “Identifying patterns of copy number variants in case control studies of human genetic disorders” research conduct to bring missing DNA data in cell, apart from them it can identify genetic disorders (Alqallaf, 2009).The different of DAN sequence are called single nucleotide polymorphisms(SNPs) . Researchers have found SNPs that may help predict an individual’s response to certain drugs, susceptibility to environmental factors such as toxins, and risk of developing particular diseases. Research like “A SNP and KEGG Based Approach to Mine Risk Pathways Associated with Bipolar Disorder” proved the SNPs are help to identify disorders (Zhang, 2008).

The Support Vector Machine (SVM) based approach uses to find out gene contributes to inherited disease. SVM is used to classify extracted features from the genes, based on gene ontology (Xie, 2012). Various data manning algorithm has been applied including linear regression, neural network and support vector regression, and create “Hierarchical learning Approach to Calibrate Allele frequencies for SNP Based Genotyping of DNA pools”. This framework explains for basic genomic to a rage of new application and studies (Hellicar, 2014). Domain-domain interaction network to detect diseases-associated nsSNPs research explains how to find genetic variation and inherited disease based on statistical based scoring method (Jiang, 2012).

Even though outcomes of research were very successful, DNA analysis and find inherited disease based researches are rare. Most of the researchers are explain the approaches to determine genetic disease. But there is no clear path explain of inherit DNA analysis and identify diseases. ”The inference of genes that are associated with human inherited diseases (diseases gene) has been a task of grate challenging in biological and medical studies” (Peng, 2012, p. 1). Making these research approaches are apply to identify inherited disease based on DNA analysis.

The purpose of “IIDDNA” project is identify inhered disease based on DNA analysis data set which applying different data mining techniques and bioinformatics algorithms to predict probability of having inherited diseases.

## PROJECT AIM

The aim of the project is to research, design, develop and evaluate a prototype outcome of the research, identify inherited diseases based on DNA analysis data. Also demonstrate the algorithms and methods which found from the research.

Further elaborate on the aim, the system will feed by standard data set and correctly classify which inherited diseases are exists. When creating classification model, machine learning methods for bioinformatics such as supervised learning and unsupervised learning algorithms will help to correctly classify the inherited diseases.

## SCOPE

First part of the project is search, and finds approaches which researchers were taken in their researches. Based on these approaches try to find out solution to identify inherited disease using visualizes, and running different algorithms on data set in different statistical analysis tools.

When successful approach found from the testing and running different algorithms on data sets, develop the prototype application with limited futures. User input sample test data with given format, the application runs the algorithm and display probability of having particular disease.

Due to the time constrains the prototype supports essential functionalities but yet sufficient to demonstrate outcome of the research. The feature like visualize data set is not supported in prototype.

For some disease heritability is directly involve but some disease are not (Nicolavich, 2010). These types of diseases not count only inheritance is major factor. So this system has only find the inherited disease such as Schizophrenia, “Schizophrenia is a chronic, severe, and disabling brain disorder that has affected people throughout history”(NIH, 2014), Asthma, “Asthma (AZ-ma) is a chronic (long-term) lung disease that inflames and narrows the airways”(NIH, 2014), Cleft lip / cleft palate, “Cleft lip and cleft palate are birth defects that occur when a baby's lip or mouth do not form properly”(NLM.NIH, 2014), pyloric stenosis, “Pyloric stenosis is a narrowing of the pylorus, the opening from the stomach into the small intestine”( NLM.NIH, 2014), ankylosing spondylitis, “Ankylosing spondylitis is a type of arthritis of the spine”(NLM.NIH, 2014).

## OBJECTIVES

1. Writing Terms of Reference report to identify the problem domain of the project and define scope, objectives and deliverables.

* **Output Artifact: Terms of Reference Document**

1. Conducting a literature survey on the following topics to gain depth knowledge and analyze past works done in those areas and estimate the future work needed to be done.
   * Different approaches for analysis DNA patterns.
   * Existing application and prototype for DNA analysis, and identify diseases.
   * Data mining techniques related to DNA analysis and classify diseases.
   * Predictions algorithms for probability of having classified disease.

* **Output Artifact: Literature Review Document**

1. Identify and analyze requirements by conduction online surveys and interviews for a prototype evaluation. This phase helps to improve prototype features.

* **Output Artifact: Software Requirement Specification (SRS)**

1. Design the application incorporating the requirements prioritized through analysis and using appropriate design methodology to guide the implementation.
2. Select appropriate development technologies and tools which are very ideal for the development of the application. Identifying the appropriate technology is significant to develop an efficient prototype.

* **Output Artifact: Design Specification**

1. Develop a prototype for demonstrating proposed features.Developing the following components is necessary to fulfill the goals of the project.

* Development prototype to demonstrate a solution found from the research.
* **Output Artifact: Application to demonstrate a solution found from the research.**

1. Test the prototype by unit testing and also by a black box testing approach to ensure all the requirements gathered in the requirement gathering phase are incorporated and functional.
2. Evaluate the prototype by allowing potential users to use the prototype and obtain feedback to ensure the outcome is good to be usable and user-friendly.
   1. Enhance prototype according to user feedbacks.
3. Submit all the project deliverables with in the dead line.

## FEATURE OF PROTTYPE

* Plugin data source or data set

The application is able to plug in different data sources with pre-defined data format to analysis, and produce the appropriate result.

* Implemented algorithms found from the research

The best approach which found from the research, implement the application with appropriate libraries and custom logic to develop the business logic of the application.

* User Interface for display result in understanding manner

Eliminate complexity in the process of analyzing data, change threshold in algorithm and easily understand the test result etc. Design simple user interface to understand the process of the analysis and plugin data sources. User will be able to understand process of the application without much hassle.

**Figure 1 High level view of the IDExplorer**

The following illustration shows the usage of different features of the Identify inherited diseases based on DNA system.

Gene Data source

Feature Extraction

Classifier

Prediction

**Figure 2: Application Flow of the IDExplorer**

## RESOURCE REQUIREMENT

### SOFTWARE REQUIREMENTS

* MATLAB or Octave

These are high level languages for prototyping purpose and data visualization. The MATLAB has tool called “BIONIFOMATIC”, it gives more features to analysis gene sequencing and prototype bioinformatics related researches.

* R software

Statistical data analysis software for calculate statistic of the data sources.

Ex: - Calculate mean distribution of the given data set.

### HARDWARE REQUIREMENTS

* Dual Core 2.0 GHz CPU
* 2GB DD2 Ram
* 5GB Hard Disk Free Space
* Intel HD Graphics Card

Note – Above requirements are subject to change

## CHAPTER WALKTHROUGH

## CHAPTER SUMMARY

# LITERATURE SURVEY

## CHAPTER OVERVIEW

The previous chapter discussed the background of the problem domain on the inherited disease identification system and overview of the project. This chapter attempts to document the findings of the literature review conducted on the problem domain. The literature review carried out on existing problem of the already existing system, invented methodologies to tackle the problem, available tools and techniques and similar solutions with advantages and limitations of existing systems.

## FOREWORD

## CHAPTER SUMMARY

# REQUIREMENTS ANALYSIS

## CHAPTER OVERVIEW

The previous chapter was discussed and summarized the findings of the literature survey conducted on the problem domain. This chapter will discuss requirements analysis phase conducted on the IDExplorer system. System requirements, requirement elicitation process and stakeholder involvement in the project are explained in this section. Expert and normal survey results are discussed for gathering requirements for the prototype development.

## METHOD OF REQUIREMENT ELICITATION

Following requirement were gathered from different source, and based on requirement engineering techniques in order to cover wide range of the domain and impartiality.

Literature reviews, online surveys, written surveys, e mail correspondence with domain experts and interviews were considered to gather requirements for this research project.

### LITERATURE REVIEW

Literature review was conduct to find out the sate-of-art techniques in SNPs (Single Nucleotide Polymorphism) based approaches for identifies inherited diseases, Gene prioritization for identify inherited diseases, Protein-Protein interaction network for identify inherited diseases.

### ONLINE SURVEY

Questionnaires were prepared for target audiences such as bioinformatician, domain experts, doctors and medical students. The survey was assist to understand aspect of the data analysis techniques of biological methods and clinical system which were not concerned by the author.

### EMAIL CORRESPONDENCES WITH DOMAIN EXPERTS

Email correspondences is a grate techniques stayed in touch with the research community and the medical experts in order to understand their opinions on particular problem domain.

### INTERVIEWS WITH DOMAIN EXPERTS

The author interviewed domain experts to gain more technical knowledge in order to identify the depth of the problem and possible solution. The domain expert group was Dr Prashanth Suravajhala (PHD studies from LJR lab at Rokilde and Aalborg University, Denmark), Lahiru Prabodha (PHD student).

The author also interviewed few medical experts to gain knowledge about diseases which are inherited and identify the relation between genes and diseases.

## SURVEY FINDING – GENERAL USER SURVEY

## SURVEY FINDING – EXPERT SURVEY

## STAKEHOLDERS

The following stakeholders have been identified for inherited diseases identification system. The role of each stakeholder is specified below in Table

Project Supervisor

Other inherited

diseases identifies

systems

(Negative)

(Functional

Beneficiary )

Research Community

Researcher

( Normal

Operator)

**Figure 3: Onion Model Showing Stakeholders of The IDExplorer**

The Wider Environment

The Containing System

The System

Identify inherited diseases based on DNA

Doctor

Doctor

Genome wide

association

National centre of

the biotechnology

information

Other researchers

(Negative)

Clinics

Online mendelian

inheritance

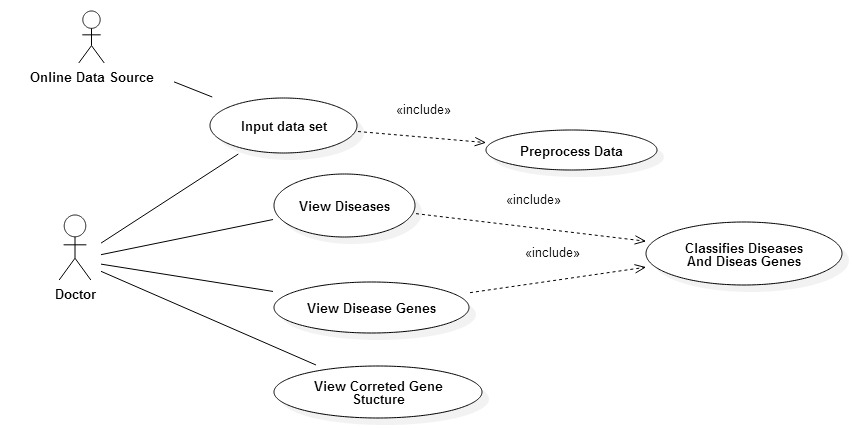
in man

|  |  |  |
| --- | --- | --- |
| **Stakeholder** | **Role** | **View point** |
| Doctor | Functional Beneficiary, Operator | Get useful information from the diseases identifies system |
| Researcher | Intellectual Beneficiary, development | Researcher and develop diseases identifies system. |
| Public | Regulator | Public will be able to get advantages from the inherited diseases identifies system. |
| Clinics | Financial Beneficiary | This system should be recommended to the doctors to find inherited diseases for patient. |
| Doctor | Functional Beneficiary, Operator | Get useful information from the diseases identifies system |
| Researcher | Intellectual Beneficiary, development | Researcher and develop diseases identifies system. |
| Public | Regulator | Public will be able to get advantages from the inherited diseases identifies system. |

## USE CASES

Use Case Diagram explains the functionality of a system and users of the system. The diagram includes the following elements

* Actors - Which represent users of the system
* Use Cases – Which represent functionality provide by the inherited diseases identification system to user



**Figure 4: Usecase Diagram**

## FUNCTIONAL REQUIREMENTS

The main functional requirements are specified bellow in Table 1

(C=Critical, D=Desirable, L=Luxury)

|  |  |  |  |
| --- | --- | --- | --- |
| **Id** | **Functional requirements** | **Priority** | **Use case(s)** |
| FR1 | The system requires genetic data set as input and data will preprocess | C | Input Data Set, Preprocess Data |
| FR2 | Diseases should be identify by the system | C | View Diseases,  Classifies Diseases And Diseases Genes |
| FR3 | Disease genes should be identify by the system | C | View Diseases, |
| FR4 | User should be able to view the corrected gene structure of the particular disease | D | View Corrected Gene Structure |
| FR5 | User should be able to ask question from the system of the particular disease. | D | View Diseases |
| FR6 | Application should generalize to identify most of the inherited diseases | L | View Diseases,  Classifies Diseases And Diseases Genes |

**Table 1: Functional Requirement**

# SYSTEM ARCHITECTURE AND DESIGN

## CHAPTER OVERVIEW

In this chapter, it is expected to reveal the design paradigm taken on the IDEplorer solution based on the concepts which were chosen in the previous chapter. The main purpose of this chapter is to discuss and architectural and design approaches in the proposed solution in detail. First part of this chapter explains the high level architecture of the IDEplorer system and relevant justification for selecting that architectural model. The later part of this chapter discusses the design of each individual module which an explanation of how they relevant to proposed solution of the IDEplorer system.

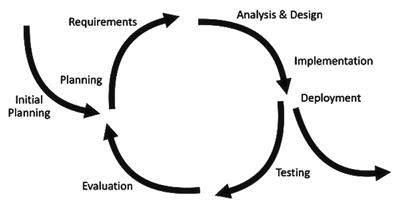
## DESIGN METHODALOGY

Design methodologies are a comprehensive set of principles of the software engineering. Though there are several design methods available, two main suitable designed method are selected for IDExplorer project.

* Iterative and Incremental Development (IID)
* Object oriented Analysis and Design Methodology (OOADM)

### ITERATIVE AND INCREMENTAL DEVELOPMENT

The main advantage of this method is to develop through the repeated cycles (iterative) and in smaller task at a time (incremental). At each iteration, enhance the system functionalities and design modification are made. The process of the IID is shown in Figure 5.



**Figure 5: Iterative and Incremental Development**

This design approach is most suitable for a research project because the requirement and design objective of the research project are changed time to time. So consideration of these factors, flexibility and suitability of iterative and incremental development method fits into IDExplorer system.

### OBJECT ORIENTED ANALYSIS AND DESIGN METHODOLOGY

Object oriented analysis and design methodology is a software engineering approach that models a system as groups of interacting objects. There are various models can be created to show the static structure, dynamic behavior and run time development interaction of the objects. There are a number of different notations for representing the sea models, such as the Unified Modeling Language ([Perera, 2008](#_ENREF_2)).

Object oriented design aspects are suitable for the design purpose of the IDExplorer system.

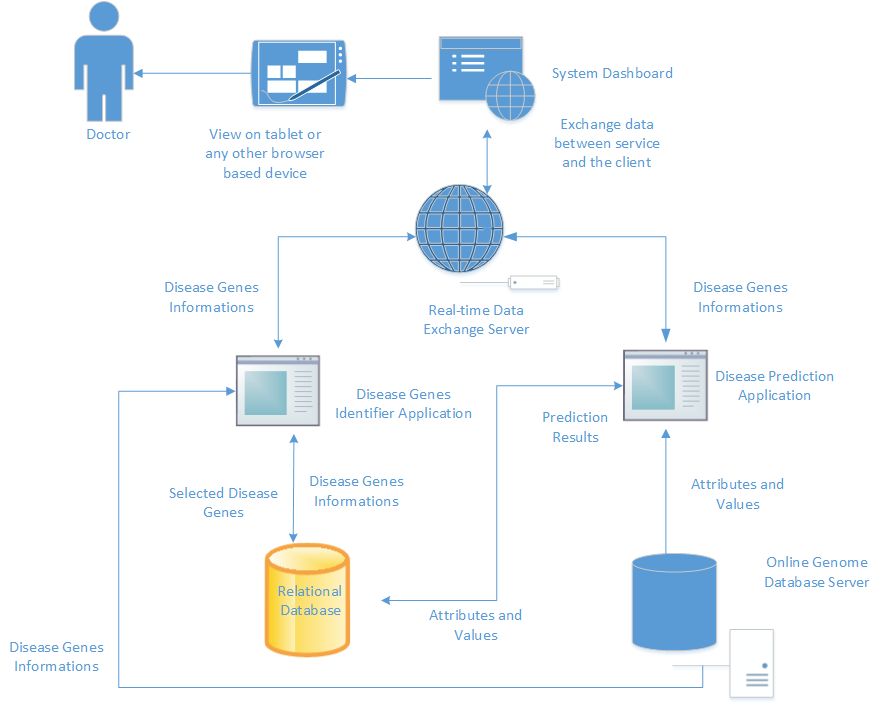
## ARCHITECTURAL GOALS AND CONSTRAINTS

Various architectural best practices and QoS (Quality of Service) factors guided the process of developing the architecture for the IDExplorer system. They are listed below Table2

|  |  |
| --- | --- |
| **Constraint** | **Description** |
| Reusability | The designed architecture will be modularizing so that each module will reuse other modules to achieve its purpose. Intercommunication between modules will be possible. |
| Customizability | The proposed system mainly considers on the inherited disease schema. But generally proposed design can be customized for another disease identification system as well. |
| Extendibility | As the system architecture is modular, it will easily support Extendibility. |
| Portability | This solution designs for a web based system so that the application uses any browser based devices. |
| Availability | The platform should be up to and running 24x7. Because modules are independent, failure of one module should not affect to another module. |

**Table 2: Architectural Best Practices**

## HIGH LEVEL ARCHITECTURE



**Figure 6: High Level Design of the System**

Figure 5 above explains the data flow through the IDExplorer system. The doctor will specify data set for the analysis that they wish to be predicting the disease and prioritize the disease genes. Disease Prediction Application is cable of predicting weather that data set related to disease or not. Disease Genes Prioritize application is cable of select optimal set of genes related to particular diseases.

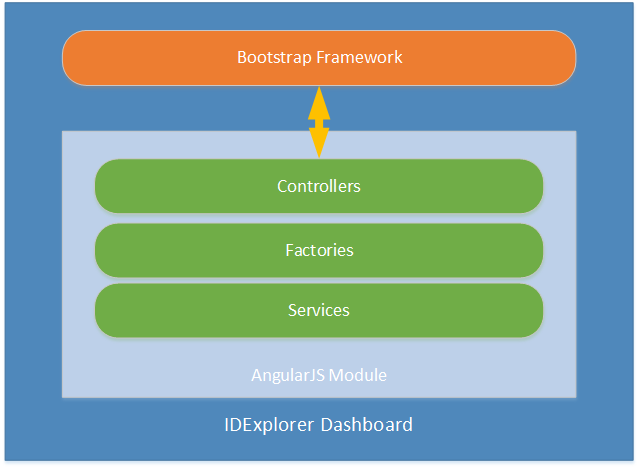
The results of the prediction application and the disease genes prioritize application will be stored in the relational database. The real time data exchange server is working as a data communicator between the dashboard and the subsystems. If prediction and prioritization process completed the data will be immediately transferred to the dashboard. The results will be shown in the tablet application or any other browser based devices.

## DETAILED MODULE DESIGN

### WEB DASHBOARD

The presentation layer of the IDEplorer divided in to server layers. The Boostrap framework application is for the UI framework as HTML and CSS. Modularized design is in the CSS with good practice. Chart components use for the gain usability experience of the application. The dashboard communicates with the real-time server to exchange application to dashboard and user preform and action redirected to the applications.

The AngularJS framework is used for the front end dashboard dynamic components build. Its goal is to augment web applications with model-view-controller (MVC) capability, in an effort to make both development and testing easier. Service layer interacts with the web service developed in the real time data exchange server. For a good design, the controller's logic divided into factories. AngularJS is used for giving rich user experience to the browser based web IDExplorer application users.

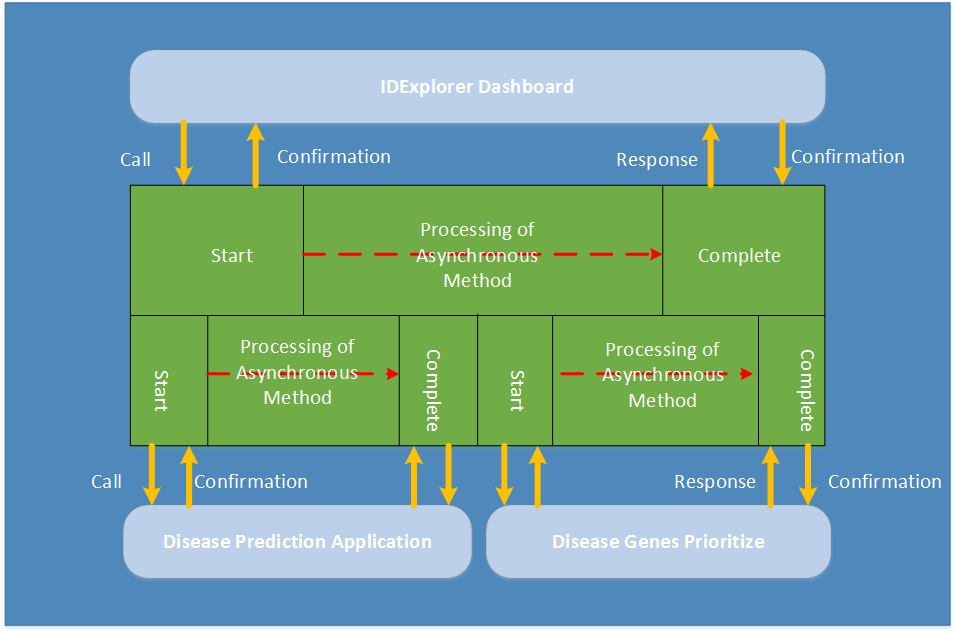


**Figure 7 IDExplorer Dashboard Design**

### REALTIME DATA EXCHANGE SERVER

The real time data exchange is used for exchange application data quickly through the dashboard. It will reduce the users waiting time because data prediction and selection algorithms get much higher time to process and produce results. This also a web service it handles the requests in asynchronous manner.

Figure 7 explains how each client application interacting with the real time data exchange server.



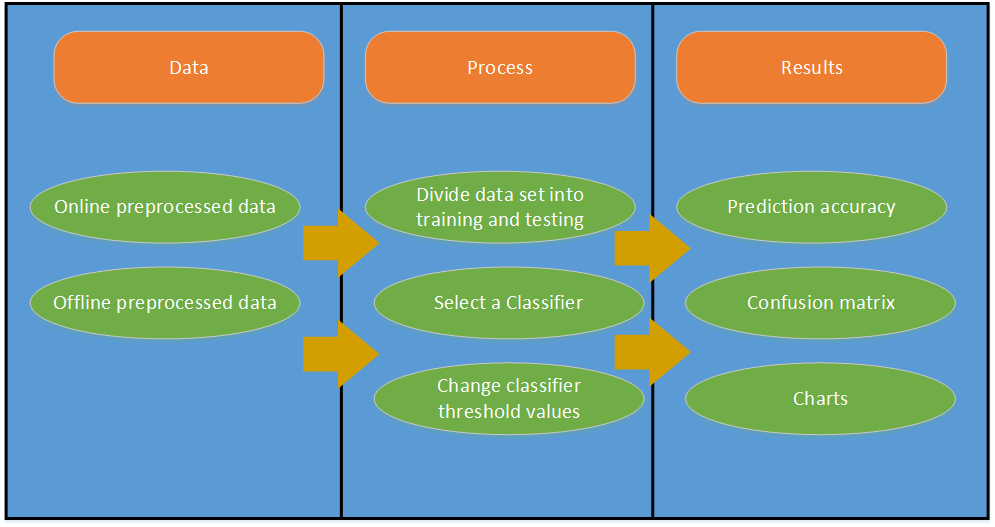
**Figure 8: Real time Data Exchange Server Interacting with Client Applications**

The asynchronous service methods are listed below.

* IDExplorer dashboard and real time data exchange service.
* Insert data set
* Get prediction results
* Change threshold values
* Get chart details
* Get disease genes results
* Real time data exchange service and Disease prediction application.
* Set prediction results
* Set chart details
* Real time data exchange service and Disease genes prioritize application.
  + Set disease genes results
  + Set threshold values

### DISEASE PREDICTION APPLICATION

The disease prediction application is the main component of the IDExplorer system. The prediction application has three major parts. Figure 8 explains the high level view of the prediction process.



**Figure 9: Disease Prediction Application Design**

* Data

Online and offline data set can used for the prediction. When online data sources are using, it needs to format first. The data preprocessing goes through a stepwise process and finally create a single data file ready for a process. Stepwise preprocessing is attribute selection and filtering, values scaling and remove anomalies.

This can be done using an algorithm or a manual way.

* Process

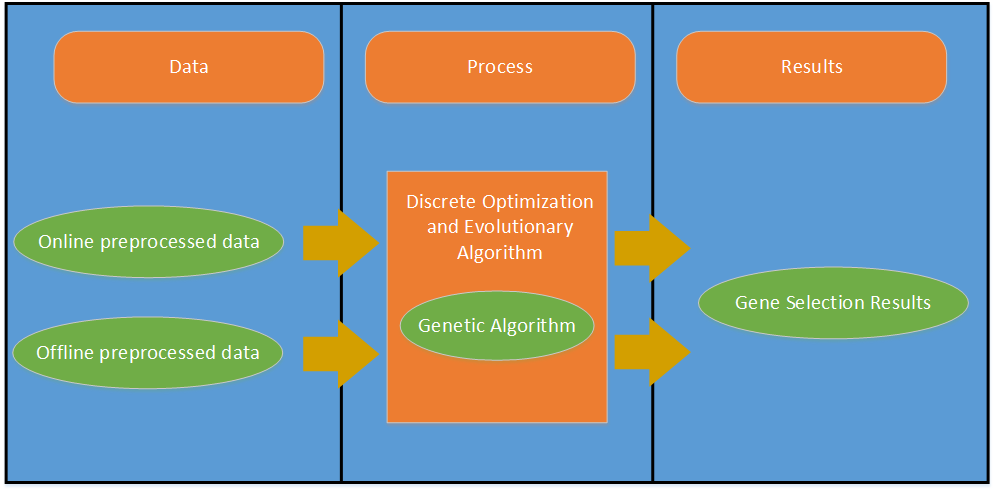
The process is divided into little steps to design purposes. Before selecting a prediction algorithm needs to divide dataset into, two part training and testing data set. Then needs to select a classification algorithm based on literature survey or experience on the trail and errors. This process is iterative process; go through again and again to find the best fit algorithm for the prediction model. Adjustment of the threshold values in a prediction algorithm will produce a better result.

* Results

The design of the result layer is a very important to understand the prediction output. The results contain main details to understand the prediction results. Prediction accuracy value is important for getting a decision about the algorithm. The confusion matrix displays the correctly classified instances. Charts are more important to get an idea about the classification errors and accuracy.

### DISEASE GENES PRIOTIZE APPLICATION

The disease genes prioritize application is a subsystem of the IDExplorer application. The main flow of the application shows below Figure 8.



**Figure 10: Disease Genes Prioritize Application Design**

* Data

The application data tier contains two ways of getting data to the application. The application is designed to read the file formats and read source form the online. The data go through the data preprocessing process and filtering method.

* Process

The process stage design using genetic discrete optimization and evolutionary algorithms based algorithm design. The genetic algorithms chose to design the process layer to respects to an evolutionary algorithm. The genetic algorithm is selected based on the previous literature mentioned in the literature survey chapter.

* Results

The results view is designed to view the selection of the disease genes for particular diseases. It contains the all information about the optimization. The main perspective of the result viewer easily identifies the prioritized set of genes.

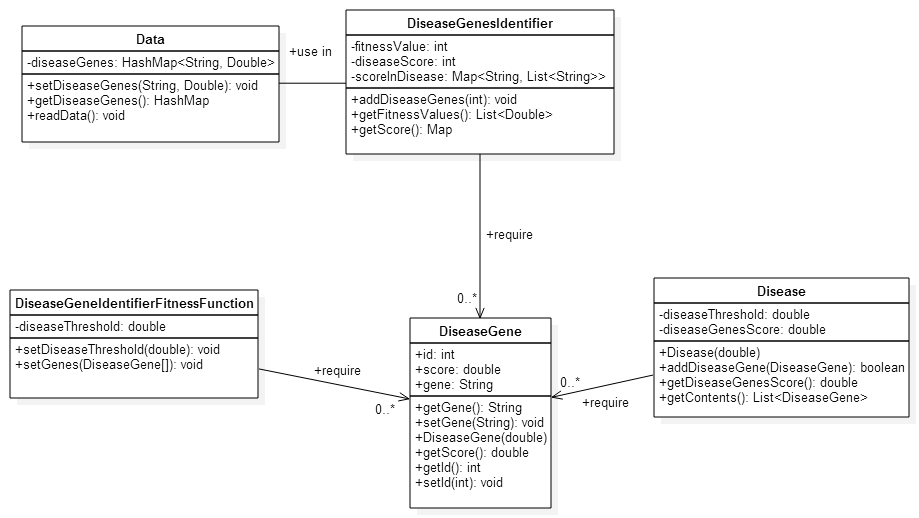
## LOW-LEVEL SYSTEM DESIGN

### CLASS STRUCTURE OF THE DISEASE PREDICTION APPLICATOIN

### FLOW OF SEQUECNE OF THE DISEASE PREDICTION APPLICATION

### CLASS STRUCTRE OF THE DISEASE GENES PRIORITIZE APPLICATION

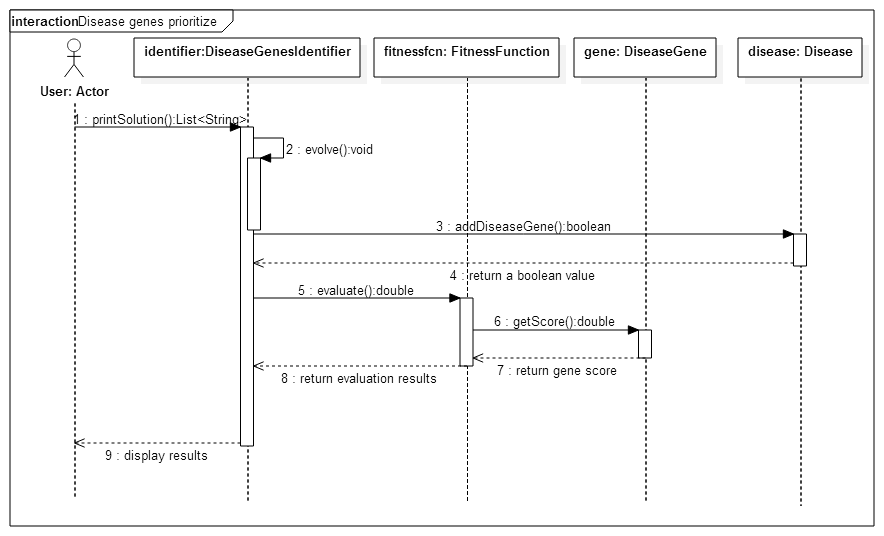
Class diagram of the disease genes prioritize application is shown in Figure 10.



**Figure 11: Disease Genes Prioritize Application Class Diagram**

### FLOW OF SEQUECE OF THE DISEASE GENES PRIOTIZIE APPLICAION

The main sequence of the disease genes prioritize application is shown in Figure 11



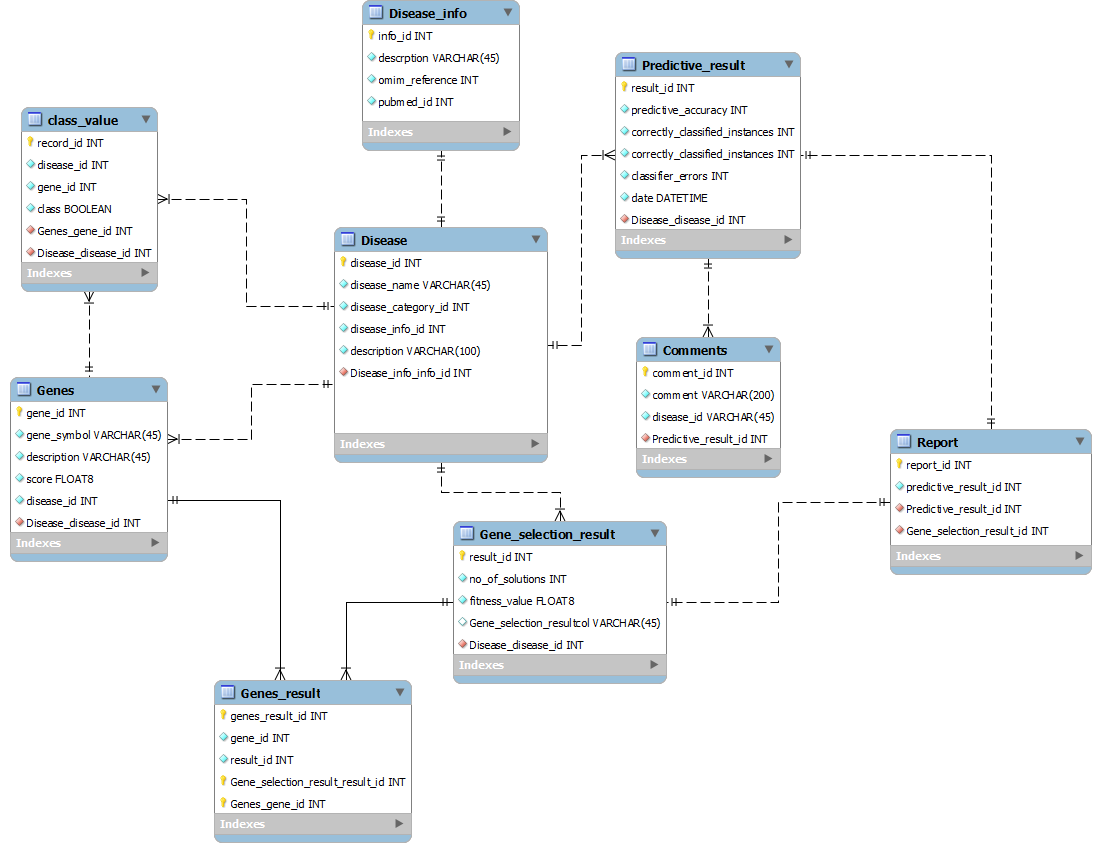
**Figure 12: Disease Genes Prioritize Application Sequence Diagram**

### CLASS STRUCTURE OF THE DASH BOARD

### FLOW OF SEQUENCE OF THE DASH BOARD

### RELATIONAL DATABASE DESIGN

Figure **7** explains how data persistence handle is in the IDExplorer system.



**Figure 13: Relational Database Design**

# SYSTEM IMPLEMENTAION

## CHAPTER OVERVIEW

The previous chapter discussed the design of the IDExplorer system. This chapter aims to discuss how each functionality of the system is implemented according to the requirements which identified in the requirements analysis phase. It will discuss the rationale for selecting the programming language and the development environment. Then explains the implementation of the each subsystem of the prototype.

## SELETION OF JAVA AS THE PROGRAMMING LANGUAGE

Before selecting a Java as a programming language for the IDExplorer system, programming language evolution done for Microsoft .NET, PHP and Python. The prediction application requires more memory and performance so shat web based programming languages are ignored.The Microsoft .NET framework based language are ignored due to no appropriate libraries of the proposed solutions. The Java programming language is open source, N-tier architecture support and optimized language for the specific tasks. Also prediction and discrete optimization framework built on the Java programming language. Therefore Java was chosen as the programming languages to be used.

## SLECTION OF DEVELOPMENT TOOLS

### SELECTION OF ECLIPSE AS THE INTEGRATED DEVELOPMENT ENVIRONMENT (IDE)

Eclipse, Netbeasns and IntelliJ IDEA are most popular IDEs for Java Language. The Netbeans IDE is mostly used in enterprise level application because it is not supported may plug-ins and features. IntelliJ IDEA is mostly used in enterprise level, but it is not distributed on the free license. Eclipse is used by the majority of the enterprise level application and open source development. Therefore Eclipse IDE was selected as the integrated development for Java applications.

### SELECTION OF MYSQL WORKBENCH AS RELATIONAL DATABASE DEVELOPMENT

SQLyog, MYSQL, phpMyAdmin and Workberch are most popular development tools for MYSQL data base development.SQLyog is a grate featured and it is mostly in commercial environment but the tool is not free. phpMyAdmin is a web based MYSQL database development tool and it is not supported most of the database development and design features. MYSQL Workbench is free and open source software; its default MYSQL development tool and its support convert designs into the database schema. Therefor MYSQL Workbench was chosen as development of MYSQL database.

## EXTERNAL LIBRIES AND TECHNOLOGIES USED IN THE IMPLEMENTATION

The external libraries and technologies used in the IDExplorer system is shown in Table 3

|  |  |  |
| --- | --- | --- |
| **Library** | **Version** | **url** |
| Chart.js |  |  |
| weka |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |

**Table 3: External Libraries and Technologies Used**

## IMPLEMENTATION OF THE PROTOTYPE

### WEB DASHBOARD

AngularJs is a front end web application development framework that respects the MVC architecture. The AngularJs framework is highly used to develop single page application (SPA). It’s used to develop dynamic HTML view more manageable way, combination of both AngularJs and Bootstrap produce more elegant user interface. Comparing other front end framework, AngularJs is more robust and modularized framework for rapid application development. They are large open source community are supporting for the AngularJs framework. Therefore AnjularJs framework was selected to develop the web dashboard.

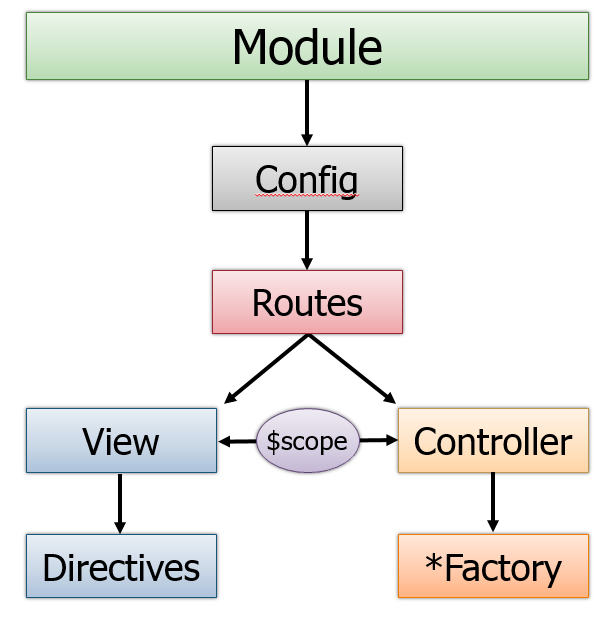
The following features were implemented in the prototype

* Generate the charts according the prediction results.
* Display the optimal set of disease genes.
* Dashboard is responsive.

The AngularJs framework was setup by using Node.js, it is helps to build the automated environment for the AngularJs application development. The initial project setup, there were couple of component installed into Node.js. First node package manager was installed from the Node console. The following components were installed from the node package manager.

* Yeoman – Yemon is used to generate the scaffolding of the application. It is created basic folders, files and configurations to get AngularJs application up and running.
* Bower – Bower is a package manager. It is helps to install AngularJs libraries and Bootstrap framework easily.
* Grunt – Grunt is a server to run the JavaScript and it is helps to update JavaScript, HTML and CSS files while developing the application.

The AnjularJs application structure was used for the development. The structure is shown in Figure 14.



**Figure 14: Components of the AngularJs Application**

Dashboard view was developed based on the Bootstrap UI components and Chart.js HTML5 Charts. The included Chart.js JavaScript library and Bootstrap are shown in the Figure 15



**Figure 15: Included Chart.js Library and Bootstrap**

<!-- Bootstrap core CSS -->

<link href=**"assets/css/bootstrap.css"** rel=**"stylesheet"**>

<!--external css-->

<link href=**"assets/font-awesome/css/font-awesome.css"** rel=**"stylesheet"** />

<link rel=**"stylesheet"** type=**"text/css"** href=**"assets/css/zabuto\_calendar.css"**>

<link rel=**"stylesheet"** type=**"text/css"** href=**"assets/js/gritter/css/jquery.gritter.css"** />

<link rel=**"stylesheet"** type=**"text/css"** href=**"assets/lineicons/style.css"**>

<!-- Custom styles for this template -->

<link href=**"assets/css/style.css"** rel=**"stylesheet"**>

<link href=**"assets/css/style-responsive.css"** rel=**"stylesheet"**>

<script src=**"assets/js/chart-master/Chart.js"**></script>

The following code explains the instance count in each attribute.

**Figure 16: Bar Chart HTML Code Snippets**

<div class=**"col-lg-6"**>

<div class=**"content-panel"**>

<h4><i class=**"fa fa-angle-right"**></i> **Bar**</h4>

<div class=**"panel-body text-center"**>

<canvas id=**"bar"** height=**"300"** width=**"400"**></canvas>

</div>

</div>

</div>

The AngularJs application was built on the several modules. The landing view is the main view of the dashboard and there are two main routes to navigate disease prediction results view and the disease genes prioritize results view. The main modules are shown in Figure 17

**Figure 17: Main AngularJs Modules in Dashboard**

**var** app **=** angular

**.**module**(**'IDExplorerApp'**,** **[**

'ngAnimate'**,**

'ngCookies'**,**

'ngResource'**,**

'ngRoute'**,**

'ngSanitize'**,**

'ngTouch'**,**

'landingViewModule'**,**

'chartModule'**,**

'loginModule'**,**

'signupModule'**,**

'diseasePredictionModule'**,**

'diseaseGenesPrioritizeModule'**,**

'navigationBar'

**])**

The each and every module is included factories and serveries to reduce the cupping of the controller. The logic of the each module is in the factories and service code is used to make a service calls to web service. The main purpose of factories and service was used for a good software engineering practices.

### REAL-TIME DATA EXCHANGE SERVICE

### DISEASE PREDICTION APPLICATION

Technology and the library were chosen for disease prediction application based on the requirement analysis chapter and the literature survey chapter. The Weka is a collection of machine learning algorithms for data mining purpose. The Weka code is free and open source and it is customizable comparing with tools like MATLAB. Weka contains tools for data-preprocessing, classification, regression, clustering, association rules and visualization.

### DISEASE GENES PRIORITIZE APPLICATION

### REALTIME DATA EXHANGE SERVER

### DISEASE PREDICTION APPLICATION

### DISEASE GENES PRIORITIZE APPLICATION

# TESTING AND EVALUATION

# CONCLUTION

*Weka 3: Data Mining Software in Java* [Online]. Available: <http://www.cs.waikato.ac.nz/ml/weka/> 2015].

PERERA, P. 2008. *Technology Adapted* [Online]. University of Moratuwa. Sri Lanka. Available: <http://dl.lib.mrt.ac.lk/bitstream/handle/123/1781/93001_3.pdf?sequence=5> 2015].