**Hybrid Feature Selection Using Correlation Coefficient and Particle Swarm Optimization on Microarray Gene Expression Data**

Abstract:

Diagnosis of cancer is one of the most emerging clinical applications in microarray gene expression data. However, cancer classification on microarray gene expression data still remains a difficult problem. The main reason for this is the significantly large number of genes present relatively compared to the number of available training samples. In this paper, we propose a hybrid feature selection approach that combines the correlation coefficient with particle swarm optimization. The process of feature selection and classification is performed on three multi-class datasets namely Lymphoma, MLL and SRBCT. After the process of feature selection is performed, the selected genes are subjected to Extreme Learning Machines Classifier. Experimental results show that the proposed hybrid approach reduces the number of effective levels of gene expression and obtains higher classification accuracy and uses fewer features compared to the same experiment performed using the traditional tree-based classifiers like J48, random forest, random trees, decision stump and genetic algorithm as well.

1. **INTRODUCTION**

The main objective of this project is to provide security using Hybridization method is used to generate DNA samples in microarray gene expression data. This process can be done in two ways. In the first method, during the hybridization process, messenger RNA (mRNA) is stained using matrices sample taken from tissue or blood stream becomes cDNA. RNA profiling can be noisy and may not be sampled unevenly over time [1].

The second method is the Affymetrix chips are hybridized using oligonucleotides on the surface of the array chip. It is possible to monitor and simultaneous measure thousands of activation levels of gene expression in a single experiment. This is considered as the key advantage of DNA microarray technology. Protein production helps identify the different types of memberships. This is achieved because the gene expression level refers to a specific protein production gene [2]. The clinical medicine progress is only possible because of valuable results produced by microarray experiments performed on a variety of issues of gene expression profile. Microarray data can be applied to the problems of classification of cancer also. This was a recent development in the field of clinical research.

Microarray data on cancer DNA is combined with statistical techniques for analyzing gene expression profiles to identify potential biomarkers for diagnosis and treatment of various types of cancer [3]. Statistical analysis of differentially expressed genes helps to assign them to different classes. This process improves the understanding of basic biological processes in the system. Using the concept of technology of microarray gene expression, it is possible to study the simultaneous activity of thousands of genes. The relative abundance of mRNA in the gene can be found by using gene expression profiles [4]. Results obtained represent the state of the cell. Discriminant analysis of microarray data is an excellent tool for medical diagnostics of diseases, treatment and prevention.

The main purpose of the classification is to build an effective model that can identify differentially expressed genes and could also be used to identify classes in the unknown samples [5]. Some of the challenges in the microarray data are the smallest number of training and testing data available, the higher dimensionality of the data and the variations that could sneak in experiments performed to estimate the levels of gene expression. The two main tasks in the analysis of gene expression microarray are feature Selection and classification. To perform the classification process with an acceptable level of accuracy, the process of feature selection becomes crucial. Microarray gene expression data contains hundreds of thousands of genes or feature information.

Only a small subset of genes exhibit strong correlation between them. Feature selection is a process that effectively selects differentially expressed genes in the dataset and forms a new subset for efficient classification. There may be situations in which a low-ranked gene could perform well in the rankings and a critical gene could be left out in the selection of functions [6].

The prediction accuracy would increase only with the best method of feature selection which otherwise would be impossible to understand. Another important measure is to avoid overfitting and build faster and cost effective models [7]. In this study, we use a hybrid approach that combines the benefits of a filter and a wrapper to perform feature selection. They are easy to use, simple and computationally efficient [8].

**1.1 Motivation**

The motivation behind the development of a project is to provide security using generate training module and J48, EML, Random forest algorithm can easily perform transactions without the fear of losing credentials.

**1.2 Problem definition**

* Now-a-days, ensuring the security for every individual is essential.
* We have only prevention methods and the security issues are a lot and not up to the mark.
* Microarray data can be applied to the problems of classification of cancer also
* This was a recent development in the field of clinical research.
* Microarray data on cancer DNA is combined with statistical techniques for analyzing gene expression profiles to identify potential biomarkers for diagnosis and treatment of various types of cancer

**1.3 Limitations of the project**

* We don’t have such kind of integrated application in different platforms.
* The list of datasets text and manual only works with the help of application (Recommended accuracy).

**2. ANALYSIS**

After gathering the requirements of the task to be performed, the next step is to analyse the problem and understand.Design models give us a template that guides us in constructing a system

**2.1 Introduction**

* Our project is Java based Standalone application (Desktop).
* With the help of our application, we can perform different dataset transactions in a secure way and we can generate training module and J48, EML, Random forest algorithm etc.
* It has a special Microarray Gene Expression Data feature of viewing Hybrid Feature Selection
* The admin can add different data sets apply the Correlation Coefficient and Particle Swarm Optimization

**2.2 Existing System**

* In the existing system, A multivariate approach is effective in identifying the correlation that exists among the different genes in the dataset.
* Pearsons correlation coefficient is very sensitive to the presence of outliers and noise.
* there is no such application which uses the concept of for providing accuracy of the datasets.
* Random accuracy is 27% now click on Run K means Algorithm

**2.2.1 Limitations of Existing System**

It includes techniques such as tricking datasets through generated accuracy percentage and detecting cluster micro array dataset installation of key loggers and screen captures.

* Only Preventive measures are used to eradicate this.
* These approaches do not provide security and leads to misuse of datasets**.**

**2.3 Proposed System**

We try to overcome the limitations of the existing system

* The information cannot be correlation and coefficient of particle datasets.
* Various feature ranking and feature selection techniques have been proposed such as Correlation-based Feature Selection (CFS), Principal Component Analysis (PCA), Gain Ratio (GR) attribute evaluation, Chi-square Feature Evaluation, Fast Correlation-based Feature selection (FCBF),

**2.4 Software Requirements Specifications**

Software Requirement Specification (SRS) is the starting point of the software developing activity. As system grew more complex it became evident that the goal of the entire system cannot be easily comprehended. Hence the need for the requirement phase arose. The software project is initiated by the client needs. The SRS is the means of translating the ideas of the minds of clients (the input) into a formal document (the output of the requirement phase.)

The SRS phase consists of two basic activities:

**Problem Analysis**

The process is order and more nebulous of the two, deals with understand the problem, the goal and constraints.

**Requirement Specification**

Here, the focus is on specifying what has been found giving analysis such as representation, specification languages and tools, and checking the specifications are addressed during this activity.

The Requirement phase terminates with the production of the validate SRS document. Producing the SRS document is the basic goal of this phase.

## Intended Audience and Reading Suggestions

This document is for better understanding for Remote desktop control. Mainly intended for Head of the Dept., Internal guide, External guide, Staff members, Users and colleagues. This detail given below guides every normal user to how to go through this document for better understanding. The sequence to follow for better understanding is here Purpose, Scope, Technologies used, Modules present in the project, References etc.

**Role of SRS**

The purpose of the Software Requirement Specification is to reduce the communication gap between the clients and the developers. Software Requirement Specification is the medium though which the client and user needs are accurately specified. It forms the basis of software development. A good SRS should satisfy all the parties involved in the system.

**2.4.1 Purpose**

The main purpose of this standalone service is to provide flexibility and use all the services of the desktop for end-users through offline or online.

**2.4.2 Scope**

As the technology is growing rapidly and many of the things are done online these days, so we have come up with an idea about creating an standalone application which acts an interface to datasets generated with Random class Algorithm by avoiding accuracy. It is very efficient and secure.

**DISTRIBUTED FILE SYSTEM**

**Introduction**

A distributed file system (DFS) is a file system with data stored on a server. The data is accessed and processed as if it was stored on the local client machine. The DFS makes it convenient to share information and files among users on a network in a controlled and authorized way. The server allows the client users to share files and store data just like they are storing the information locally. However, the servers have full control over the data and give access control to the clients.

There has been exceptional growth in network-based computing recently and client/server-based applications have brought revolutions in this area. Sharing storage resources and information on the network is one of the key elements in both local area networks (LANs) and wide area networks (WANs). Different technologies have been developed to bring convenience to sharing resources and files on a network; a distributed file system is one of the processes used regularly.

One process involved in implementing the DFS is giving access control and storage management controls to the client system in a centralized way, managed by the servers. Transparency is one of the core processes in DFS, so files are accessed, stored, and managed on the local client machines while the process itself is actually held on the servers. This transparency brings convenience to the end user on a client machine because the network file system efficiently manages all the processes. Generally, a DFS is used in a LAN, but it can be used in a WAN or over the Internet. A DFS allows efficient and well-managed data and storage sharing options on a network compared to other options. Another option for users in network-based computing is a shared disk file system. A shared disk file system puts the

access control on the client’s systems so the data is inaccessible when the client system goes offline. DFS is fault-tolerant and the data is accessible even if some of the network nodes are offline.

**Client**

Client more than one client may access the same data simultaneously, the server must have a mechanism in place (such as maintaining information about the times of access) to organize updates so that the client always receives the most current version of data and that data conflicts do not arise.

**Server**

Server is a system which receives request or commands from client and gives back the response according to the request. Server can run on any type computer.

**Challenges in HDFS**

* DFS due to failure of hardware components data do not reach the destination point.
* Data in node can get altered or corrupted.
* Lack of performance and scalability.
* Lack of flexible resource management.
* Lack of application deployment support.
* Lack of quality of service.

Lack of multiple data source support.

**2.4.3 Overall Description**

**Introduction to Python**

Python is interpreted object-oriented language. It is also referred to as a high-level programming language. It is designed by Guido Van Rossum. It was initially released in the year 1991. The filename extensions are .py, .pyc, .pyd, .pyo. It enables the straight and easy programming for both small and large applications. It mainly emphasizes on code reusability, readability and using white space. Python uses expressions mainly similar to C language and its methods and typing. Python supports multiple programming paradigms like functional programming, imperative and procedural as well.

**Importance of Python to the Internet**

Python can be more user-friendly because of its advantages. Please find below the uses of python language for different reasons:

#### 1. Applications:

Python can be used to develop different applications like web applications, graphic user interface based applications, software development application, scientific and numeric applications, network programming, Games and 3D applications and other business applications. It makes an interactive interface and easy development of applications.

#### 2. Multiple Programming paradigms:

Python is also used because of its providing continuous support to several programming paradigms. As it [supports object-oriented programming](https://www.educba.com/functional-programming-vs-oop/) and structured programming. Python has features, which also support various concepts of functional programming language. It is used for dynamic type system and automatic memory management. Python language features and programming paradigms allow you for developing the small as well as large applications. It can be used for [complex software applications](https://www.educba.com/what-is-application-software-its-types/).

#### 3. Robust Standard Library:

Python has a large and robust standard library to use for developing the applications. It also makes the developers use Python over other languages. The standard library helps in using the different range of modules available for Python. As this module helps you in adding the functionality without writing any more code. To get the information about various modules, documentation on python standard library can be referred. While [developing any web application](https://www.educba.com/how-to-build-web-applications-using-mongodb/), implementing web services, performing string operations and other usages like interface protocol, the standard library documentation helps.

4. Compatible with Major Platforms and Systems:

Python is mainly compatible with major platforms and systems because of which it is used mainly for developing applications. With help of python interpreters, python code can be run on specific platforms and tools as it supports many operating systems. As python is an interpreted high-level programming language and it allows you to run the code on multiple platforms. The new and modified code can be executed without recompiling and its impact can be monitored or checked. It means it’s not required to recompile the code after every change. This feature helps in saving the development time of the developers.

5. Access of Database:

Uses of Python also helps in accessing the database easily. Python helps in customizing the interfaces of different databases like MySQL, Oracle, Microsoft SQL Server, PostgreSQL, and other databases. It has an object database like Durus and ZODB. It is used for [standard database](https://www.educba.com/careers-in-database-administration/) API and freely available for download.

6. Code Readability:

Python code is easy to read and maintained. It is easily reusable as well wherever it is required. Python’s having simple syntax, which allows the different concepts to develop without writing any additional code. The code should be of good quality and easy to maintain the source code and simplify the maintenance, which is required to develop the software application. It also emphasizes code readability, which is the great feature, unlike other programming languages. It helps in building custom applications and clean code helps in maintaining and updating the software applications without putting extra effort on the same code.

7. Simplify Complex Software Development:

Applications of Python is used to simplifying the [complex software development process](https://www.educba.com/software-development-vs-web-development/) as it is a general-purpose programming language. It is used for developing the complex application like scientific and numeric application, and for both desktop and web applications. Python has features like analyzing data and visualization, which helps in creating custom solutions without putting extra effort and time. It helps you to visualize and present data in an effective way.

8. Many Open Source Frameworks and Tools:

Python is open source and easily available. This also helps in costing the [software development](https://www.educba.com/what-is-software-development/) significantly. There are many [open source applications of python frameworks](https://www.educba.com/python-frameworks/), libraries, and development tools for developing the application without putting extra cost. Python frameworks simplify and make the process faster for web application development and the frameworks are Django, Flask, pyramid etc. Python GUI frameworks are available for developing the GUI based application.

9. Adopt Test Driven Development:

Python makes coding easier as well as testing with help of adopting Test Driven Development approach. The test cases can be easily written before any code development. Whenever the code development started, the written test cases can start testing the code simultaneously and provides the result. These can also be used for checking or testing the pre-requirements based on the source code.

10. Other applications for which python is used:

There are other applications for which python is used that are Robotics, web scraping, scripting, artificial intelligence, data analysis, machine learning, face detection, color detection, 3D CAD applications, console-based applications, audio-based applications, video-based applications, enterprise applications, and applications for Images etc. These are some major applications used.

Conclusion – Uses of Python

In this Uses of Python article, we have seen that python is one of the major languages are used for the development of both desktop and web applications. Python has features that take care of common programming tasks. Python is simple to learn and easy to use. Sometimes, python marks as slower than other widely used programming languages like Java. Python applications can speed up by simply maintaining the code and using custom runtime.

Python does support the modules and packages, which encourages program modularity and code reuse. Python provides an increase in productivity, which makes the first choice of developers. It has a great learning curve as it supports functional and procedural programming language. It is open source and can be freely distributed. The programming language mainly selected based on the requirement and compatibility with platforms and database.

In today’s world, [Python](https://www.educba.com/course/python-training-course-bundle/) is one of the most used language. It’s not just a language, it’s a way to do things in a proper, simple and compact manner. Python is one of the most known high-level language especially in open source. It sure is a vast python language. Point being, the more you know about Python Programming, the more you will know that there are still things that you don’t know. I know that the statement is ironic, but that’s how Python is.

Python Programming is language, which is fast, compact and can be easily rendered into any operating system. Besides, if you look around, you will see that majority of the things run on Python; you can take examples of Google or YouTube. There are frameworks which exclusively support Python such as [Django](https://www.educba.com/course/django-unchained-with-python/) and Pyramid. There are even micro-frameworks such as Flask and Bottle. Python’s standard library consists of many internet protocols such as [HTML](https://www.educba.com/course/online-html-5-training/), [XML](https://www.educba.com/course/xml-training-course/), JSON, E-mail Processing, Support for FTP, IMAP and its Easy-to-use [Socket interface](https://www.educba.com/what-is-python-socket-programming/). Most common purpose of usage of Python is for scientific and numeric computing. For example, SciPy, Pandas and iPython.

So now you know what Python is and how it works, let’s get to know some of the reasons why I support Python more than any other language. I have been around Python since a long time. Long enough to understand the importance of basic and advanced interpretations of the python language. But still there are times, when even people who have worked long enough with Python will forget some of the most basic important stuff in Python Programming. So, to give you guys a refresher, I have written this article.

Whether you an Expert or a Beginner, these things will always remind you why Python Programming is still the most flexible, simple and creative python language of all time. I have taken my time to explain these examples in detail, so if you still have a hard time understanding it, you can simply Google the titles I have given to these examples and you will easily understand what I am talking about. So, without much further ado, let’s start.

So these are few tricks I have learnt over the past couple of years, which, I keep saved in a text file. So, anytime I am stuck at any piece of code, I just take a glimpse of these things and most of my problems are solved many of the times.

The best thing of python is that you can create your own modules. So, for example, I can create my own function and modules and place them all together in a separate folder. So what I do is I write down specific codes, which I know I would be using in common in most of my work, then convert them in a module and keep it aside in a separate folder. By doing this, I save a lot of time, from writing them again and debugging them to check for errors. Another reason for that is, you need to keep your program efficient and manageable if they are big in size. So, to manage this, you can break them into separate files, put multiple functions and definitions into a file and use them by importing them into scripts and programs. Note that these files will have a \*.py extension. And once you import them, it will auto-create a \*.pyc extension file that will load much quicker than the normal \*.py file.

### 2. True and False

This is also one of the most used method. If you have ever played games, and by games, I mean high-end Games, you must have noticed that sometimes you have to lower the graphics. But again, sometimes, even you cannot find these options in the game. So, what you normally do is, find the config file in the documents folder and change it. For example, you change the Vsync = True, or False as per the situation. But the one which I am trying to explain here is somewhat different but still related to it. True equals to 1 and False equals to 0 in python. In short, true means you agree and false means you disagree. These things are used most often in python. So, you can either assign True and False statements by using “=” sign or you can check the equality by using “==” sign. As simple as that.

**3. Python Performance debug**

When writing a program, our main goal is to make the program efficient, fast and compact. But there are times, when you simply cannot make the program compact. So at these times, you may not actually want to make the program compact to make it faster. What you can do is, for example, when handling codes in a dictionary, you can try an alternate method of dictating an item. Confused? Let me explain this. You can simply choose an option to add an item directly and then check whether the inserted items exist or needs to be updated. So, by doing this, you don’t need to check each and every other item to match with it and then update it which kind of will make the application slow. Following is a famous example of that:

*p = 16*

*myDiction =*{}

*for i in range(0, p):*

*char = 'abcd'[i%4]*

*if char not in myDict:*

*myDiction[char] = 0*

*myDiction[char] += 1*

*print(myDiction)*

The above example is the normal way of writing it. Now here is how it will make the code run faster,  
*p = 16*  
*myDiction =*{}  
*for I in range(0, p):*  
*char = ‘abcd’[i%4] try:*  
*myDiction[char] += 1*  
*except KeyError:*  
*myDiction[char] = 1*  
*print(myDiction)*

### 4. Py2exe

Another useful tip I can think of is py2exe. Normally when writing a code in any language, sometimes it can be a hassle to actually compile them into an executable, especially if you are using windows. But for python Programming, it’s actually very simple. You can simply download py2exe which again is an [open source software](https://www.educba.com/myths-misconceptions-about-open-source-software/)which you can download from sourceforge.net. Using this app, you can simply convert even your modules into an exe, unlike [C](https://www.educba.com/course/c-bundle/) or [C++](https://www.educba.com/course/online-c-plus-plus-programming-training/) which is actually a stress of mind when compiling into an exe.

**Recommended courses**

* [IT Security Training Bundle](https://www.educba.com/course/it-security-fundamentals)
* [Ruby on Rails Certification Course](https://www.educba.com/course/ruby-on-rails-from-scratch)
* [CSS Training Bundle](https://www.educba.com/course/css-training-tutorials)
* [Professional HTML5 and CSS3 Training](https://www.educba.com/course/online-html5-and-css3-training)

### 5. Sets

If you are a maths freak, you would surely love this next tip. You may have used sets in your lower classes. Remember something? Yeah, exactly, Unions and stuff. So, there are people like me, who don’t like to use automated software’s sometimes. The reason for that is Security. Let’s take a simple example of [Microsoft Excel](https://www.educba.com/course/microsoft-excel-2016-basic-course/). Some people tend to use excel, only to group and create a database. They just need that and a good security for that. They are not interested in formatting the text, colour and stuff. So, what I do at those times, is I create my own python Programming software stack and create my own database. For, some of my security reasons, I prefer [Python](https://www.educba.com/course/python-programming-zero-to-hero/) over [MYSql](https://www.educba.com/course/mysql-training/). So, coming back to my point of sets, Sets are extremely useful when creating databases. Especially, when you want to find matches, create groups and other similar tasks. Following is a simple example of that.

*>>> A =*{*1, 2, 3, 3*}  
*>>> A*  
*set([1, 2, 3])*  
*>>> B =*{*3, 4, 5, 6, 7*}  
*>>> B*  
*set([3, 4, 5, 6, 7])*  
*>>> A | B*  
*set([1, 2, 3, 4, 5, 6, 7])*  
*>>> A & B*  
*set([3])*  
*>>> A - B*  
*set([1, 2])*  
*>>> B - A*  
*set([4, 5, 6, 7])*  
*>>> A ^ B*  
*set([1, 2, 4, 5, 6, 7])*  
*>>> (A ^ B) == ((A - B) | (B - A))*  
*True*

### 6. Merging Python and Shell Scripts

Now this is something you can’t do with [C or C++](https://www.educba.com/course/learn-c-and-c-online-course/). If you are an open source guy, then you would surely be using [Linux](https://www.educba.com/course/linux-training-course-bundle/) as the main Operating OS, or at least a Dual Boot. So, [Linux already includes python](https://www.educba.com/course/linux-system-administration-with-python/). And python is extremely compatible with Linux. This gives us a benefit of compiling and merging them together. You can simply create a script which can work as a normal [unix](https://www.educba.com/course/unix-1/) script as well as an interpreted python code at the same time. When writing a [shell script](https://www.educba.com/shell-scripting-in-linux/), you need a four quote characters and an empty string to the shell, but in python, you need to do that with a triple-quoted string with a quote character. Remember that, the first string in a script can be easily stored as a doc string for a module, but after that, the python interpreter will simply ignore it.  
Example is as follows:

*#!/bin/sh*  
*\_\_doc\_\_ = """*  
*Demonstrate how to mix Python + shell script.*  
*"""*  
import*sys*  
*print "Hello World!"*  
*print "This is Python", sys.version*  
*print "This is my argument vector:", sys.argv*  
*print "This is my doc string:", \_\_doc\_\_*  
*sys.exit (0)*

### 7. JSON-esque

Python has a lot of hidden stuff underneath. It only takes a person and his time to find out what all magical operators and stuff are hidden inside. One among all the other stuff is the famous JSON-esque. You can create nested dictionaries without explicitly creating sub-dictionaries. They magically come into existence as we reference them.  
Example as follows:  
*users = tree()*  
*users['harold']['username'] = 'hrldcpr'*  
*users['handler']['username'] = 'matthandlersux'*  
  
Now you can print the above as JSON with:  
*>>>print(json.dumps(users))*

And it will look like this:  
{*"harold":*{*"username": "hrldcpr"*}*, "handler":*{*"username": "matthandlersux"*}}

### 8. Pip

Pip is something maybe most people know of. But still it is an awesome stuff that you need to know if you are starting with python. Sometimes, you need to inspect the source of a package before installing it. Most of the times, it’s for installing a newer version of some package. So, you can simply install pip and do the following:

*>>> pip install --download sqlalchemy\_download sqlalchemy*  
*>>>pip install --no-install sqlalchemy*  
*>>>pip install --no-download sqlalchemy*  
  
If you want to install the bleeding-edge version of a package, you can directly check it from the GIT repository.

*>>>pip install git+https://github.com/simplejson/simplejson.git*  
*>>>pip install svn+svn://svn.zope.org/repos/main/zope.interface/trunk*

### 9. Virtualenv

Another important function of python is the Virtualenv. Virtualenv means Virtual Environment. This, now my friends, is a very awesome function of python. Basically, to test python in different conditions, you would normally you would have to change the global python environment. But, one of the key benefits of sandboxing your python environment is that you can easily test one code under different python versions and package dependencies. To install virtualenv, you need to install pip first.  
You can do as follows:

*easy\_install pip*  
*pip install virtualenv*  
*virtualenv python-workspace*  
*cd python-workspace*  
*source ./bin/activate*  
*python*

### 10. Zen of Python

Last but not the least, it’s the Zen of python. Zen of python is a mini-guide for python programming. Even if you don’t program python, it’s still an interesting thing to read. Just simply go to the python interpreter and type:

*>>>*import*this*

And I won’t be explaining this one. Because it’s simply amazing to try it out yourself.

And now, we have come to the end of the road. These are just my experiences, but if you find some, sure do post it so that the world knows how exactly awesome python is.

First Image Source: pixabay.com

**Feasibility Study**

Feasibility study is an important phase in the software development process. It enables the developer to have an assessment of the product being developed. It refers to the feasibility study of the product in terms of outcomes of the product, operational use and technical support required for implementing it.

Feasibility study should be performed on the basis of various criteria and parameters.

The various feasibility studies are:

* Operational feasibility
* Technical feasibility
* Economic feasibility
* Feasibility Report

**Operational Feasibility**

This study is carried out to check the technical feasibility, that is, the technical requirements of the system. Any system developed must not have a high demand on the available technical resources. This will lead to high demands on the available technical resources. This will lead to high demands being placed on the client. The developed system must have a modest requirement, as only minimal or null changes are required for implementing this system.

**Technical Feasibility**

The technical issue usually raised during the feasibility stage of the investigation includes the following:

* Does the necessary technology exist to do what is suggested?
* Do the proposed equipments have the technical capacity to hold the data required to use the new system?
* Will the proposed system provide adequate response to inquiries, regardless of the number or location of users?
* Can the system be upgraded if developed?
* Are there technical guarantees of accuracy, reliability, ease of access and data security?

Earlier no system existed to satisfy the user needs with less delay. The current system developed is technically feasible. Thus it provides an easy access to the users. Therefore, it provides the technical guarantee of accuracy, reliability and security. The software and hard requirements for the development of this project are not many and are already available in-house at NIC or available as free as open source. The work for the project is done with the current equipment and existing software technology. Necessary bandwidth exists for providing a fast feedback to the users irrespective of the number of users using the system.

**2.4.5 Economic Feasibility**

This study is carried out to check the economic impact that the system will have on the organization. The amount of fund that the company can pour into the research and development of the system is limited. The expenditures must be justified. Thus the developed system as well within the budget and this was achieved because most of the technologies used are freely available. Only the customized products had to be purchased.

**Feasibility Report**

After collecting user requirements and generating all information from existing system, feasibility report having some specifications and limitations are submitted to guide for approval.

**2.4.6 Technologies used**

**Hardware Requirements**

* Hardware : Pentium based system with a minimum of P4
* RAM : 1GB (minimum)

**Software Requirements**

* Operating System : Windows 7
* Technology : Python

**3. MODULES**

**3.1 Modules Description**

Our application consists of three modules

The main aim of this project is to select features from dataset by applying PSO features selection algorithm to reduce dataset size.

el

This project consists of following modules

1. Pre-processing: using this module we will remove out empty values from dataset and convert non digits attribute such as ‘?’ to 0 digits.
2. Generate Model: using this module will convert dataset into train model
3. Features selection: using this module we will apply PSO and Correlation algorithm to select features.
4. Run algorithm module: using this module we will run various algorithms such as J48 tree, random forest and propose EML (Extreme Machine Learning) algorithms
5. Graphs: using this will compare accuracy of all algorithms in graphs

**3.2 PROJECT ARCHITECTURE**

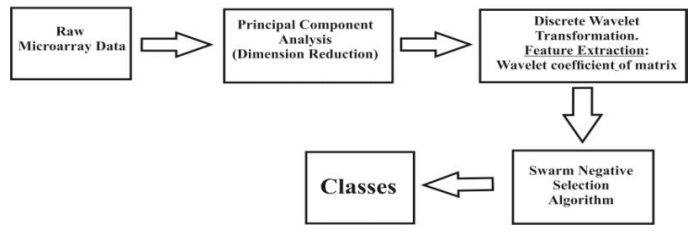


Fig.3.2: Architecture.

**4. DESIGN**

4.1 Introduction

The Unified Modeling Language (UML) is a standard language for writing software blue prints. The UML is a language for

i. Visualizing

ii. Specifying

iii. Constructing

Document the artifacts of a software intensive system.

The UML is a language which provides vocabulary and the rules for combining words in that vocabulary for the purpose of communication. A modeling language is a language whose vocabulary and the rules focus on the conceptual and physical representation of a system. Modeling yields an understanding of a system.

Building Blocks of the UML

The vocabulary of the UML encompasses three kinds of building blocks

1. Things
2. Relationships
3. Diagrams

Things are the abstractions that are first-class citizens in a model; relationships tie these things together; diagrams group interesting collections of things.

Things in the UML

There are four kinds of things in the UML

i. Structural Things

ii. Behavioral Things

iii. Grouping Things

iv. Annotational Things

Structural things are the nouns of UML models. The structural things used in the project design are:

First, a class is a description of a set of objects that share the same attributes, operations, relationships and semantics.

|  |
| --- |
| Window |
| origin  size |
| open()  close()  move()  display() |

##### Fig.4.1.1: Classes.

Second, a use case is a description of set of sequence of actions that a system performs that yields an observable result of value to particular actor.



Fig.4.1.2: Use Cases.

Third, a node is a physical element that exists at runtime and represents a computational resource, generally having at least some memory and often processing capability.

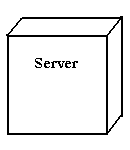


Fig.4.1.3: Nodes.

Behavioral things are the dynamic parts of UML models. The behavioral things used are:

**Interaction**

An interaction is a behavior that comprises a set of messages exchanged among a set of objects within a particular context to accomplish a specific purpose. An interaction involves a number of other elements, including messages, action sequences (the behavior invoked by a message, and links (the connection between objects).



Fig.4.1.4: Messages.

**Relationships in the UML**

There are four kinds of relationships in the UML

* Dependency
* Association
* Generalization
* Realization

A dependency is a semantic relationship between two things in which a change to one thing may affect the semantics of the other thing (the dependent thing).



Fig.4.1.5: Dependencies.

An association is a structural relationship that describes a set links, a link being a connection among objects. Aggregation is a special kind of association, representing a structural relationship between a whole and its parts.



Fig.4.1.6: Association.

A generalization is a specialization/ generalization relationship in which objects of the specialized element (the child) are substitutable for objects of the generalized element (the parent).



Fig.4.1.7: Generalization.

A realization is a semantic relationship between classifiers, where in one classifier specifies a contract that another classifier guarantees to carry out.



Fig.4.1.8: Realization.

**Sequence Diagrams**

UML sequence diagrams are used to represent the flow of messages, events and actions between the objects or components of a system. Time is represented in the vertical direction showing the sequence of interactions of the header elements, which are displayed horizontally at the top of the diagram.

Sequence Diagrams are used primarily to design, document and validate the architecture, interfaces and logic of the system by describing the sequence of actions that need to be performed to complete a task or scenario. UML sequence diagrams are useful design tools because they provide a dynamic view of the system behavior which can be difficult to extract from static diagrams or specifications.

**Actor**

Represents an external person or entity that interacts with the system

Sequence diagram actor element

Fig.4.1.9: User.

**Object**

It is an object in the system or one of its components.

**Sequence diagram object element**

Fig.4.1.10: Object.

**Unit**

Represents a subsystem, component, unit, or other logical entity in the system (may or may not be implemented by objects).

**Sequence diagram unit element**

Fig.4.1.11: Unit.

**Separator**

Represents an interface or boundary between subsystems, components or units (e.g., air interface, Internet, network)

**Sequence diagram separator element**

Fig.4.1.12: Separator.

**Group**

Groups related header elements into subsystems or components

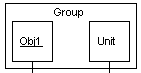
****

Fig.4.1.13: Group.

## Sequence Diagram Body Elements

**Action**

Represents an activity by an actor ,object or unit.

**Sequence diagram action element**

Fig.4.1.14: Action.

**Asynchronous Message**

An asynchronous message between header elements.

**Sequence diagram asynchronous message element**

Fig.4.1.15: Asynchronous message.

**Block**

A block representing a loop or conditional for a particular header element.

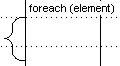
****

Fig.4.1.16: Block.

**Call Message**

A call (procedure) message between header elements.

**Sequence diagram call message element**

Fig.4.1.17: Call Message .

**Create Message**

A "create" message that creates a header element (represented by lifeline going from dashed to solid pattern).

**Sequence diagram create message element**

Fig.4.1.18: Create Message.

**Diagram Link**

Represents a portion of a diagram being treated as a functional block. Similar to a procedure or function call that abstracts functionality or details not shown at this level. Can optionally be linked to another diagram for elaboration.

**Sequence diagram diagram link element**

Fig.4.1.19: Link.

Else Block Represents an "else" block portion of a diagram block.

**Sequence diagram else block element**

Fig.4.1.20: Else.

**Message**

A simple message between header elements.

**Input Design**

The input design is the link between the information system and the user. It comprises the developing specification and procedures for data preparation and those steps are necessary to put transaction data in to a usable form for processing can be achieved by inspecting the computer to read data from a written or printed document or it can occur by having people keying the data directly into the system. The design of input focuses on controlling the amount of input required, controlling the errors, avoiding delay, avoiding extra steps and keeping the process simple. The input is designed in such a way so that it provides security and ease of use with retaining the privacy. Input Design considered the following things

1. What data should be given as input?
2. How the data should be arranged or coded?
3. The dialog to guide the operating personnel in providing input.
4. Methods for preparing input validations and steps to follow when error occur.

**Objectives**

1. Input Design is the process of converting a user-oriented description of the input into a computer-based system. This design is important to avoid errors in the data input process and show the correct direction to the management for getting correct information from the computerized system.

2. It is achieved by creating user-friendly screens for the data entry to handle large volume of data. The goal of designing input is to make data entry easier and to be free from errors. The data entry screen is designed in such a way that all the data manipulates can be performed. It also provides record viewing facilities.

3. When the data is entered it will check for its validity. Data can be entered with the help of screens. Appropriate messages are provided as when needed so that the user will not be in maize of instant. Thus the objective of input design is to create an input layout that is easy to follow.

**Output Design**

A quality output is one, which meets the requirements of the end user and presents the information clearly. In any system results of processing are communicated to the users and to other system through outputs. In output design it is determined how the information is to be displaced for immediate need and also the hard copy output. It is the most important and direct source information to the user. Efficient and intelligent output design improves the system’s relationship to help user decision-making.

1. Designing computer output should proceed in an organized, well thought out manner; the right output must be developed while ensuring that each output element is designed so that people will find the system can use easily and effectively. When analysis design computer output, they should Identify the specific output that is needed to meet the requirements.

2. Select methods for presenting information.

3. Create document, report, or other formats that contain information produced by the system.

The output form of an information system should accomplish one or more of the following objectives.

1. Convey information about past activities, current status or projections of the
2. Future.
3. Signal important events, opportunities, problems, or warnings.
4. Trigger an action.
5. Confirm an action.

**Dataflow Diagram**

1. The DFD is also called as bubble chart. It is a simple graphical formalism that can be used to represent a system in terms of input data to the system, various processing carried out on this data, and the output data is generated by this system.
2. The data flow diagram (DFD) is one of the most important modeling tools. It is used to model the system components. These components are the system process, the data used by the process, an external entity that interacts with the system and the information flows in the system.
3. DFD shows how the information moves through the system and how it is modified by a series of transformations. It is a graphical technique that depicts information flow and the transformations that are applied as data moves from input to output.

**4.2 UML Diagrams**

**Use Case Diagram**

Use case diagrams are usually referred to as behavior diagrams used to describe a set of actions (use cases) that some system or systems (subject) should or can perform in collaboration with one or more external users of the system (actors).

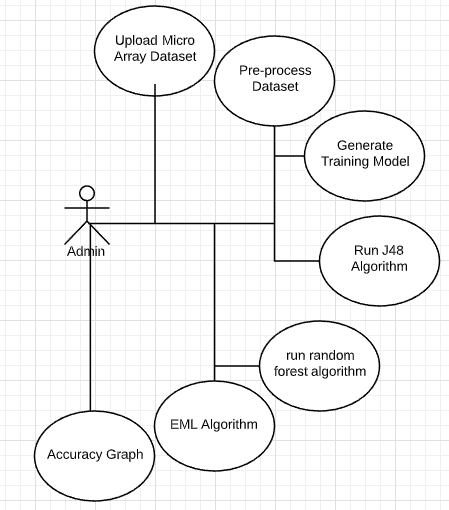


Fig.4.2.1: Use Case diagram.

**Class Diagram**

In software engineering, a class diagram in the Unified Modeling Language (UML) is a type of static structure diagram that describes the structure of a system by showing the system's classes, their attributes, operations (or methods), and the relationships among objects.

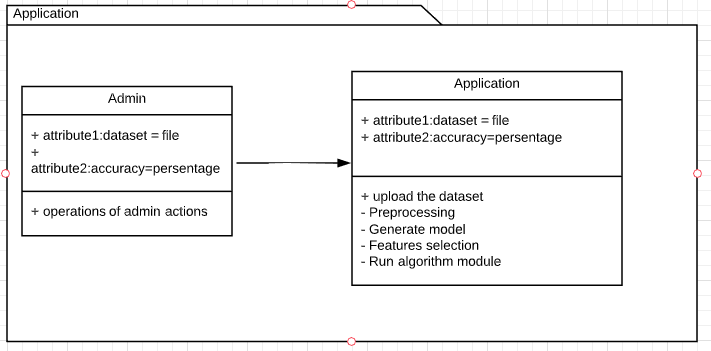


Fig.4.2.2: Class diagram.

**Sequence Diagram**

A Sequence diagram is an interaction diagram that shows how processes operate with one another and in what order. It is a construct of a Message Sequence Chart. A sequence diagram shows object interactions arranged in time sequence.

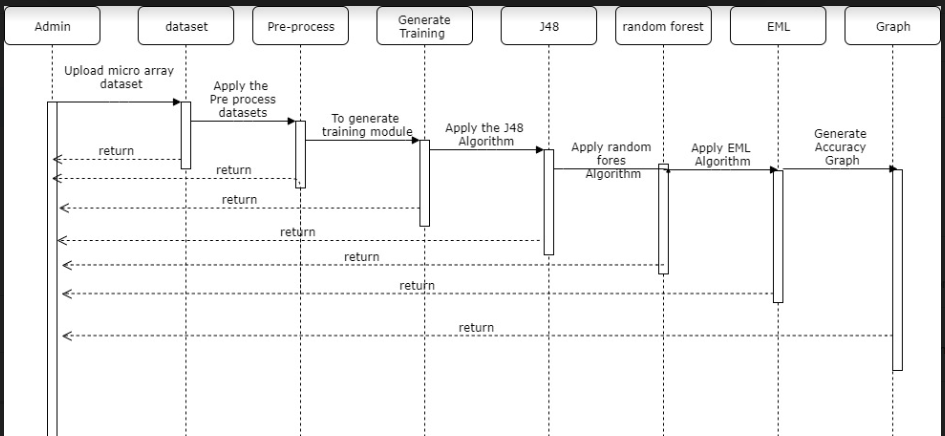


Fig.4.2.3: Sequence diagram.

**5. IMPLEMENTATION**

**Sample Code**

**6. TESTING**

**6.1 Introduction**

The purpose of testing is to discover errors. Testing is the process of trying to discover every conceivable fault or weakness in a work product. It provides a way to check the functionality of components, sub assemblies, assemblies and/or a finished product. It is the process of exercising software with the intent of ensuring that the Software system meets its requirements and user expectations and does not fail in an unacceptable manner. There are various types of test. Each test type addresses a specific testing requirement.

**Types of Tests**

**Unit testing**

Unit testing involves the design of test cases that validate that the internal program logic is functioning properly, and that program inputs produce valid outputs. All decision branches and internal code flow should be validated. It is the testing of individual software units of the application .it is done after the completion of an individual unit before integration. This is a structural testing, that relies on knowledge of its construction and is invasive. Unit tests perform basic tests at component level and test a specific business process, application, and/or system configuration. Unit tests ensure that each unique path of a business process performs accurately to the documented specifications and contains clearly defined inputs and expected results.

**Integration testing**

Integration tests are designed to test integrated software components to determine if they actually run as one program. Testing is event driven and is more concerned with the basic outcome of screens or fields. Integration tests demonstrate that although the components were individually satisfied, as shown by successfully unit testing, the combination of components is correct and consistent. Integration testing is specifically aimed at exposing the problems that arise from the combination of components.

**Functional testing**

Functional tests provide systematic demonstrations that functions tested are available as specified by the business and technical requirements, system documentation, and user manuals.

Functional testing is centered on the following items:

Valid Input : identified classes of valid input must be accepted

Invalid Input : identified classes of invalid input must be rejected

Functions : identified functions must be exercised

Output : identified classes of application outputs must be exercised

Systems/Procedures : interfacing systems or procedures must be invoked

Organization and preparation of functional tests is focused on requirements, key functions, or special test cases. In addition, systematic coverage pertaining to identify Business process flows; data fields, predefined processes, and successive processes must be considered for testing. Before functional testing is complete, additional tests are identified and the effective value of current tests is determined.

**System Testing**

System testing ensures that the entire integrated software system meets requirements. It tests a configuration to ensure known and predictable results. An example of system testing is the configuration oriented system integration test. System testing is based on process descriptions and flows, emphasizing pre-driven process links and integration points.

**White Box Testing**

White Box Testing is a testing in which in which the software tester has knowledge of the inner workings, structure and language of the software, or at least its purpose. It is used to test areas that cannot be reached from a black box level.

**Black Box Testing**

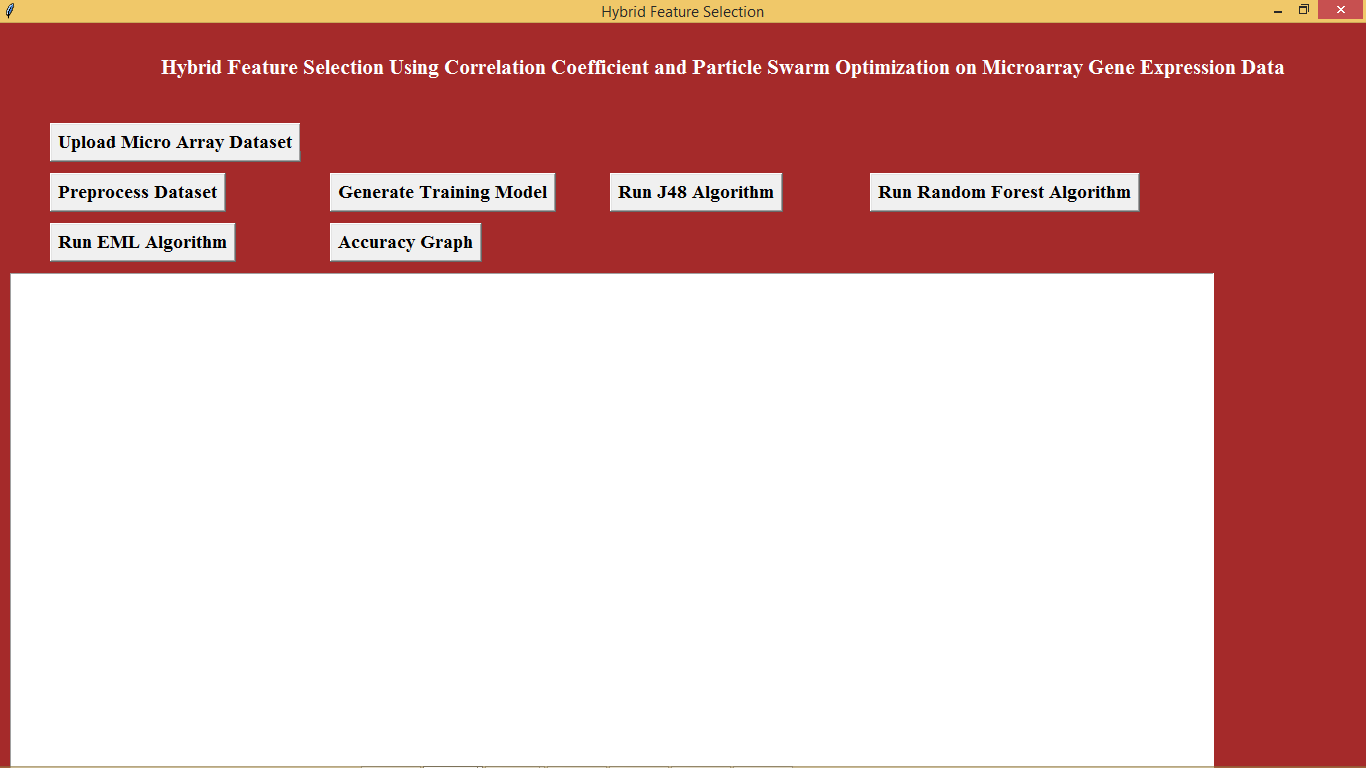
Black Box Testing is testing the software without any knowledge of the inner workings, structure or language of the module being tested. Black box tests, as most other kinds of tests, must be written from a definitive source document, such as specification or requirements document, such as specification or requirements document. It is a testing in which the software under test is treated, as a black box .you cannot “see” into it. The test provides inputs and responds to outputs without considering how the software works.

**6.2. Test cases**

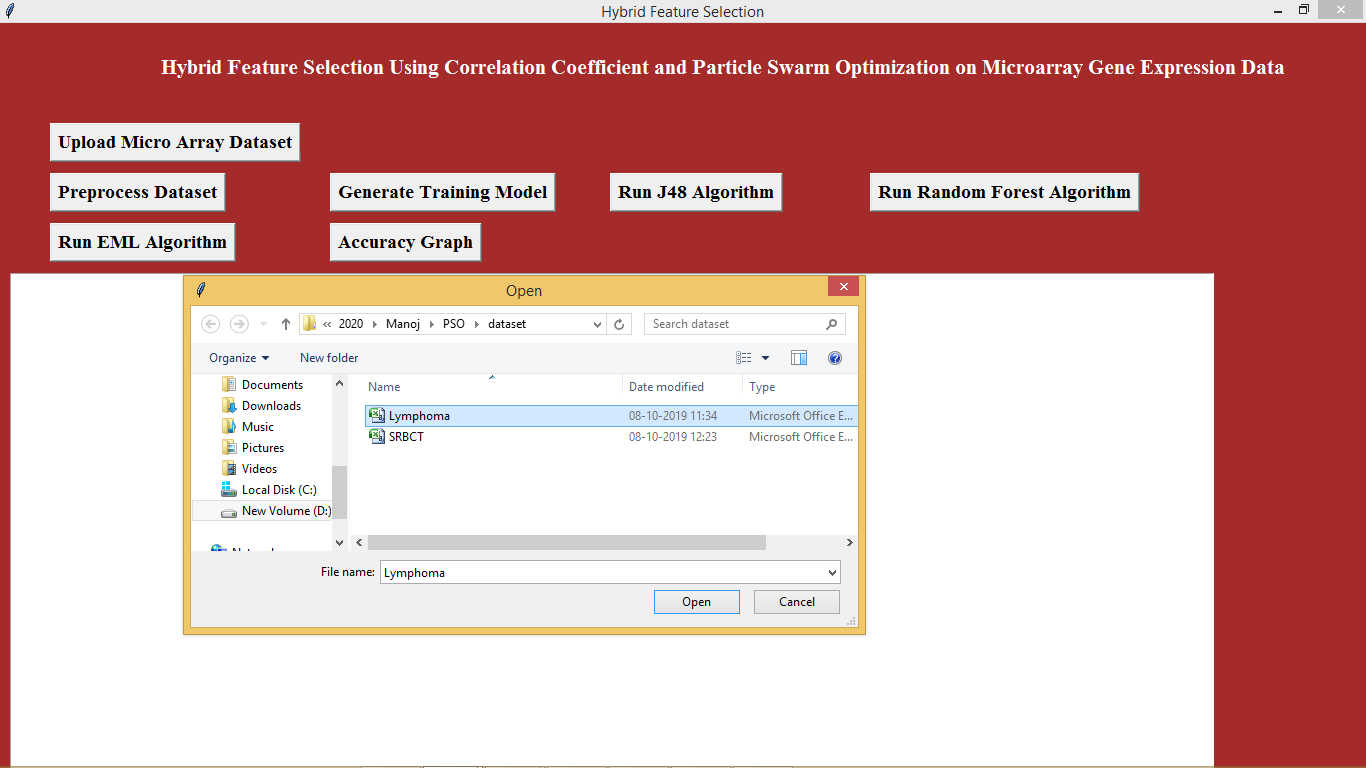
|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Test case ID** | **Test case Name** | **Functionality** | **Initial state** | **Input** | **Expected output** | **Status** |
| TC001 | Register | Register into the site | Homepage | Entering the necessary values to register | Registration complete | Success |
| TC002 | Login | Login into his/her account | Login page | Enter username, account no and password | Asks for OTP  Which is sent to mail | Success |
|  | Login | Login into his/her account | Login page | valid username, account no and password | Login page | success |
| TC003 | Enter OTP | Enter OTP which is sent to registered mail id | login page | Enter valid OTP | navigates to captcha page | Success |
|  | Enter OTP | Enter OTP which is sent to registered mail id | Login page | Entering OTP | valid OTP | Success |
| TC004 | Captcha Verification | Asks customer to enter captcha | login page | Verifies the Captcha | Customer Home Page | Success |
|  | Captcha Verification | Asks customer to enter captcha | Login page | Verifies the Captcha | Re-enter the Captcha | Success |
| TC005 | Transfer | Verifies Security answer | Customer page | Enter the same security answer at the time of registration | Sends OTP to mail | Success |
| TC006 | Mini Statement | Views all the performed transactions. | Customer page | Click on view mini statement | Shows all the incoming and outgoing transactions performed. | Success |
| TC007 | Balance | View Balance | Customer page | Click on view Balance | Displays the current balance | Success |
| TC008 | Fraud Details | View Fraud Details | Customer page | Click on fraud details | Displays fraud details(if any) | Success |
| TC009 | Requested Customers | Accepts or Rejects the requested customers | Branch Manager page | Accept the customer | Accepted successfully | Success |
| TC010 | Add Branch Managers | Adds branch manager | Bank Owner page | Enter the branch name, id and password | Branch Manager added successfully | Success |
| TC011 | Balance | View Balance | Branch Manager page | Click on view Balance | Displays the total balance  In the branch | Success |
| TC012 | Fraud Details | View Fraud Details | Bank Owner page | Click on fraud details | Displays fraud details(if any) | Success |

**7. SCREEN SHOTS**

Double click on ‘run.bat’ file to get below screen



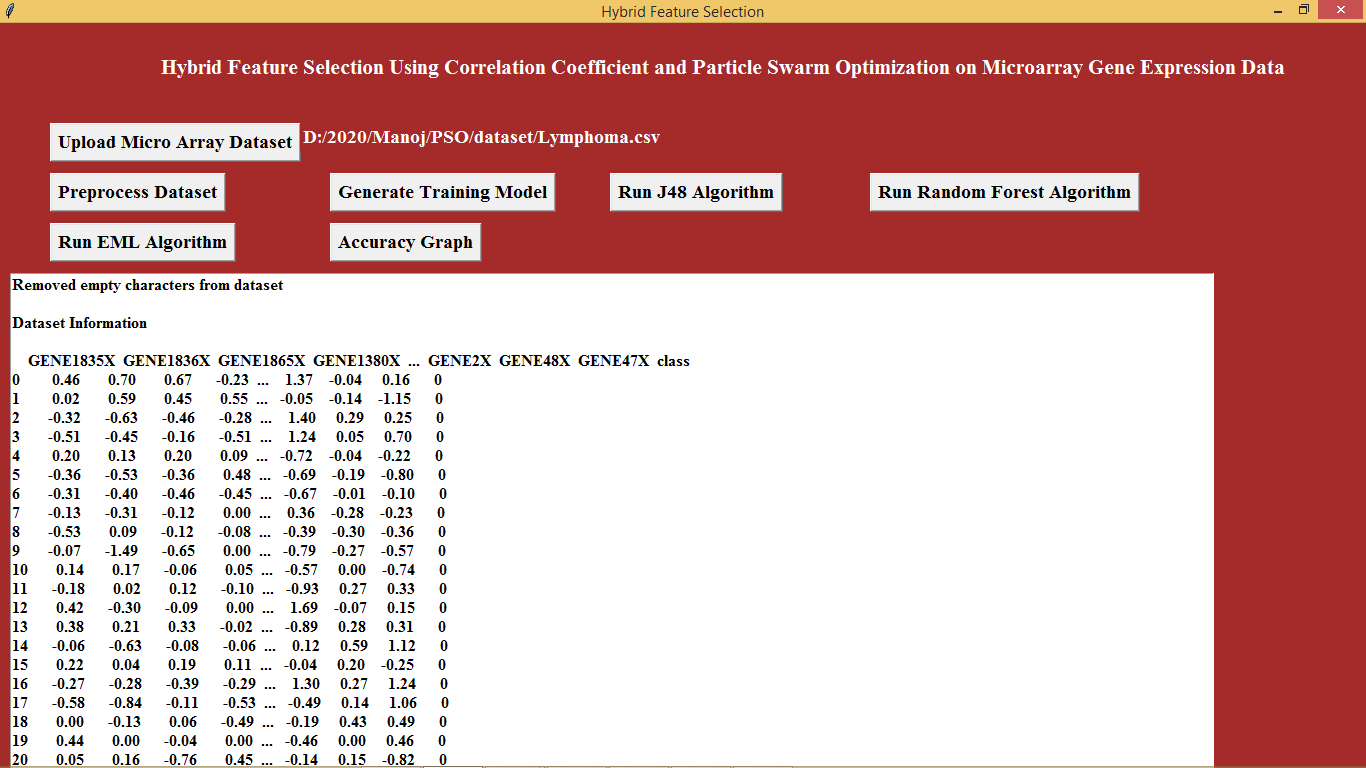
In above screen click on ‘Upload Micro Array Dataset’ button to upload genes dataset



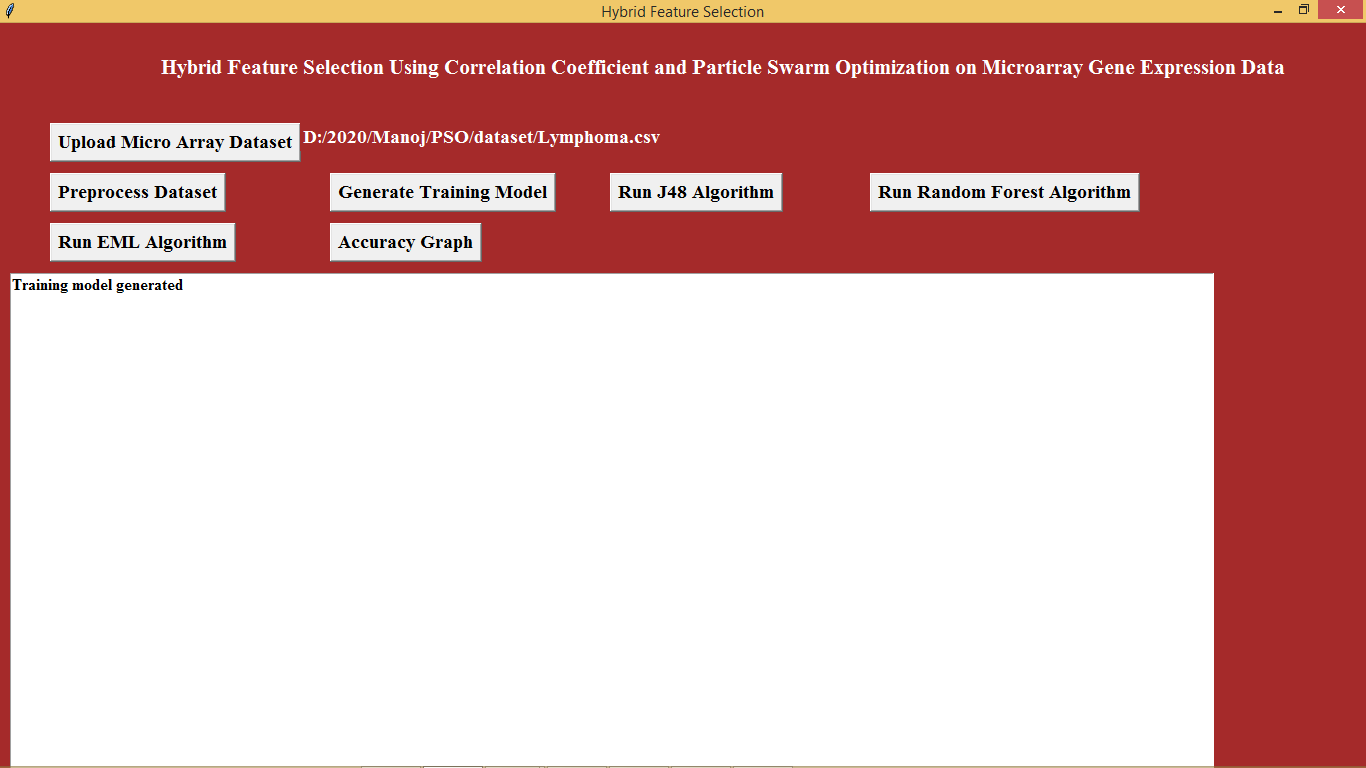
In above screen we are uploading Lymphoma dataset after upload dataset will get below screen



Now click on ‘Pre-process Dataset’ to remove empty strings



In above screen we are seeing come values from the dataset now click on ‘Generate Training Model’ to create train and test dataset.



After training model click on ‘Run J48 Algorithm’ to run j48 decision tree classifier



In above screen for J48 we got 42% accuracy now run random forest algorithm



In above screen for random forest accuracy is 57% now click on Run EML Algorithm



In above screen selected text we can see total 4027 attributes are there but after applying PSO we got 52 and in below lines we can see accuracy is 64% for ELM which is higher than other algorithms

Now click on “Accuracy Graph” button to view accuracy graph



In above graph we can see accuracy for J48, random forest and EML algorithms. In above graph x-axis represents algorithm name and y-axis represents accuracy

**8. CONCLUSION**

we have discussed the classifier accuracy of the proposed hybrid approach that combines the correlation coefficient with particle swarm optimization. This is compared with the traditional tree based classifiers like J48, Random Forest, Random Tree, Decision Stump and Genetic Algorithm as well. It is evident that the extreme learning machines classifier produces more or comparatively better accuracy than the other tree based classifiers available in literature. The proposed hybrid method that has higher potential in aiding further research in the area of feature selection simplified the process of gene selection which is evident from the experimental results. The proposed method significantly reduces the number of genes needed for classification and has also contributed to the improvement in classifier accuracy. The proposed method has greater scope of application to problems in other domains in future.

**10. BIBLIOGRAPHY**

**Text books Referred:**

UML : Unified Modeling Language by Iyar Jacobson, Ram Baugh, Grady Booch

Applying Use Case Driven Object Modeling with

UML by **Doug Rosenberg, Kendall Scott**

PYTHON:

**\**