

A Tool for Anticancer Peptide Prediction Using Feature Subspacing Ensemble



H.M. Fazlul Haque , Student Id: 011 141 169
Fariha Arifin , Student Id: 011 151 025
Fatima Islam Mouri , Student Id: 011 151 108
Kamrul Islam Tushar , Student Id: 011 141 139
Rana Kumar Ghosh , Student Id: 011 132 149

Department of Computer Science and Engineering
United International University

A thesis in the Department of Computer Science and Engineering presented in
partial fulfillment of the requirements for the Degree of
BSc in Computer Science & Engineering
September 2018

Declaration

We, declare that this thesis titled, Thesis Title and the work presented in it are our own. We confirm that:

- This work was done wholly or mainly while in candidature for a BSc degree at United International University.
- Where any part of this thesis has previously been submitted for a degree or any other qualification at United International University or any other institution, this has been clearly stated. Where we have consulted the published work of others, this is always clearly attributed.
- Where we have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work. We have acknowledged all main sources of help.
- Where the thesis is based on work done by ourselves jointly with others, we have made clear exactly what was done by others and what we have contributed ourselves.

fazlul haque

[H.M. Fazlul Haque,ID:011 141 169,Department:CSE]

fariha

[Fariha Arifin,ID:011 151 025,Department:CSE]

Mouri

[Fatima Islam Mouri,ID:011 151 108,Department:CSE]

Kamrul Islam Tushar

[Kamrul Islam Tushar,ID:011 141 139,Department:CSE]

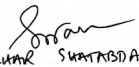
Rana kumar Ghosh

[Rana Kumar Ghosh,ID:011 132 149,Department:CSE]

Certificate

I do hereby declare that the research works embodied in this thesis/project entitled **A Tool for Anticancer Peptide Prediction Using Feature Subspacing Ensemble** is the outcome of an original work carried out by [H.M. Fazlul Haque student Id: 011 141 169, Fariha Arifin student Id: 011 151 025, Fatima Islam Mouri student Id: 011 151 108, Kamrul Islam Tushar student Id: 011 141 139 and Rana Kumar Ghosh student Id: 011 132 149] under my supervision.

I further certify that the dissertation meets the requirements and the standard for the degree of [BSc] in Computer Science and Engineering.


SWATIKUMAR SWATIKUMAR ASSOCIATE PROFESSOR

[Name and designation of Supervisor]

Abstract

Anticancer peptides have emerged as new effective method for treatment of cancer due to their high precision. As they are not toxic to normal cells they are preferred over traditional methods and therapy and hence discovery of anticancer peptides has found importance among biologists. However, the laboratory methods are still time consuming and cost-ineffective. In this project, we create a tool and propose a novel ensemble classification algorithm to predict anticancer property of peptides based on their primary amino acid sequences. Our ensemble divides the feature space into subspaces and thus each weak classifiers based on the subspaces and provides the final prediction using majority voting. We have tested the performance of the method on standard benchmark dataset. Our method is significantly better than most other state-of-the-art methods for anticancer peptide prediction.

Contents

| | |
|--|-------------|
| List of Figures | viii |
| List of Tables | ix |
| 1 Project Overview | 1 |
| 1.1 Introduction | 1 |
| 1.2 Basic Biological Entities | 2 |
| 1.2.1 Pros and Cons | 2 |
| 1.2.2 Tools | 2 |
| 1.3 Motivation | 2 |
| 1.4 Our Project | 3 |
| 1.4.1 Description of the Project | 3 |
| 1.4.2 Difficulties | 3 |
| 1.5 Methodology | 3 |
| 1.6 Summary | 3 |
| 2 Background | 5 |
| 2.1 Introduction | 5 |
| 2.2 Existing Work and Web on Biological Entities | 5 |
| 2.3 Comparison of Recent Workable Tools | 6 |
| 2.4 Software Requirement Specification (SRS) | 7 |
| 2.4.1 Overview | 7 |
| 2.4.2 Purpose | 7 |
| 2.4.3 Scope | 7 |
| 2.4.4 Goals | 7 |
| 2.4.5 Overall Description | 8 |
| 2.4.5.1 Users | 8 |
| 2.4.5.2 Functionality | 8 |
| 2.4.5.3 Platform | 8 |
| 2.4.5.4 Development and Responsibility | 8 |
| 2.4.5.5 Functional Requirements | 8 |
| 2.4.5.6 Requirements Specification | 9 |
| 2.4.5.7 Technical Process | 10 |

| | | |
|----------|--|-----------|
| 2.5 | Summary | 10 |
| 3 | Machine Learning Methodology | 11 |
| 3.1 | Introduction | 11 |
| 3.2 | Benchmark Dataset | 11 |
| 3.3 | Feature Collection | 12 |
| 3.3.1 | Monomer Composition | 12 |
| 3.3.2 | Dipeptide Composition | 12 |
| 3.3.3 | Tripeptide Composition | 13 |
| 3.3.4 | 1-gapped Di-mono Composition | 14 |
| 3.3.5 | 1-gapped Mono-Di Composition | 14 |
| 3.3.6 | Feature Collection Summary | 14 |
| 3.4 | Classification Algorithm | 15 |
| 3.5 | Used Algorithm | 15 |
| 3.6 | Parameter of Using Algorithms | 16 |
| 3.7 | Performance Evaluation | 16 |
| 3.8 | Summary | 17 |
| 4 | Experimental Results and Discussion | 18 |
| 4.1 | Introduction | 18 |
| 4.2 | Result of used Algorithm | 18 |
| 4.2.1 | Comparison with other methods | 19 |
| 4.3 | Summary | 20 |
| 5 | Standards and Impacts | 21 |
| 5.1 | Impacts | 21 |
| 5.1.1 | Introduction | 21 |
| 5.1.2 | Biological Impact | 21 |
| 5.1.3 | Machine Learning | 21 |
| 5.1.4 | Social Impact | 21 |
| 5.1.5 | Economic Impact | 21 |
| 5.1.6 | Ethical Impact | 22 |
| 5.2 | Standard | 22 |
| 5.2.1 | Introduction | 22 |
| 5.2.2 | Standards | 22 |
| 5.3 | Challenges | 22 |
| 5.4 | Summary | 22 |
| 6 | Web Application | 23 |
| 6.1 | Development | 23 |
| 6.2 | User manual | 23 |

CONTENTS

| | |
|---------------------------|-----------|
| 7 Conclusion | 28 |
| 7.1 Summary | 28 |
| 7.2 Limitation | 28 |
| 7.3 Future Work | 28 |
| Bibliography | 29 |

List of Figures

| | | |
|------|---|----|
| 1.1 | Central Dogma of Molecular Biology | 2 |
| 1.2 | Methodology | 4 |
| 2.1 | SPRINT-Mal | 6 |
| 2.2 | ACPred-FL | 6 |
| 2.3 | Use Case Diagram of SRS | 9 |
| 3.1 | Feature Collection | 12 |
| 3.2 | Monopeptide Composition | 13 |
| 3.3 | Dipeptide Composition | 13 |
| 3.4 | Tripeptide Composition | 14 |
| 3.5 | (a) 1-gapped Mono-Di Composition and (b) 1-gapped Di-mono Composition | 14 |
| 3.6 | Block diagram of the feature subsampling ensemble classifier. | 15 |
| 6.1 | Home Page | 24 |
| 6.2 | Read Me Page | 24 |
| 6.3 | Downloads Page | 24 |
| 6.4 | Contributors Page | 25 |
| 6.5 | Contributors Page | 25 |
| 6.6 | Server main Page | 25 |
| 6.7 | Server Page(with example data) | 26 |
| 6.8 | Server Page(result of example data) | 26 |
| 6.9 | Server Page(with other data) | 26 |
| 6.10 | Server Page(result of other data) | 27 |

List of Tables

| | | |
|-----|--|----|
| 1.1 | Tools | 2 |
| 2.1 | Comparison of Recent Workable Tools | 6 |
| 3.1 | Summary of the dataset. | 12 |
| 3.2 | Summary of features and feature subspacing. | 14 |
| 3.3 | Parameter of Algorithm | 16 |
| 4.1 | Performance of single classifiers on the dataset. | 18 |
| 4.2 | Performance of different ensemble classifiers on the dataset. | 19 |
| 4.3 | Performance comparison of different state-of-the-art predictors on the benchmark dataset. | 19 |

Chapter 1

Project Overview

Anticancer is a method which prevent the development of cancer. Anticancer is basically used for cancer treatment.

1.1 Introduction

Cancer, the emperor of all maladies [1], is the mostly deadly of all diseases and has been spreading as an epidemic for the last few decades. Due to advances in radiation and chemotherapy, nowadays cancer is being treated in most of the cases. However, most of these treatments come with side effects that damages normal cells. Thus anticancer peptides (ACP) have emerged as more effective means to treat cancer due to high precision [2]. ACPs are short sequences of amino acids, generally of length varying from 5 to 30 monomers. However, the discovery of anticancer peptides using *in vitro* methods is expensive and time consuming.

Prediction of many biological entities related to genomics, transcriptomics and proteomics have been formulated as supervised learning problem and hence many machine learning methods are applied to solve them in recent times [3–7]. Similar to those problems, the prediction of anticancer peptides can be formulated as a binary classification task where given an unknown peptide sequence the task of the predictor is to predict whether that is an anticancer peptide or not based on a dataset collected from already verified instances of anticancer peptides. Several methods in the literature are found that address the anticancer peptide prediction as a binary classification task [8–12].

Most of these methods are dependent on derived features (structural and evolutionary) that are computationally expensive to generate. Many ensemble methods like Adaboost [5] and Random Forest [4] are applied to solve these problems where single classifiers fails to produce good results. Random Forest classifier randomly selects features and builds ensemble of decision trees. On the other hand, Adaboost learns classifiers in an iterative manner by adjusting weights of the misclassified instances.

1.2 Basic Biological Entities

Biological entities means details about life and living beings. Every life and living being considered as individual organisms. For example two human may affect each other but their bodies are independent to each other. They have their own energy, unique set of genes and body metabolism. If we consider bacteria they are also considered as distinct organism, even if they are living inside us or inside our food.

In most organisms the genetic information are stored in DNA (Deoxyribonucleic acid). First, By dividing cells DNA double strands became single and cells have completely new genomes. This process is known replication. Then RNA (Ribonucleic acid) is transcribed from DNA, and the protein is translated from RNA. The cell life is described from amino acid, which contains proteins.

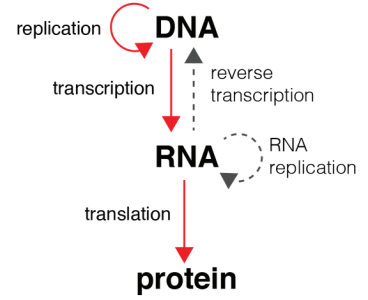


Figure 1.1: Central Dogma of Molecular Biology [13]

1.2.1 Pros and Cons

Biological entity plays a vital role in our life. Predicting attributes or functionality will be helpful for biologist and clinical researcher. But we need large, balanced data set and strong prediction algorithm to predict or the result may become harmful for human being.

1.2.2 Tools

Table following table shows the technical tools & programming languages which we will be used for this tool.

Table 1.1: Name of tools & languages.

| No. | Technical Tools | Programming Languages |
|-----|-----------------|-----------------------|
| 1. | Latex | Python |
| 2. | Anaconda | PHP |
| 3. | PyCharm | JavaScript |
| 4. | Notepad++ | Mysql |

1.3 Motivation

Biological entities are the source of the information of cell. Predicting the attribute or functionality is helpful for biologist and clinical researcher. Especially to find the cure

of cancer and diabetes or to find new diseases or help to find reasons of that diseases. By using machine learning algorithms, new features, and methods for detection and prediction will be invented. That will play a vital role on our social and economics especially human life.

1.4 Our Project

We are working with biological entities to predict attributes based on biological entities. And also develop a web based tool which will be able to predict results from given protein sequence.

1.4.1 Description of the Project

The idea behind the project is to develop a web based tool that relies on a sequence which will be able to detect biological entity. In this project, we propose a new ensemble classifier for anticancer peptide prediction. Our ensemble classifier divides the feature space into subspaces and weak classifiers are learned on the subspaces. The ensemble classifier then provides a majority voting classification based on the predictions given by each weak classifiers. We have used simple sequence based features and divided them into three groups. Each group was then used to be learnt using weak classifiers. Tested on a standard benchmark dataset for anticancer peptides, our method significantly outperforms most other state-of-the-art methods for anticancer peptide prediction. Our method is different from Random Forest and Adaboost in the way it divides the feature space and learns the single weak classifiers.

1.4.2 Difficulties

First challenge is to collect balanced dataset, after collecting data set, feature collection is also a challenging task. Because there's a need to find out more features which will help to predict attributes with more accuracy using different algorithms. At last development of the tool with created model will also be challenging for us.

1.5 Methodology

First we formulate the problem, then we collect data according to our problem formulation. Then we extract some feature from the data set and use some classifier algorithm to predict biological entities from the data, then we validate it and create a model for our website tool. At last we will finish our tool development using the created model.

1.6 Summary

In this chapter we know about the basics of biological entities, methods which we are following, pros and cons of predicting biological entities, Tools which we are using,

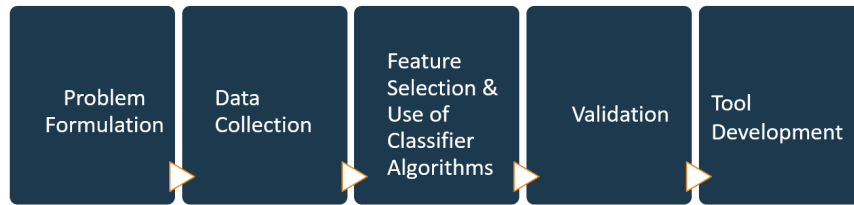


Figure 1.2: Methodology - Steps of methodology

our proposed project idea, difficulties which we may face to complete our project and motivation of making this project. The thesis is organized as : Chapter 2 provides related works, Chapter 3 presents methods, Chapter 5 discusses about the impact, Chapter 4 presents the result and Chapter 7 presents conclusion.

Chapter 2

Background

2.1 Introduction

In this section, we present a brief literature review on predictive tools for anticancer peptide prediction and attribute of other biological attributes and processes.

2.2 Existing Work and Web on Biological Entities

We have studied some already developed project tools, there are some works for predicting biological entities like The first computational method to formulate anticancer peptide prediction as a machine learning problem was by Tyagi et al. in [8]. In their work, Tyagi et al. used amino-acid composition and binary profiles as input vectors and Support Vector Machines as the operating engine for classification. However, Vijaykumar et al. [9] proved that on the dataset used by Tyagi et al. there were no significant difference among anticancer peptides and non-anticancer peptides in terms of amino-acid composition. Hence they proposed new effective features based on protein relatedness of amino acid distribution. They too used Support Vector Machines as classification algorithm.

In another work, Hajisharifi et al. [10] used Support Vector Machines with Chou's pseudo-amino acid composition as input features. Later, Chen et al. proposed iACP [11] that used g-gapped dipeptide composition as feature vector and Support Vector Machines as classification algorithm. In a very recent work, Wei et al. [12] also used Support Vector Machines and employed binary profiles, amino-acid composition, g-gapped dipeptide composition, composition-transition-distribution based features and used feature selection to provide a web server that produces the so far best results. They also introduced a new updated benchmark dataset for training purpose. One thing to note that, most of these methods use single classifier which is Support Vector Machine in this case and uses wide range of feature vectors. All of them provides web servers for use. Ensemble based classifiers based on feature subsampling are previously used in solving drug-target interaction prediction [5] and promoter identification problems [6]. Other ensemble based classifiers like Random Forest and Adaboost are also used for

2.3 Comparison of Recent Workable Tools

solving various related problems [4, 14].

SPRINT-Mal and AntiCancer peptide prediction(ACPred-FL) are example of prediction tool. Sprint-Mal predicts lysine malonylation of sites of proteins using sequence and predicted structural features. It can predict lysine malonylation of sites of proteins from the protein sequence of Human and mouse [15]. Anti cancer peptides works with two different dataset. One is to train the model and another one is to rest results. The training dataset is balanced. At predicts or classifies anti cancer peptides from protein sequence [16].

<http://sparks-lab.org/server/SPRINT-Mal/>
server.malab.cn/ACPred-FL/

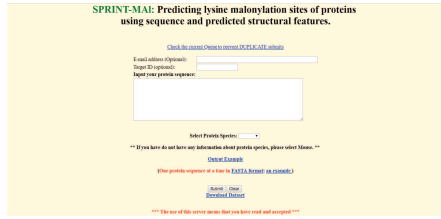


Figure 2.1: SPRINT-Mal
[15]

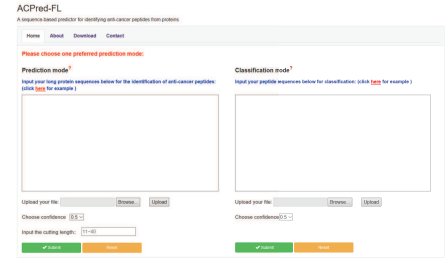


Figure 2.2: ACPred-FL
[16]

2.3 Comparison of Recent Workable Tools

We have found some difference, when we studied the tools. Some comparison between the tools are given below :

Table 2.1: Sprint-Mal & ACPred-FL

| No. | Sprint-Mal | ACPred-FL |
|-----|--|---|
| 1. | There is option to input as mail for tracking user | There is option to take any input for tracking user |
| 2. | There are some rule and regulation | There is no rule and regulation |
| 3. | It works on two data set so there is a choice option of prediction | It works only one data set so there is no option |
| 4. | There is only one option of prediction | There is two option for Identification and prediction |
| 5. | There is no option to set up confidence level | There is option of choose confidence level |

2.4 Software Requirement Specification (SRS)

This document lays out a project plan for the development of "A tool for Prediction of Anticancer Peptide Using Sub spacing Ensemble Classifier". The plan will include, but is not restricted to a summary of the system functionality, the scope of the project from the perspective of the "Anticancer Prediction Problem" team (me, my team members and my supervisor), scheduling and delivery estimates, project risks and how those risks will be mitigated, the process by which we will develop the project will be recorded throughout the project.

2.4.1 Overview

Anticancer Peptide Prediction using laboratory method is time consuming and expensive. With laboratory method prediction is fully dependent on laboratorians and financial capabilities of individuals. We aim to develop an application that would enable them to save their valuable time and money with nearly perfect prediction system.

2.4.2 Purpose

The purpose of SRS document is to present a detailed description of constraint of anticancer prediction. It will explain the purpose and features of our system used here and how the system will work, what type of algorithms will be used here and how this system will be operated. This document is intended for both stakeholders and developers of the system.

2.4.3 Scope

In our system anticancer peptides prediction can be hazard free and cheap. Biologists spends their valuable time and laboratory resources for the prediction of anticancer peptides. Those methods are highly expensive. Our system will help them to predict anticancer peptides absolutely free. Thus will help them to invent cure for cancer patients.

2.4.4 Goals

After the completion of this project we aspire to fulfill some specific goals. Some of the goals are listed below.

- Predict anticancer peptide using our developed tool
- Help biologists/clinical researcher in the field of finding cancer cure
- Reducing the cost of anticancer peptide prediction
- Help saving time of researchers/biologists

2.4.5 Overall Description

Here we have described the overall process elaborately to provide as much as information about our project we can in an organized way.

2.4.5.1 Users

There will be mainly two kind of users of our web tool who are Clinical Researcher and Biologists.

2.4.5.2 Functionality

It is important to understand how our web tool will function. Down below we have listed the functionality of our web tool.

- User would be able to predict anticancer peptides
- User will provide protein sequence and our tool will predict attribute according to that
- After prediction, predicted result will be shown to the user as a response message.

2.4.5.3 Platform

Our project will be launched as a Web-based application which will be accessed by a web browser which has an internet connection.

2.4.5.4 Development and Responsibility

We would be developing the software and we are responsible for the creation of related interfaces, server connections and support.

2.4.5.5 Functional Requirements

The functional requirement is describing the behavior of the system as it relates to the system's functionality. A finely designed system's usability is always satisfactory. This anticancer peptides prediction systems portal and each of its pages are very much easy to use. User will find it comfortable and some good set of directives are given to guide the user doing different set of activities. The category of user interfaces depends on the privileges given to the users. All the basic user functions that the user can perform are shown at the homepage and they are just one click away to access those functions. Making it user friendly is one of our prime goals and there will be a feedback screen for the user if there are any issues to address.

2.4 Software Requirement Specification (SRS)

Step-By-Step Description: Clinical researchers or biologists uses the web tool to predict anticancer peptides.

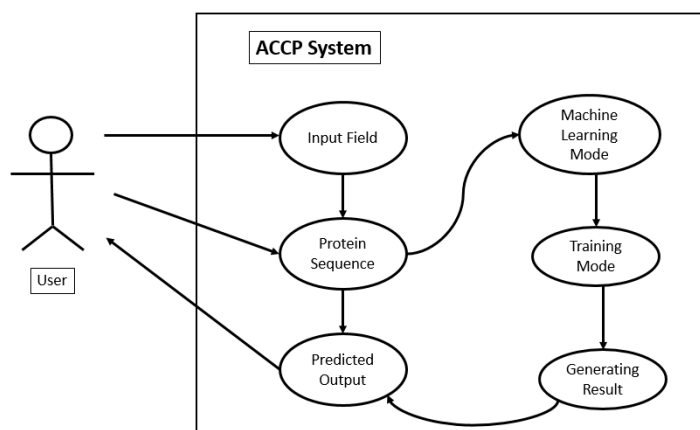


Figure 2.3: Use Case Diagram of SRS

Use case diagram 2.3 also given which help us to visualize the whole process.

- Researchers/Biologists/User inserts protein sequence in the text field
- That sequence is sent to the machine learning algorithm
- Machine learning algorithm predicts anticancer peptide based on previously saved trained model
- Prediction result is show to the user/researchers/biologists after predictions.

2.4.5.6 Requirements Specification

Let us discuss the non-functional requirements of our routine management system.

System Properties : The system properties are listed below which describes which tools were used to make this project and what are the requirements to run this tool.

- (a) A web based anticancer peptides prediction system runs on the internet
- (b) Runs on Linux server
- (c) HTML/CSS, Bootstrap, JavaScript based UI design
- (d) Python flask based functioning system.

Storage Requirements : The storage requirements for the project are described down below.

- (a) Different contents uploaded to the system are preserved
- (b) Minimum and maximum requirements of disk space are considered with care
- (c) Unnecessary and obsolete contents are removed to free the disk space.

Accessibility : Here we have stated the means by which someone can access this web tool.

- (a) This system is accessible from mobile/desktop/laptop devices
- (b) A stable internet connection is required.

Documentation : Proper documentations make it easier to use any system. We will provide visual directions and instructions to use this web tool.

- (a) User guidelines to use the system with screen-shots will be provided
- (b) User privileges and access management guidelines will be added in the documentation.

Availability : The system will be available to its specific users as stated below.

- (a) Any user can access this system 24/7
- (b) A minimum amount of downtime will be taken during maintenance period.

2.4.5.7 Technical Process

Following would be the languages we would like to use for the development of our application within the stipulated time period.

Front-end development : HTML, CSS, Bootstrap, JavaScript.

Back-end development :

- Programming Language : Python
- IDE : Anaconda, Notepad++, Spyder
- Virtual Machine : Vagrant Box.

2.5 Summary

In this chapter we know about existing works and SRS. Which inspired us and helped us to develop our project.

Chapter 3

Machine Learning Methodology

3.1 Introduction

This section provides the details of materials and methods used in this project. As suggested by K. C. Chou in [3], we have followed the famous five steps: i) selection of a proper dataset; ii) representation of peptides using feature extraction; iii) selection of a classification algorithm; iv) evaluation methodology and v) establishment of a web server as a prediction tool. In the testing phase, we generate a large number of features based on the peptide sequences taken from the dataset. After that, the feature space is divided into three parts and each are learnt using a weak classifier. The majority voting technique is used to provide with the final prediction decision from the weak classifiers. The ensemble is saved for the testing phase and used to predict anticancer property for any unknown peptide sequence. Rest of this section delineates the steps followed in this work.

3.2 Benchmark Dataset

In this project, we have used the training dataset constructed and presented by Wei et al. in [12]. One major problem in the previous datasets prior to this work was the imbalance with the negative instances. This newly proposed dataset was based on previously available databases of ACPs [8, 11, 17]. Among all the available ACPs, CD-HIT [18] was used to remove instances with 90% or more similarity. From that 250 positive and 250 negative samples were selected to construct the training dataset. Table 3.1 provides a summary pdf the dataset. Formally, the dataset \mathbb{S} can be shown as below:

$$\mathbb{S} = \mathbb{S}^+ \cup \mathbb{S}^-$$

Here, \mathbb{S}^+ and \mathbb{S}^- is the set of positive and negative instances respectively.

| Type | Number of Instances | Label |
|------------------------|---------------------|-------|
| Anticancer Peptide | 250 | +1 |
| Non Anticancer Peptide | 250 | -1 |
| Total | 500 | |

Table 3.1: Summary of the dataset.

3.3 Feature Collection

After selection of the dataset, now we have to choose a proper way to represent each peptide $P \in \mathbb{S}$. Each peptide instance in the dataset is replaced by a set of features and thus converted into a feature vector formally shown as below:

$$P = [f_1 f_2 f_3 \cdots f_n]$$

Note that, here each peptide sequences $P \in \mathbb{S}$ is a small string of amino-acids from the alphabet, Σ that contains 20 different symbols. Here, f_i is a feature extracted from the peptide instance. In this project, we generate different sequence based features that are easy to generate and also follows the common view that information flows from the genetic code and must be hidden in the peptide sequence itself. We provide a brief description of the features generated for our work



Figure 3.1: Feature Collection
[19]

3.3.1 Monomer Composition

Monomer composition is simply the normalized frequency of the count of different amino acid monomers in the given peptide sequence. Since there are only 20 different amino acids, the number of features is 20. Here monomer composition is represented by f_1 .

3.3.2 Dipeptide Composition

Dipeptide composition is the normalized frequency of the different dipeptides. Dipeptides are of 400 types and so is the number of features generated. Here monomer composition is represented by f_2 .

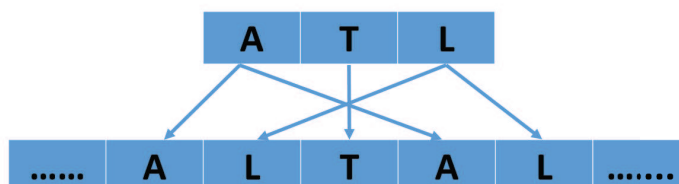


Figure 3.2: Monopeptide Composition

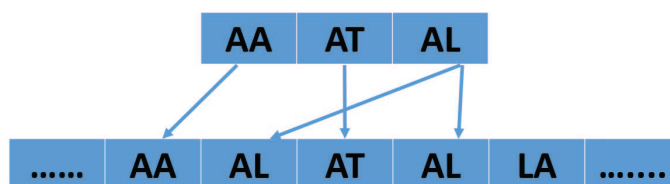


Figure 3.3: Dipeptide Composition

3.3.3 Tripeptide Composition

We have also considered tripeptide composition which is the normalized frequency of tripeptides. The total number of tripeptides are 8000. Here monomer composition is represented by f_3 .

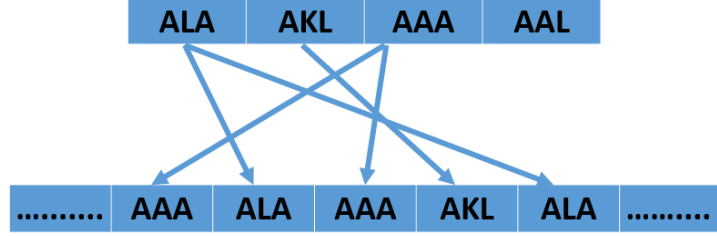


Figure 3.4: Tripeptide Composition

3.3.4 1-gapped Di-mono Composition

We have also considered the normalized frequency of tripeptides with a single gap in them. The particular patterns are in the form of XX_X . The number of features are 8000.

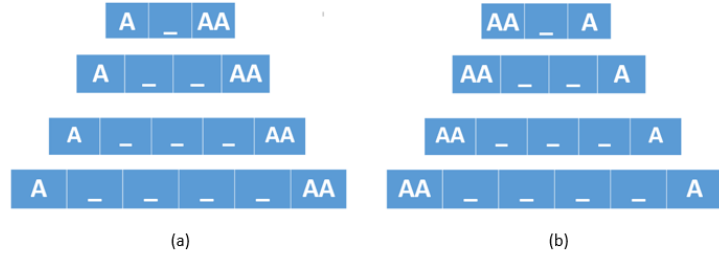


Figure 3.5: (a) 1-gapped Mono-Di Composition and (b) 1-gapped Di-mono Composition

3.3.5 1-gapped Mono-Di Composition

Similarly we have also considered normalized frequency of tripeptides with a single gap in them in the form of X_XX . The number of features are 8000.

3.3.6 Feature Collection Summary

We have divided the generated features into three non-overlapping groups. Summary of the features and subspacing are given in Table 3.2.

| Feature Subspace | Features | Type | Number of features |
|------------------|----------|------------------------------|--------------------|
| G_1 | f_1 | Monomer Composition | 20 |
| | f_2 | Dipeptide Composition | 400 |
| | f_3 | Tripeptide Composition | 8000 |
| G_2 | f_4 | 1-gapped Di-mono Composition | 8000 |
| G_3 | f_5 | 1-gapped Mono-Di Composition | 8000 |

Table 3.2: Summary of features and feature subspacing.

3.4 Classification Algorithm

We have used a feature subsampling based ensemble classification algorithm as a classifier in this work. Fig. 3.6 shows a block diagram for the classifier used. After the feature extraction phase, the total number of feature are divided into three groups as shown in the figure and also in Table 3.2. Each of these feature vectors are then zipped with the peptide labels to train individual classifiers. Since there are three parts, we will need three different classifiers for training.

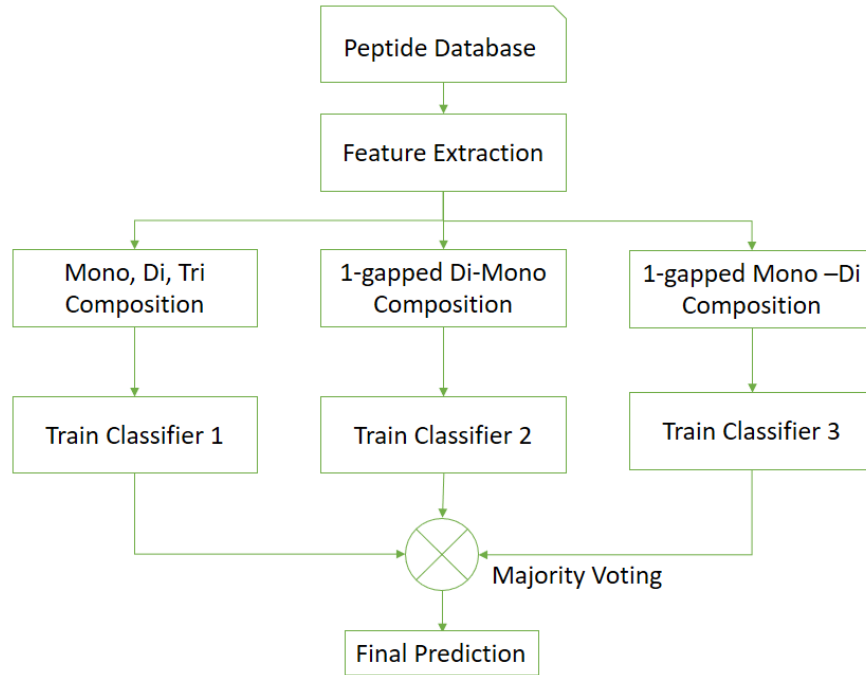


Figure 3.6: Block diagram of the feature subsampling ensemble classifier.

Note that three single classifiers used to train three subspaces of the feature space are not required to be of the same type. After the training each will learn their own model and predictions from each of these single classifiers will be aggregated using a simple majority voting for the final prediction. In the testing phase three models are to be saved in order to be used along with the voting scheme. Note that, we have not utilized any weighted voting here. Also the choice of dividing the feature space into three parts was also done arbitrarily. Weighted schemes, overlapped feature subsampling and number of subspaces could be left for future exploration.

3.5 Used Algorithm

We used K nearest neighbor[20], naive bayesian[21][22], support vector machine[23], decision tree[24] and logistic regression[25] for predicting the result.

3.6 Parameter of Using Algorithms

The table shows the parameter of algorithms.

Table 3.3: Parameter of Algorithm

| Classifier | Parameters | Types |
|------------------------|--|---------------------------|
| K-Nearest Neighbor | Number of Neighbors, Training set, test instances and defaults | - |
| Naïve Bayes | Training set, test instances and defaults | Bernoulli |
| Support Vector Machine | Training set, test instances and defaults | Support Vector Classifier |
| Decision Tree | Training set, test instances and defaults | - |
| Logistic Regression | Training set, test instances and defaults | - |

3.7 Performance Evaluation

In the literature of supervised learning methods for any prediction task, it is shown that selection of performance evaluation methods is very important [3]. For the sake of comparison, we have used the standard set of metrics and evaluation methods for the performance evaluation of our method as well. We have used 10-fold cross validation technique for the sampling on the train set. Here, the train set is divided into 10 folds or parts shuffled randomly and each time, a single fold is used as testing while the rest of the dataset is used for training.

We have used eight standard metrics for as performance measure: Accuracy (Acc), Sensitivity (Sn), Specificity (Spc), Precision, Recall, F1-Score, Matthew's Correlation Coefficient (MCC) and area under Receiver Operating Characteristic curve (auROC). For any binary classification task, we assume TP be the number of true positives or number of correctly predicted positive instances and TN be the number of true negatives or number of correctly predicted negative instances. Let FP and FN denote number of false positive and false negatives. They are respectively negative instances incorrectly predicted as positive and positive instances incorrectly predicted as negative. Now the metrics are defined as in below:

$$Acc = \frac{TP + TN}{TP + FP + FN + TN} \times 100\%$$

$$Sn = \frac{TP}{TP + FN} \times 100\%$$

$$Spc = \frac{TN}{TN + FP} \times 100\%$$

$$Precision = \frac{TP}{TP + FP}$$

$$Recall = \frac{TP}{TP + FN}$$

$$F1 - Score = \frac{2 \times precision \times recall}{precision + recall}$$

$$MCC = \frac{(TP \times TN) - (FN \times FP)}{\sqrt{(TP + FN) \times (TN + FP) \times (TP + FP) \times (TN + FN)}}$$

We have also considered auROC or area under Receiver Operating Characteristic (ROC) curve. ROC curve is the plot of TPR against FPR for different thresholds from a probabilistic classifier. Here note that each of these measures have the values in the range $[0,1]$ except MCC. Here 0 means a worst classifier and 1 means a perfect classifier. In the case of MCC the values are in the range of $[-1,+1]$.

3.8 Summary

In this chapter we have discussed about our collected features, our used algorithms.

Chapter 4

Experimental Results and Discussion

4.1 Introduction

In this section, we present the experimental results and analysis of our method. All the experiments were run five times and only the averages are reported as the classifiers used here are stochastic in nature. The classifiers were implemented using Scikit-learn library and Python 3.

4.2 Result of used Algorithm

The main goal of this work was to show that feature subspace ensembles works better in comparison to the single classifiers. Without changing the feature subspace as shown in the last section, we have used five different weak classifiers and their combinations in the ensemble. They are: Decision Tree (DT), Support Vector Machines (SVM), K-Nearest Neighbor (KNN), Logistic Regression (LR) and Naive Bayesian Classifier (NB). Note that we are performing 10-fold cross validation. Firstly, in Table 4.1, we present the results achieved by each of these single classifiers on the dataset. Here, we note that performance of all the classifiers are similar. However, KNN performs little worse to other single classifiers. Note that the full set of features were fed to these single classifiers.

| Classifier | Precision | Recall | F1 | Acc | MCC | Sn | Spc | AuROC |
|------------|-----------|--------|--------|-------|--------|-------|-------|--------|
| SVM | 0.8386 | 0.7152 | 0.7567 | 77.08 | 0.5686 | 80.61 | 74.41 | 0.8528 |
| NB | 0.6821 | 0.9440 | 0.7978 | 75.88 | 0.5614 | 68.84 | 91.32 | 0.7588 |
| KNN | 0.7431 | 0.3632 | 0.5007 | 64.50 | 0.3434 | 82.12 | 59.10 | 0.7650 |
| LR | 0.8605 | 0.7768 | 0.7869 | 78.92 | 0.5840 | 79.18 | 78.68 | 0.8657 |
| DT | 0.6760 | 0.7288 | 0.7302 | 73.36 | 0.4710 | 73.61 | 73.15 | 0.7336 |

Table 4.1: Performance of single classifiers on the dataset.

Next, we applied the ensemble classifier proposed in this project using each of these single classifiers member of the ensemble. The first five rows of Table 4.2 shows the experimental

4.2 Result of used Algorithm

| Classifier | Precision | Recall | F1 | Acc | MCC | Sn | Spc | AuROC |
|-------------|-----------|--------|---------------|--------------|---------------|-------|-------|--------|
| SVM+SVM+SVM | 0.9245 | 0.9512 | 0.8247 | 79.00 | 0.6296 | 70.89 | 97.24 | 0.9345 |
| NB+Nb+Nb | 0.8234 | 0.9376 | 0.8234 | 79.80 | 0.6230 | 73.29 | 91.39 | 0.8530 |
| KNN+KNN+KNN | 0.9001 | 0.384 | 0.8151 | 78.64 | 0.6036 | 71.94 | 91.22 | 0.9166 |
| LR+LR+LR | 0.8739 | 0.9220 | 0.8242 | 80.24 | 0.6083 | 74.35 | 89.96 | 0.8969 |
| DT+DT+DT | 0.8199 | 0.8288 | 0.8202 | 81.88 | 0.6376 | 81.01 | 82.46 | 0.8596 |
| SVM+Nb+LR | 0.8201 | 0.9736 | 0.7994 | 75.48 | 0.5687 | 67.73 | 94.10 | 0.8497 |
| Nb+LR+SVM | 0.8164 | 0.9736 | 0.7931 | 74.28 | 0.5527 | 66.80 | 95.14 | 0.8464 |
| LR+SVM+Nb | 0.8415 | 0.9776 | 0.8032 | 75.52 | 0.6539 | 67.81 | 95.99 | 0.8488 |
| DT+SVM+DT | 0.7735 | 0.8944 | 0.8058 | 78.60 | 0.5912 | 73.49 | 87.29 | 0.8206 |
| SVM+DT+DT | 0.7416 | 0.8928 | 0.8038 | 78.28 | 0.5854 | 73.18 | 87.04 | 0.7616 |
| LR+LR+DT | 0.8730 | 0.8816 | 0.8360 | 82.52 | 0.6603 | 79.51 | 86.52 | 0.8819 |
| SVM+LR+DT | 0.8186 | 0.9488 | 0.7969 | 77.48 | 0.5882 | 70.39 | 92.13 | 0.8379 |
| SVM+Nb+DT | 0.7530 | 0.9528 | 0.8098 | 77.44 | 0.6087 | 70.23 | 92.66 | 0.7895 |

Table 4.2: Performance of different ensemble classifiers on the dataset.

results. Note that, in each case accuracy and other measures were enhanced by the ensemble compared to the single classifier. The best performing ensemble was the decision tree ensemble.

We have also tried several mixtures of single classifiers in the ensemble. Eight different combinations of the five single classifiers were tried and the results are shown in the last eight rows of Table 4.2. Here note that a combination of logistic regression and decision tree is the best performing ensemble combination among all in terms of F1, MCC and Accuracy. Thus we select this combination as the best ensemble for our method.

4.2.1 Comparison with other methods

We have also compared the performance of the best ensemble combination to that of the previous methods in the literature. We have used four other methods for the sake of comparison: AntiCP [8], Hajisharifi’s method [10], iACP [11] and ACPred-FL [12]. Note that, we have not run their predictors, rather reported the 10-fold cross validation results as presented in the work by Wei et al. [12]. The results are shown in Table 4.3.

| Predictors | Sn | Spc | Acc | MCC | AuROC |
|----------------------|------|------|------|-------|-------|
| AntiCP_AAC | 67.2 | 86.0 | 76.6 | 0.542 | 0.84 |
| AntiCP_DC | 73.6 | 85.2 | 79.4 | 0.592 | 0.87 |
| Hajisharifi’s method | 68.0 | 86.4 | 77.2 | 0.553 | 0.84 |
| iACP | 72.8 | 86.0 | 79.4 | 0.593 | 0.86 |
| ACPred-FL | 84.8 | 98.0 | 91.4 | 0.835 | 0.94 |
| Our Method | 79.5 | 86.5 | 82.5 | 0.660 | 0.88 |

Table 4.3: Performance comparison of different state-of-the-art predictors on the benchmark dataset.

Note that, the performance of our method is superior to all other methods except ACPred-FL. Now, ACPred-FL have used a large number of features compared to ours and also did

perform feature selection that could have enhanced their performance.

4.3 Summary

In this chapter we know about results by using many algorithms and also know about the comparison of that used algorithms.

Chapter 5

Standards and Impacts

5.1 Impacts

5.1.1 Introduction

This section is contain the details about the impacts on various section of our project.

5.1.2 Biological Impact

This work has a large impact on economy. It increase of probability to find new cure, new formula can be generated from the attributes. It may help to find new & unknown diseases and also may help to find reason of a diseases.

5.1.3 Machine Learning

By doing this work new features can be generated in machine learning. Also generated new machine learning algorithm, new methods for detection & prediction and user friendliness.

5.1.4 Social Impact

Cancer is a disease which has no proper treatment. And available treatments doesn't always guarantees of full recovery due to the unavailability of proper cures/medicine. It is also expensive. Most of the people cannot bear the expense due to high expenses. Our developed tool may help biologist/clinical researchers to find cures. Which will help the cancer patients & save lives.

5.1.5 Economic Impact

This work has a large impact on economy. Cancer treatment is extremely costly and time consuming. As patient's needs to go through a lots of diagnosis. Our developed tool may help the biologist/clinical researcher to find a better cure which will helpful for treatment. Biologist spends lots of money and time for detecting anticancer peptide using laboratory method. Those laboratory methods are highly expensive which also causes cancer treatments being expensive. But our system will be fully free and thus will help clinical researchers/biologists detect anticancer peptides with absolutely less amount.

5.1.6 Ethical Impact

Anti Cancer peptide dataset is open to all for use, so we can use the datasets without needing any permission. For coding we are using Anaconda cloud, which is a free software and using spider and jupyter IDE which is also free. For documentation latex is being used. It is also a free software. All software's are free.

5.2 Standard

5.2.1 Introduction

Every project should maintain some standards. We are developing our project by following some standards in coding, UI, web development, ethics and software process.

5.2.2 Standards

- Coding Standard: PEP 8
- UI Standard: ISO 9241
- Web Development Standard: w3
- Ethics: IEEE/ ACM software engineering code of ethics
- Software Process: ISO/IEC 12207

As we are working on python so we have used PEP 8. ISO 9241 which is an international standard for UI, many country take it as their national standard so we took it as our UI standard. For web development we followed w3 because it is well organized platform. In ethical standard we have followed IEEE/ACM as they maintain professional standard. And lastly in software process we followed ISO/IEC 12207 because it gives us a common framework on the process of software life cycle.

5.3 Challenges

We have developed our tool using python for machine learning algorithm python flask package as API for parsing data from the input field to the machine learning algorithm. And we have also used HTML and CSS for the front end of the tool. The developed system is little bit different from other systems available right now. The system can't be run like other web tools just by coping the data on the server and hitting the web route. To run this system Linux server is must which allows installation of 3rd party application on which the system relies on. Linux server allows root level bash scripting which enables us to install dependencies for the system. But the issue with Linux server is it's not chip. But our university has helped us by providing us with a Linux server which helped us overcome the financial issue on buying a Linux server. To be a capstone project every project needs to solve some complex issue. We think previous problem was one of them. And we have solved that problem with the help of our university authority. So this satisfies the condition of a project being a capstone project.

5.4 Summary

In this chapter we know about our projects social, biological, machine learning, economic, ethical impacts and constraint of our project and standards which we are following.

Chapter 6

Web Application

6.1 Development

The tool has been developed using flask framework. Flask is a framework of python. With this framework we have used some method of flask like Send, Post, Request and bash scripting. When user submit protein sequence on the input filed flask request helper passes that data to the controller. And using that controller the data is sent to the machine learning algorithm with the help of Post method. After the machine learning processes that data the predicted result is sent to the view/web page with the help of Send method.

6.2 User manual

Visit <http://fseacp.pythonanywhere.com> for the home page. After visiting the home page there will be menu buttons for different pages. To detect anticancer peptides go to <http://fseacp.pythonanywhere.com/server>. Insert your anticancer peptide sequence you want to predict in the text area as fasta format. After inserting your anticancer peptide sequence you can press clear button to clear out the text area or you can click submit button to check results. After clicking the submit button the input data will be passed to machine learning algorithm using controller and after processing the data then a post method will be sent to the server page with predicted result. If anticancer peptides are detected it send accuracy, "Anti-Cancer Peptide Detected" if not then "No Anticancer Peptides are Detected". If you enter a random string that doesn't match to an anticancer peptide the send response will be "Invalid Sequence". Here we also add some screen-shots. To see the contributors who worked for this tool click on the contributors menu button on the top right corner. To download datasets you can go to <http://fseacp.pythonanywhere.com/downloads>. Here we also add read me section <http://fseacp.pythonanywhere.com/readme> to make our website more user friendly.

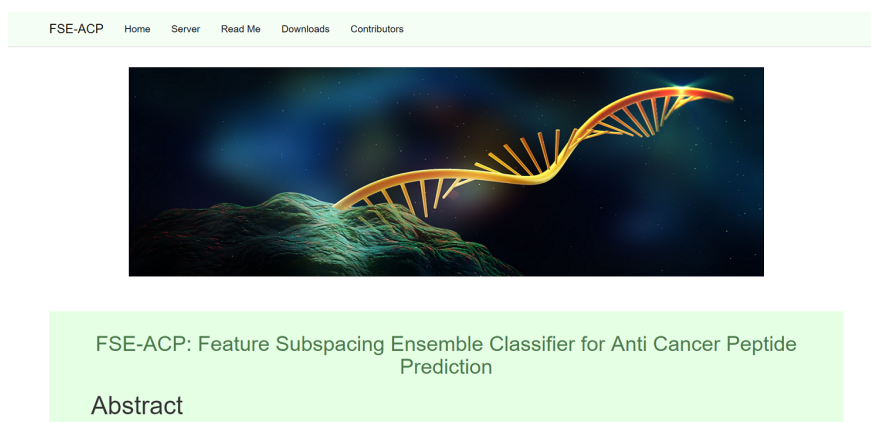


Figure 6.1: Home Page

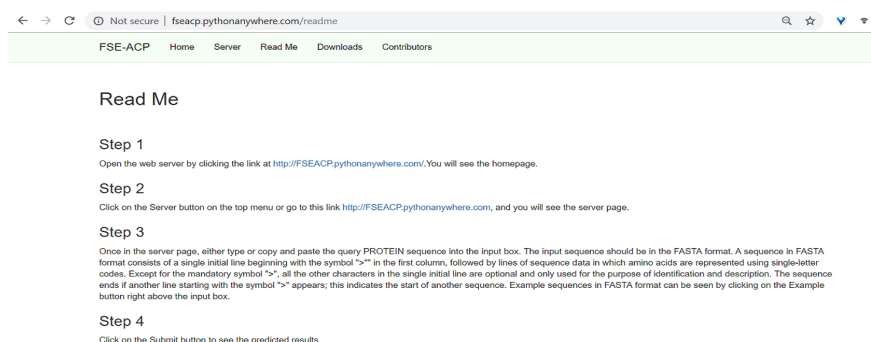


Figure 6.2: Read Me Page

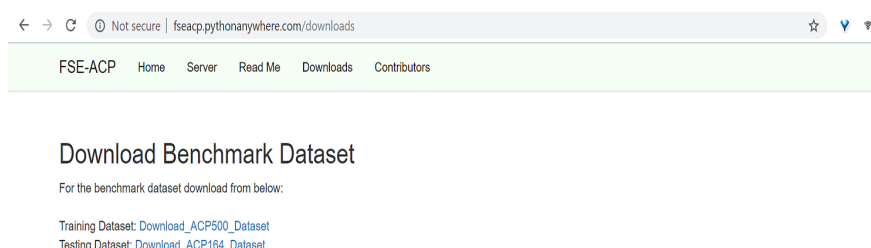


Figure 6.3: Downloads Page

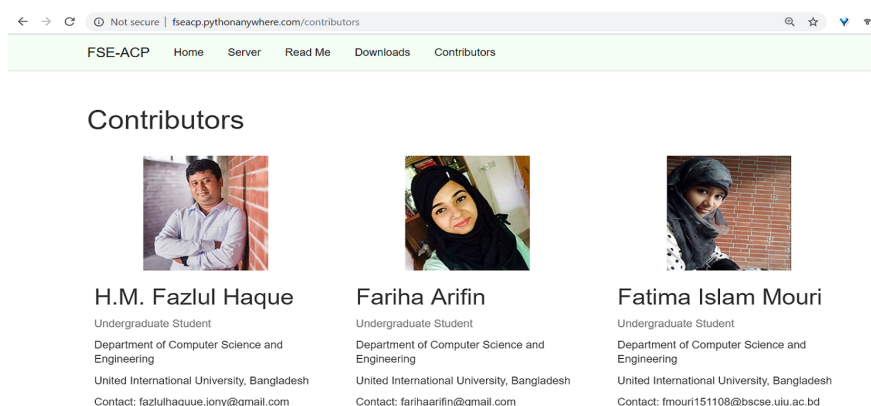


Figure 6.4: Contributors Page

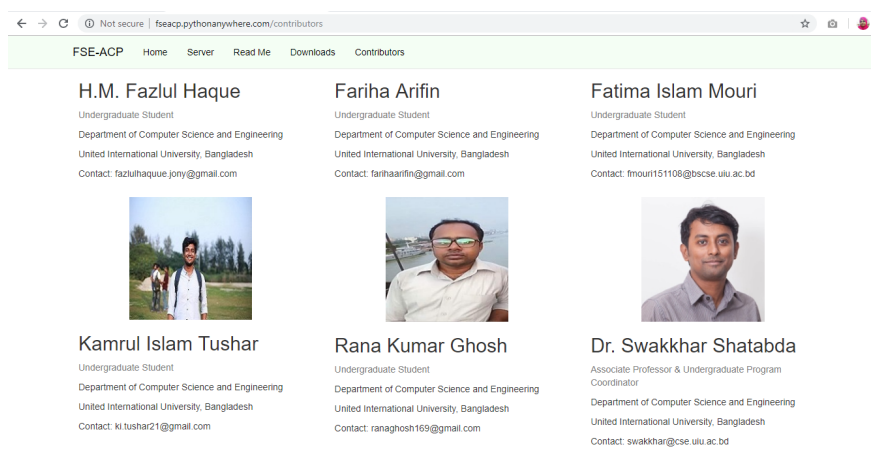


Figure 6.5: Contributors Page

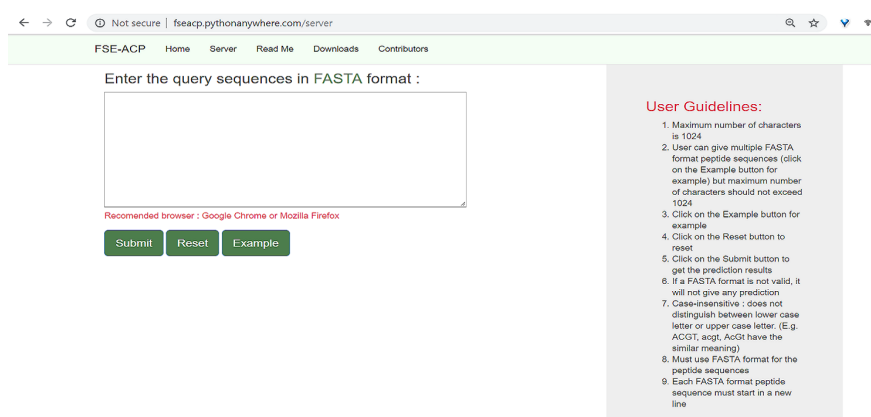


Figure 6.6: Server main Page

6.2 User manual

The screenshot shows the 'Server' page of the FSE-ACP application. At the top, there is a navigation bar with links: FSE-ACP, Home, Server, Read Me, Downloads, and Contributors. Below the navigation bar, the main content area is divided into two sections. On the left, there is a text input field labeled 'Enter the query sequences in FASTA format :'. Inside the input field, there are two sample sequences: '>Sample-1' followed by 'ZMNFNKLFFVFLVAVLCIGOSEAGWLKKGKIERVGHTRDATIONGVAQAAANVAATLKG' and '>Sample-2' followed by 'SRPSLDIDAGFEFGKEYKNGFIKQSEVORGGRLSPYFGINGGFRF'. Below the input field, there is a red text label 'Recommended browser : Google Chrome or Mozilla Firefox' and three green buttons: 'Submit', 'Reset', and 'Example'. On the right, there is a grey box titled 'User Guidelines:' containing a list of 9 rules for FASTA format sequences.

Enter the query sequences in FASTA format :

```
>Sample-1
ZMNFNKLFFVFLVAVLCIGOSEAGWLKKGKIERVGHTRDATIONGVAQAAANVAATLKG
>Sample-2
SRPSLDIDAGFEFGKEYKNGFIKQSEVORGGRLSPYFGINGGFRF
```

Recommended browser : Google Chrome or Mozilla Firefox

Submit Reset Example

User Guidelines:

1. Maximum number of characters is 1024
2. User can give multiple FASTA format peptide sequences (click on the Example button for example) but maximum number of characters should not exceed 1024
3. Click on the Example button for example
4. Click on the Reset button to reset
5. Click on the Submit button to get the prediction results
6. If a FASTA format is not valid, it will not give any prediction
7. Case-insensitive : does not distinguish between lower case letter or upper case letter. (E.g. ACGT, acgt, AcGt have the similar meaning)
8. Must use FASTA format for the peptide sequences
9. Each FASTA format peptide sequence must start in a new line

Figure 6.7: Server Page(with example data)

The screenshot shows the 'Send' page of the FSE-ACP application. At the top, there is a navigation bar with links: FSE-ACP, Home, Server, Read Me, Downloads, and Contributors. Below the navigation bar, the main content area has a heading 'All FASTAS are not valid' and a subtext 'Please go back and input valid fasta'. Below this, there is a table with two columns: 'Fasta' and 'Validity'. The table contains two rows of data.

| | Fasta | Validity |
|---|-----------|---------------|
| 1 | >Sample-1 | Invalid FASTA |
| 2 | >Sample-2 | Valid FASTA |

Figure 6.8: Server Page(result of example data)

The screenshot shows the 'Server' page of the FSE-ACP application, similar to Figure 6.7 but with different sample data. The input field contains two sample sequences: '>Sample-1' followed by 'MNFNKLFFVFLVAVLCIGOSEAGWLKKGKIERVGHTRDATIONGVAQAAANVAATLKG' and '>Sample-2' followed by 'SRPSLDIDAGFEFGKEYKNGFIKQSEVORGGRLSPYFGINGGFRF'. The rest of the page, including the navigation bar, recommended browser, buttons, and user guidelines, is identical to Figure 6.7.

Enter the query sequences in FASTA format :

```
>Sample-1
MNFNKLFFVFLVAVLCIGOSEAGWLKKGKIERVGHTRDATIONGVAQAAANVAATLKG
>Sample-2
SRPSLDIDAGFEFGKEYKNGFIKQSEVORGGRLSPYFGINGGFRF
```

Recommended browser : Google Chrome or Mozilla Firefox

Submit Reset Example

User Guidelines:

1. Maximum number of characters is 1024
2. User can give multiple FASTA format peptide sequences (click on the Example button for example) but maximum number of characters should not exceed 1024
3. Click on the Example button for example
4. Click on the Reset button to reset
5. Click on the Submit button to get the prediction results
6. If a FASTA format is not valid, it will not give any prediction
7. Case-insensitive : does not distinguish between lower case letter or upper case letter. (E.g. ACGT, acgt, AcGt have the similar meaning)
8. Must use FASTA format for the peptide sequences
9. Each FASTA format peptide sequence must start in a new line

Figure 6.9: Server Page(with other data)

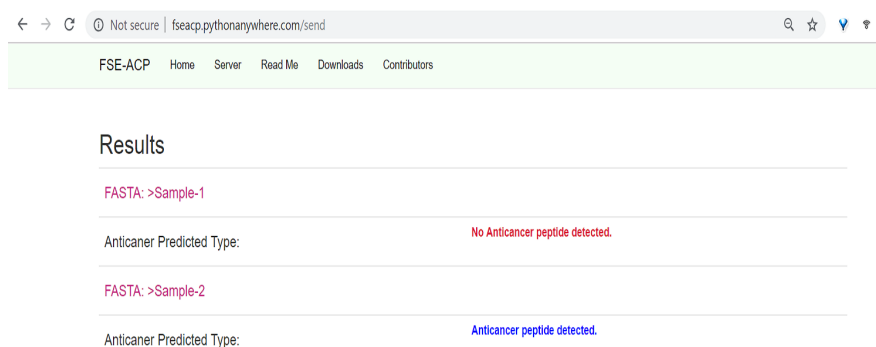


Figure 6.10: Server Page(result of other data)

Chapter 7

Conclusion

7.1 Summary

In this project, we present a novel ensemble based classification method and a web base tool based on feature subspacing and apply it for solving anticancer peptide prediction problem. On the standard benchmark dataset, our method produces significantly better results compared to single weak classifiers and outperforms most other state-of-the-art predictors. However, there are still a room to improve in terms of accuracy. We believe, more effective feature space and feature selection might result into more effective classifier.

7.2 Limitation

We have time constraints, that we may need more time to more feature selection to achieve the higher accuracy we want but we have time limitation to finish the work. We are using free software so there is no constraint on that part. Even for using dataset we do not need to take any permission as it is open for anyone to use.

7.3 Future Work

In future we will apply feature selection to increase our present accuracy and try to apply our model on independent dataset. We will try to find more features and try to apply this subspacing method on various types of datasets. We will increase tool's usability and also try to publish it as a journal.

Bibliography

- [1] S. Mukherjee, *The emperor of all maladies: a biography of cancer*. Simon and Schuster, 2010. 1
- [2] L. Otvos, “Peptide-based drug design: here and now,” in *Peptide-Based Drug Design*. Springer, 2008, pp. 1–8. 1
- [3] K.-C. Chou, “Some remarks on protein attribute prediction and pseudo amino acid composition,” *Journal of theoretical biology*, vol. 273, no. 1, pp. 236–247, 2011. 1, 11, 16
- [4] M. R. Jani, M. T. K. Mozlish, S. Ahmed, N. S. Tahniat, D. M. Farid, and S. Shatabda, “irecspot-ef: Effective sequence based features for recombination hotspot prediction,” *Computers in biology and medicine*, vol. 103, pp. 17–23, 2018. 1, 6
- [5] F. Rayhan, S. Ahmed, D. M. Farid, A. Dehzangi, and S. Shatabda, “Cfsboost: Cumulative feature subspace boosting for drug-target interaction prediction,” *Journal of Theoretical Biology*, 2018. 1, 5
- [6] M. S. Rahman, U. Aktar, M. R. Jani, and S. Shatabda, “ipromoter-fsen: Identification of bacterial $\sigma 70$ promoter sequences using feature subspace based ensemble classifier,” *Genomics*, 2018. 5
- [7] S. Y. Chowdhury, S. Shatabda, and A. Dehzangi, “Idnaprot-es: Identification of dna-binding proteins using evolutionary and structural features,” *Scientific Reports*, vol. 7, no. 1, p. 14938, 2017. 1
- [8] A. Tyagi, P. Kapoor, R. Kumar, K. Chaudhary, A. Gautam, and G. Raghava, “In silico models for designing and discovering novel anticancer peptides,” *Scientific reports*, vol. 3, p. 2984, 2013. 1, 5, 11, 19
- [9] S. Vijayakumar and P. Lakshmi, “Acpp: a web server for prediction and design of anti-cancer peptides,” *International Journal of Peptide Research and Therapeutics*, vol. 21, no. 1, pp. 99–106, 2015. 5
- [10] Z. Hajisharifi, M. Piryaiee, M. M. Beigi, M. Behbahani, and H. Mohabatkar, “Predicting anticancer peptides with chou’s pseudo amino acid composition and investigating their mutagenicity via ames test,” *Journal of Theoretical Biology*, vol. 341, pp. 34–40, 2014. 5, 19
- [11] W. Chen, H. Ding, P. Feng, H. Lin, and K.-C. Chou, “iacp: a sequence-based tool for identifying anticancer peptides,” *Oncotarget*, vol. 7, no. 13, p. 16895, 2016. 5, 11, 19

-
- [12] L. Wei, C. Zhou, H. Chen, J. Song, and R. Su, "Acpred-fl: a sequence-based predictor based on effective feature representation to improve the prediction of anti-cancer peptides," *Bioinformatics*, 2018. 1, 5, 11, 19
- [13] W. Commons, "File:centraldogma_nodetails.png — wikimedia commons, the free media repository," 2014, [Online; accessed 29-September-2018]. [Online]. Available: https://commons.wikimedia.org/w/index.php?title=File:Centraldogma_nodetails.png&oldid=140301972 2
- [14] M. S. Rahman, S. Shatabda, S. Saha, M. Kaykobad, and M. S. Rahman, "Dpp-pseaac: A dna-binding protein prediction model using chou's general pseaac," *Journal of theoretical biology*, vol. 452, pp. 22–34, 2018. 6
- [15] G. Taherzadeh, Y. Yang, H. Xu, Y. Xue, A. W.-C. Liew, and Y. Zhou, "Predicting lysine-malonylation sites of proteins using sequence and predicted structural features," *Journal of computational chemistry*, 2018. 6
- [16] L. Wei, C. Zhou, H. Chen, J. Song, and R. Su, "Acpred-fl: a sequence-based predictor based on effective feature representation to improve the prediction of anti-cancer peptides," *Bioinformatics*, 2018. 6
- [17] A. Tyagi, A. Tuknait, P. Anand, S. Gupta, M. Sharma, D. Mathur, A. Joshi, S. Singh, A. Gautam, and G. P. Raghava, "Cancerppd: a database of anticancer peptides and proteins," *Nucleic acids research*, vol. 43, no. D1, pp. D837–D843, 2014. 11
- [18] W. Li and A. Godzik, "Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences," *Bioinformatics*, vol. 22, no. 13, pp. 1658–1659, 2006. 11
- [19] W. Commons, "File:glycylglycine.png — wikimedia commons, the free media repository," 2017, [Online; accessed 29-September-2018]. [Online]. Available: <https://commons.wikimedia.org/w/index.php?title=File:Glycylglycine.png&oldid=260084342> 12
- [20] N. S. Altman, "An introduction to kernel and nearest-neighbor nonparametric regression," *The American Statistician*, vol. 46, no. 3, pp. 175–185, 1992. 15
- [21] S. J. Russell and P. Norvig, *Artificial intelligence: a modern approach*. Malaysia; Pearson Education Limited, 2016. 15
- [22] G. H. John and P. Langley, "Estimating continuous distributions in bayesian classifiers," in *Proceedings of the Eleventh conference on Uncertainty in artificial intelligence*. Morgan Kaufmann Publishers Inc., 1995, pp. 338–345. 15
- [23] C. Cortes and V. Vapnik, "Support-vector networks," *Machine learning*, vol. 20, no. 3, pp. 273–297, 1995. 15
- [24] S. R. Safavian and D. Landgrebe, "A survey of decision tree classifier methodology," *IEEE transactions on systems, man, and cybernetics*, vol. 21, no. 3, pp. 660–674, 1991. 15
- [25] D. W. Hosmer Jr, S. Lemeshow, and R. X. Sturdivant, *Applied logistic regression*. John Wiley & Sons, 2013, vol. 398. 15