A Machine Learning based approach for Subcellular Localization of Proteins using generic feature set

by

Paramita Basak Upama

MASTER OF SCIENCE

IN

INFORMATION AND COMMUNICATION TECHNOLOGY

Institute of Information and Communication Technology

BANGLADESH UNIVERSITY OF ENGINEERING AND TECHNOLOGY

August, 2018

The thesis titled "*A Machine Learning based approach for Subcellular Localization of Proteins using generic feature set*" submitted by Paramita Basak Upama, Roll No.: 1015312027, Session: October 2015, has been accepted as satisfactory in partial fulfillment of the requirement for the degree of Master of Science in Information and Communication Technology on 14th August, 2018.

BOARD OF EXAMINERS

1. Shahin 27.11.18

Dr. Shahin Ahkter — Chairman (Supervisor)
Assistant Professor
IICT, BUET

2.

Prof. Dr. Md. Saiful Islam — Member (Ex- officio)
Professor and Director
IICT, BUET

3.

Dr. Hossen Asiful Mustafa — Member
Assistant Professor
IICT, BUET

4. 29.11.18

Dr. Khondaker Abdullah Al Mamun — Member (External)
Professor
Department of Computer Science and Engineering,
United International University, Dhaka

CANDIDATE'S DECLARATION

It is hereby declared that this thesis or any part of it has not been submitted elsewhere for the award of any degree or diploma.

Paramita Basak Upama

Paramita Basak Upama

Student ID: 1015312027

I dedicate this thesis to my dearest parents and respected teachers for their love and support during this journey

TABLE OF CONTENTS

LIST OF TABLES AND FIGURES

Figures:

Tables:

Acknowledgement

I would like to express my deep and sincere gratitude to both of my supervisors Dr. Shahin Akhter, Assistant Professor, Institute of Information and Communication Technology, Bangladesh University of Engineering and Technology (BUET), Dhaka. I owe her for her constant supervision, encouragement and personal guidance during the progress of my thesis. Her in depth knowledge in machine learning techniques and logical ways of thinking have been helpful for the successful completion of this work. I am grateful to her for such cooperation.

I would like to thank the members of my thesis examination committee for their presence and patience in understanding my work. I warmly thank Prof. Dr. Md. Saiful Islam, Dr. Hossen Asiful Mustafa and Dr. Khondaker Abdullah Al Mamun for their valuable suggestions.

I am thankful to all my teachers, colleagues and classmates for their supports during the period of my thesis.

Finally, I owe my thanks to my parents and friends. Without their encouragement it would have been impossible for me to complete this thesis work.

Abstract

Protein subcellular localization is defined as predicting the functioning location of a given protein inside the cell. It is considered an important step towards protein function prediction and drug design. The task of protein subcellular localization from primary protein sequences is crucial for understanding genome regulation and functions. Support vector machine (SVM) based learning methods are shown to be effective for predicting protein subcellular and subnuclear localizations. Extraction of informative features cooperating with SVM plays an important role in designing an accurate system for predicting protein subnuclear localization. Proteins are large, complex molecules that are required for the structure, function, and regulation of a body‘s tissues and organs. Subcellular localization of proteins within a cell of the body is a mean of achieving functional diversity of protein. The process determines the access of protein‘s interacting partners and enables the integration of proteins into functional biological networks. To gain access to appropriate molecular interaction partners, protein must be at the right place at the right moment. Therefore, the process of protein subcellular localization is crucial for protein synthesis and drug discovery for a broad range of medical conditions and diseases.

The current study described here introduces a novel machine learning approach in Bioinformatics for classifying 361 protein sequences found inside a cell. The sequences were in string (text) format, and a set of characteristics were extracted out of them. The feature set includes 8 physicochemical properties of the protein found in 6 target locations of a cell. A support vector machine (SVM) based model has been developed to learn these properties of proteins and test the model on an independent dataset, considering the well-known application of SVM in this field. The algorithm developed during this work selects an optimal range of parameters of SVM and adopts feature selection for obtaining the best performance of the algorithm. The proposed algorithm achieved an average accuracy of 90% in classifying proteins on the target locations. It shows better performance compared to several similar algorithms presented in the literature. The technique proposed here can further be extended for protein sequences found in any part of the body.

CHAPTER 1
Introduction

1.1 Introducing the thesis work

Localization of proteins inside a cell is a budding topic of research in both of bioinformatics and machine learning fields [1]. Subcellular localization of proteins is the process of predicting protein sequence within a cell. The process is crucial for protein synthesis [2] and drug discovery for a broad range of medical conditions and diseases. A protein‘s function depends on its shape, and when protein formation goes awry, the resulting misshapen proteins cause problems that range from bad, when proteins neglect their important work, to ugly, when they form a sticky, clumpy mess inside of cells. Current research suggests that the world of proteins is far from pristine. Protein formation is an error-prone process, and mistakes along the way have been linked to a number of human diseases [3].

Several neurodegenerative diseases are the results of deleterious gain-of-function or dominant negative effects caused by aberrant subcellular localization of misfolded proteins [4]. Amyotrophic lateral sclerosis (ALS) is a fatal neurodegenerative disorder that affects mainly motor neurons. The mechanisms underlying the disease are not completely understood, but one – protein mislocalization – may play an important role in ALS. In fact, proteins that belong to the nucleus of cells but are wrongfully located outside of it, in the cytoplasm, are known to contribute to other neurodegenerative diseases [5]. Deregulation of the spatiotemporal dynamics of signaling proteins promotes tumorigenesis and metastasis. Wrong localization of several essential enzymes causes different metabolic diseases. Aberrantly localized proteins have also been linked to Alzheimer‘s disease, kidney stones and cancer. The changes from normal cells to cancer cells are primarily regulated by genome instability, which foster hallmark functions of cancer through multiple mechanisms including protein mislocalization. Mislocalization of these proteins, including oncoproteins, tumor suppressors, and other cancer-related proteins, can interfere with normal cellular function and cooperatively drive tumor development and metastasis [6].

Most of the popular methods for protein subcellular localization use images, or some use protein sequences with so many complex features as input data [7] [8] [9] [10]. But the sequence of a protein also contains valuable information about its characteristics [1] [11]. The method of predicting the location of any protein sequence in the divisions of a cell is important for researches in the field of Bioinformatics. So a simple test regarding subcellular localization of protein can reveal the reason of many diseases and the doctors can proceed to the treatments. Again, misfolded or altered protein sequences of a diseased body can be treated genetically to eliminate the risk of related diseases from the next generation. For this job, finding the actual location and functions of that protein sequence is important. Moreover, picking a protein sequence from a fossil, analyzing its functions and comparing them with the known species‘ protein functions can reveal new knowledge. For all these works a simple, easier and more accurate method will be better rather than using complicated, time consuming, not so accurate previous methods.

1.2 Present state of the problem and motivation for this thesis

At present, machine learning algorithms [9] [12] [13] [14] [15] are considered as popular methods for localizing proteins within a cell based on the physiochemical properties of protein sequences.Various sequence-based prediction methods have proved this fact by classifying proteins according to specific properties [16]. This has given rise to various relevant and frequently used bioinformatics tools, offered by a growing number of easily accessible web services [11] [17] [18].

Numerous techniques have been proposed in literature [7][8][9][10]. Among them, Hidden Markov model (HMM) [12], artificial neural networks [13], K-nearest neighbor [14] support vector machine (SVM) [19][15][20][21] and etc. where several variants of SVM (e.g., linear [19] or non-linear [20][15]) are mostly used in the literature. Algorithms using SVM considers a line of separation between objects of different classes which are known as hyper planes. The technique can be modified for nonlinear classification problem as well by introducing the kernel functions to the maximum-margin hyper planes in a transformed high dimensional feature space. Such tasks are required for maximizing decision boundaries to categorize objects from different classes which are nonlinear in the original input space, and SVM is considered as well suited for this task.

Employing the kernel tricks for SVM in a multi-class classification system, the study in [19] achieved 65% accuracy for an independent set compared to the 16.7% accuracy of random prediction. Performance of SVM based algorithm can be further improved by incorporating discriminative feature set to the classifier. For instance, the use of local evolutionary-based features together with SVM improved the accuracy to 86% [20]. The researchers in [15] proposed an evolutionary support vector machine (ESVM) based approach with automatic selection of features for subcellular localization of proteins where the accuracies were 56.37% and 72.82% respectively for two different dataset. Another group of researchers in [22] considered a set of SVMs, and trained it to predict the subcellular location of a given protein based on its amino acid, amino acid pair, and gapped amino acid pair compositions. The final accuracy is calculated to be 78-79%.

The high performance of the computational prediction based methods may arise from the use of a large number of complex feature set and more complex training model. However, considering the inclusion of large number of features [20][15][22] and complexity of the algorithms stated above [19][20][21][22] with respect to the classification accuracy, it requires further improvement in classifier design and feature selection for better performance in labeling protein classes.

Moreover, a number of location prediction algorithms have been developed based on amino acid compositions or on the N-terminal characteristics (signal peptides) of proteins. However, such approaches lead to a loss of contextual information. Moreover, where information about the physicochemical properties of amino acids has been used, the methods employed to exploit that information are less than optimal and could use the information more effectively [23].

Some of these works have been enlisted in table 1.1 along with their performance outcomes and problems.

Table 1.1: Some existing work along with their performance outcomes showing present state of the problem of concern

	Features Type	Classifier Type	Accuracy	Problems
1	K-peptide encoding vectors [19]	SVM with newly proposed kernel	50%	Checking sequence similarity only
2	Combination of physicochemical features [15]	SVM with inheritable genetic algorithm	56.37%	Too much features, complex computations
3	Different amino acid compositions [22]	Several SVMs	78-79%	Using only amino acid composition
4	Sequence-based occurrences [20]	SVM	79-86%	Complex feature extraction
5	GO termsin [24]	Novel genetic algorithm with SVM	90.06%	Limitations of accessing GO terms

	Features Type	Classifier Type	Accuracy	Problems
6	GO terms, amino acid composition [25]	SVM	81.1%	Limitations of accessing GO terms
7	Image [26]	SVM	81.02%	Loss of information in image, Complexity of image processing
8	Different amino acid compositions [27]	KNN	64-99%	Result completely depends on dataset
9	Image [28]	ensemble classifier by combining BR and CC classifiers	77.83%	Loss of information in image, Complexity of image processing

Hence this thesis was to develop an algorithm which should be simple and can be implemented with minimal calculations compared to the existing algorithms. This algorithm will include easily extractable features of protein sequences- which are easy to extract, easy to use for computation and results in a less-complex classification algorithm that shows increased accuracy compared to algorithms stated above.

1.3 Objectives

Motivated by the present state of machine learning based subcellular localization of proteins, the objective of this thesis work has become developing an algorithm for better localization of proteins inside a cell. To fulfill this objective, the following aims have been considered:

a) To identify a suitable feature set for subcellular localization of protein sequences found within different areas of a cell

b) To design a simple and efficient classification algorithm for identifying proteins in target locations of a cell based on their features

c) To compare the algorithm with existing works in the literature, so that the proposed method gets effectively accepted in its field

1.4 Contribution to the knowledge

This thesis work contains some novel steps in the proposed algorithm. These are:

a) Working with a new set of features, which can be easily extracted from protein sequences

b) Introducing some novel features to be used for the first time as predictors in subcellular localization of proteins

c) Developing a classification algorithm for predicting subcellular localization of proteins, which provides better performance compared to the established ones

1.5 Organization of the thesis

The rest of this thesis is organized as follows:

In Chapter 2, a brief overview of protein molecules and their functions, surveys on existing works and related literature review are carried out.

Chapter 3 describes the data collection, feature extraction, measures for feature selection and final feature set construction. This research considered standard protein dataset available in the web for 6 different locations inside the cell. A set of generic features was derived from these protein sequences for ease of feature computation and simplicity of the method.

Chapter 4 presents the steps of developing the proposed algorithm using the final dataset. An SVM based classification algorithm has been developed considering its wide range of applicability and superior performances in subcellular localization of proteins compared to other algorithms in the literature.

Chapter 5 presents the experimental results of the algorithm stated in the previous chapters. The performance of the algorithm (e.g., sensitivity, specificity and accuracy) was evaluated on independent training and testing dataset and was compared with existing similar algorithms.

Chapter 6 makes a brief discussion on the work and then concludes the thesis with future vision.

CHAPTER 2

Literature Review

2.1 Machine learning

Machine learning, now-a-days, has been popular in various fields of computer science [9] [12] [13] [14] [15]. Because of new computing technologies, machine learning today is not like machine learning of the past. It was born from pattern recognition and the theory that computers can learn without being programmed to perform specific tasks; researchers interested in artificial intelligence wanted to see if computers could learn from data.

2.1.1 Emergence of machine learning

The iterative aspect of machine learning is important because as models are exposed to new data, they are able to adapt independently. They learn from previous computations to produce reliable, repeatable decisions and results. It‘s a science that‘s not new – but one that has gained fresh momentum. Figure 2.1 below is a block diagram of the working procedure of machine learning.

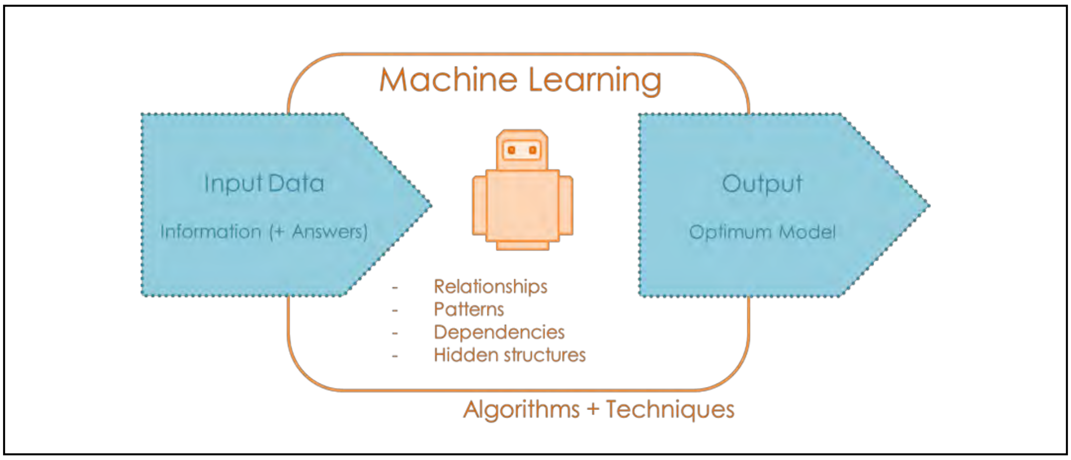


Figure 2.1: The technique of machine learning (Source:[2])

2.1.2 Applications and usefulness of machine learning

While many machine learning algorithms have been around for a long time, the ability to automatically apply complex mathematical calculations to big data – over and over, faster and faster – is a recent development: [9] [12] [13] [14] [15] [29]. Here are a few application areas of machine learning people are familiar with:

1. ***Data Security*** - predict malware
2. ***Personal Security*** - spot things human screeners might miss
3. ***Financial Trading*** - predict what the stock markets will do
4. ***Healthcare*** - spotted cancers a year before they were officially diagnosed
5. ***Marketing Personalization*** - lead consumers reliably towards a sale
6. ***Fraud Detection*** - PayPal is using machine learning to fight money laundering

Resurging interest in machine learning is due to the same factors that have made data mining and Bayesian analysis more popular than ever. Things like growing volumes and varieties of available data, computational processing that is cheaper and more powerful, and affordable data storage.

All of these things mean it's possible to quickly and automatically produce models that can analyze bigger, more complex data and deliver faster, more accurate results, even on a very large scale. And by building precise models, an organization has a better chance of identifying profitable opportunities or avoiding unknown risks.

Most industries working with large amounts of data have recognized the value of machine learning technology. By gleaning insights from this data, often in real time, organizations are able to work more efficiently or gain an advantage over competitors. For example: Financial services, Health care, Oil and gas, Government, Marketing and sales, Transportation etc.

Some of the fields who use machine learning applications are shown in the figure 2.2 below.

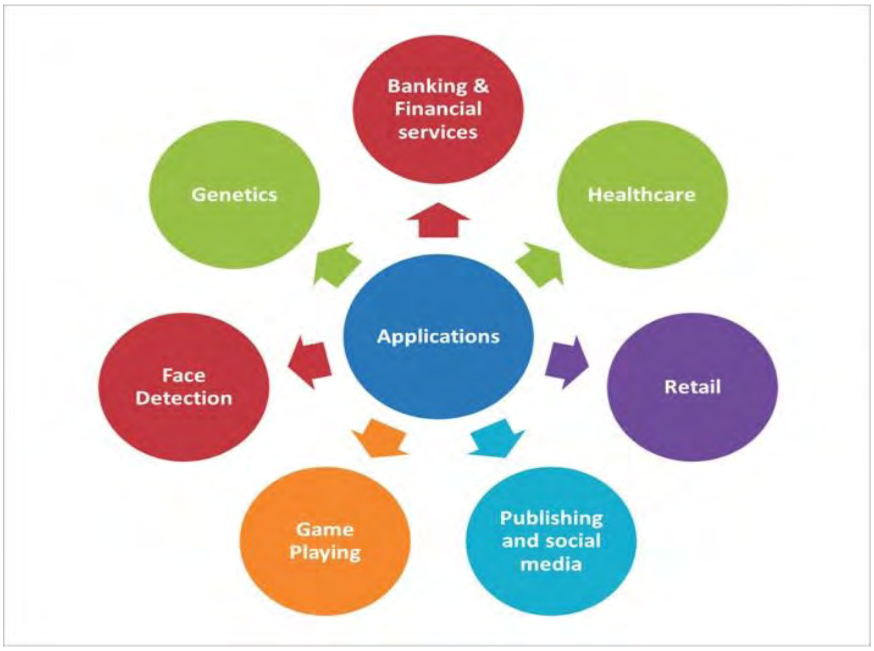


Figure 2.2:Application areas of machine learning (Source: [29])

2.1.3 Types of machine learning

Some popular types of machine learning algorithms are:

a. ***Supervised learning*** algorithms [30] [31] [32] are trained using labeled examples, such as an input where the desired output is known (figure 2.3). For example, a piece of equipment could have data points labeled either ―F” (failed) or ―R” (runs). The learning algorithm receives a set of inputs along with the corresponding correct outputs, and the algorithm learns by comparing its actual output with correct outputs to find errors. It then modifies the model accordingly. Through methods like classification, regression, prediction and gradient boosting, supervised learning uses patterns to predict the values of the label on additional unlabeled data.

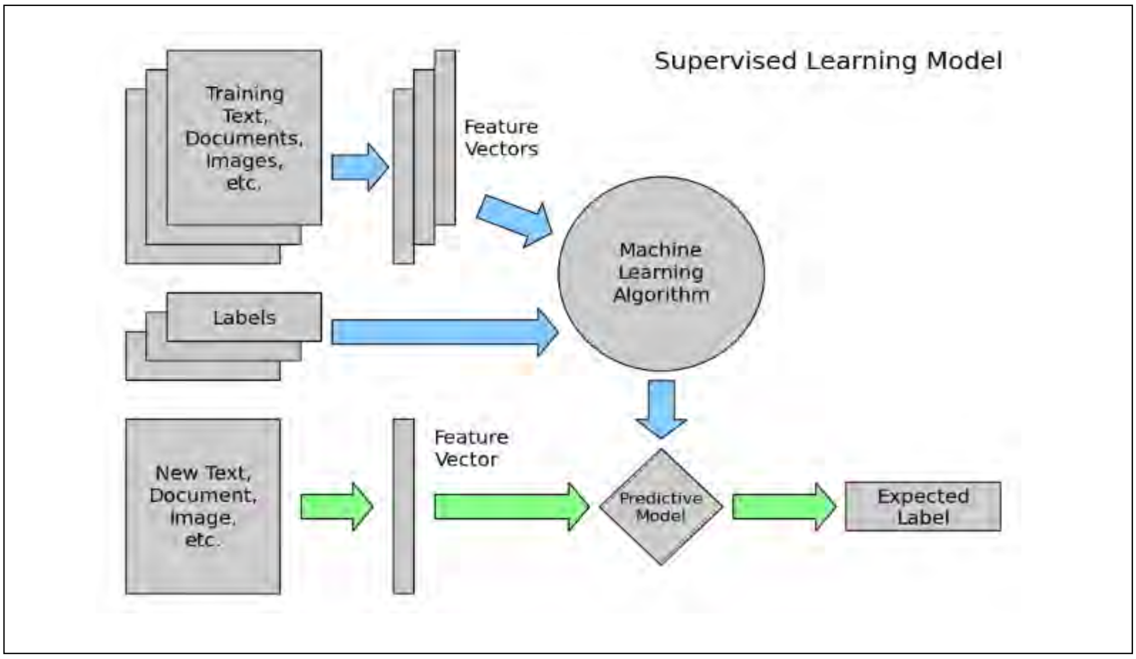

Supervised Learning Model
Training Text, Documents, Images, etc.
Feature Vectors
Machine Learning Algorithm
Labels
New Text, Document, Image, etc.
Feature Vector
Predictive Model
Expected Label

Figure 2.3: Block diagram of supervised machine learning algorithm (Source:[30])

Supervised learning is commonly used in applications where historical data predicts likely future events. For example, it can anticipate when credit card transactions are likely to be fraudulent or which insurance customer is likely to file a claim.

b. ***Unsupervised learning*** [30] [31] [32] is used against data that has no historical labels (figure 2.4). The system is not told the "right answer." The algorithm must figure out what is being shown. The goal is to explore the data and find some structure within. Unsupervised learning works well on transactional data. For example, it can identify segments of customers with similar attributes who can then be treated similarly in marketing campaigns. Or it can find the main attributes that separate customer segments from each other.

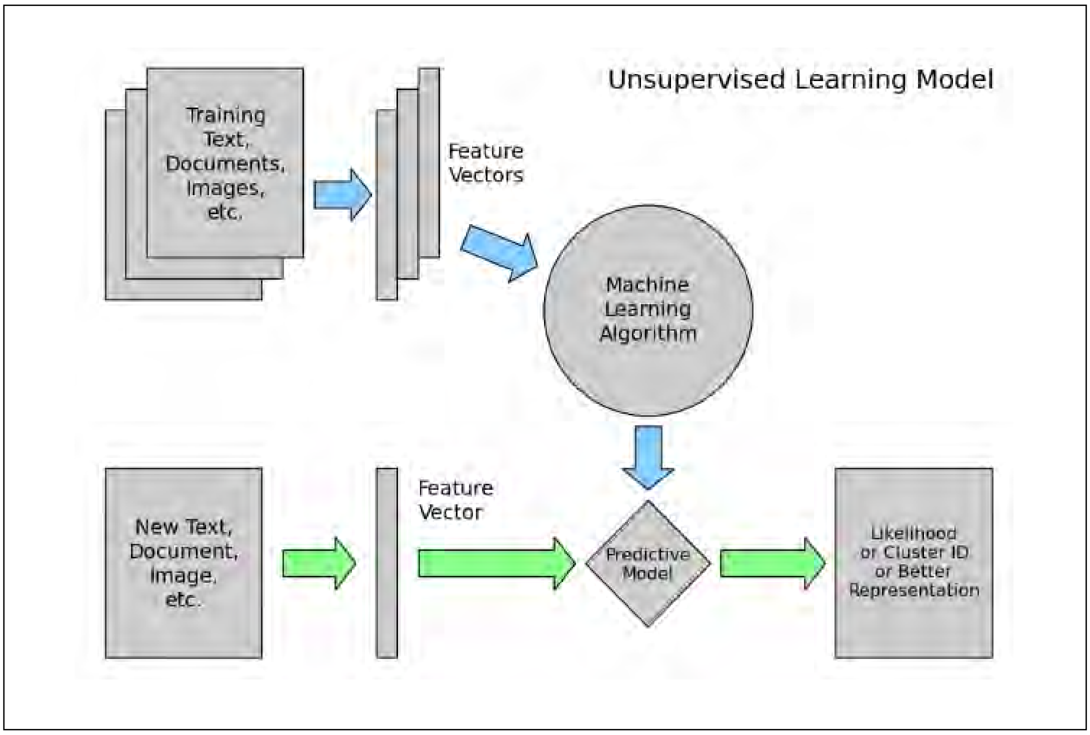

Unsupervised Learning Model
Training Text, Documents, Images, etc.
Feature Vectors
Machine Learning Algorithm
New Text, Document, Image, etc.
Feature Vector
Predictive Model
Likelihood or Cluster ID or Better Representation

Figure 2.4: Block diagram of unsupervised learning algorithm (Source:[30])

Popular techniques of unsupervised learning include self-organizing maps, nearest-neighbor mapping, k-means clustering and singular value decomposition. These algorithms are also used to segment text topics, recommend items and identify data outliers.

c. ***Semi supervised learning*** [30] [31] is used for the same applications as supervised learning. But it uses both labeled and unlabeled data for training (figure 2.5). Typically it uses a small amount of labeled data with a large amount of unlabeled data (because unlabeled data is less expensive and takes less effort to acquire).This type of learning can be used with methods such as classification, regression and prediction.

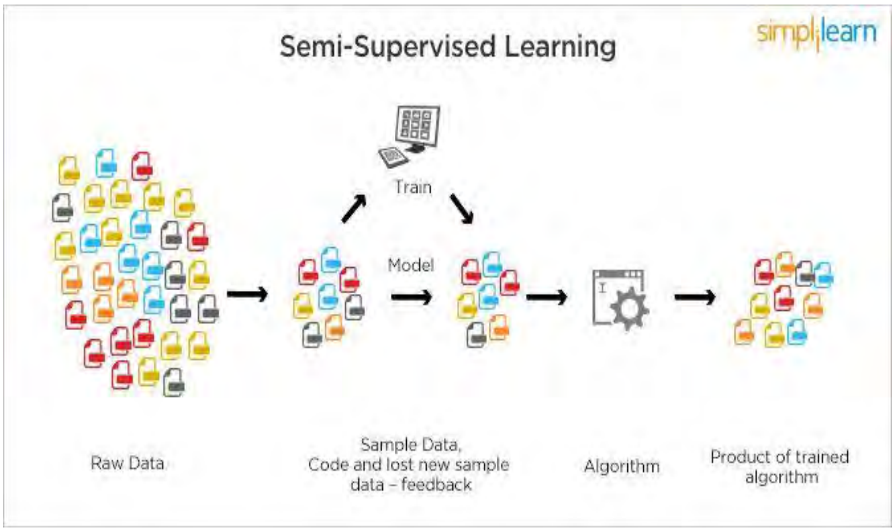


Figure 2.5: Block diagram of semi-supervised learning algorithm (Source:[31])

Semi supervised learning is useful when the cost associated with labeling is too high to allow for a fully labeled training process. Early examples of this include identifying a person's face on a web cam.

d. ***Reinforcement learning*** [30] [31] [32] is often used for robotics, gaming and navigation. With reinforcement learning, the algorithm discovers through trial and error which actions yield the greatest rewards (figure 2.6).

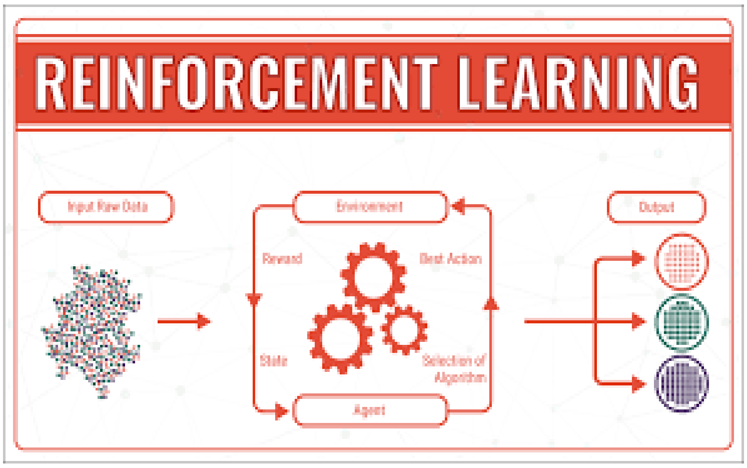


Figure 2.6: Block diagram of reinforcement learning algorithm (Source:[32])

This type of learning has three primary components: the agent (the learner or decision maker), the environment (everything the agent interacts with) and actions (what the agent can do). The objective is for the agent to choose actions that maximize the expected reward over a given amount of time. The agent will reach the goal much faster by following a good policy. So the goal in reinforcement learning is to learn the best policy.

2.2 Protein in a living cell

Proteins are large biological molecules that contain carbon, hydrogen, oxygen and nitrogen. Some proteins also contain sulphur. Proteins are made up of hundreds or thousands of smaller units called amino acids, which are attached to one another in long chains.

2.2.1 The protein molecule

The word 'protein' is derived from the Greek word ―proteios", meaning "primary" or "first" [33]. Proteins are vital for the growth and repair, and their functions are endless. Each and every property that characterizes a living organism is affected by proteins, whether it is a bacteria or a human body. They do most of the work in cells and are required for the structure, function, and regulation of the body‘s tissues and organs. Structure of amino acid is shown in figure 2.7.

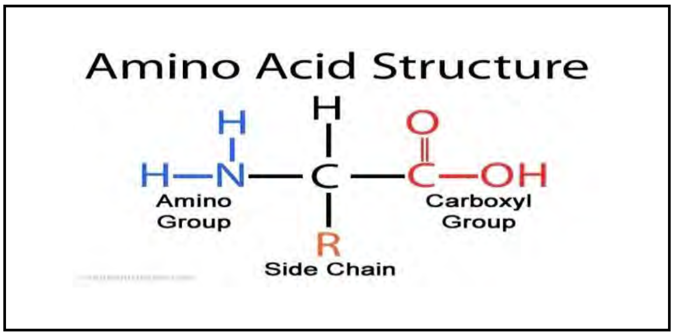


Figure 2.7: Structure of an amino acid (Source:[34])

2.2.2 Protein synthesis

Each of the 20 different amino acids has a different side chain which gives the amino acid its distinctive chemical identity. Amino acids can be represented by a letter, thus proteins can be represented by a sequence of letters. The process in which proteins are assembled is called synthesis [35] and this process is shown in figure 2.8.

Most genes contain the information needed to make functional molecules called proteins. (A few genes produce other molecules that help the cell assemble proteins.) The journey from gene to protein is complex and tightly controlled within each cell. It consists of two major steps: transcription and translation. Together, transcription and translation are known as gene expression.

During the process of transcription, the information stored in a gene's DNA is transferred to a similar molecule called RNA (ribonucleic acid) in the cell nucleus. Both RNA and DNA are made up of a chain of nucleotide bases, but they have slightly different chemical properties. The type of RNA that contains the information for making a protein is called messenger RNA (mRNA) because it carries the information, or message, from the DNA out of the nucleus into the cytoplasm [35].

Translation, the second step in getting from a gene to a protein, takes place in the cytoplasm. The mRNA interacts with a specialized complex called a ribosome, which "reads" the sequence of mRNA bases. Each sequence of three bases, called a codon, usually codes for one particular amino acid. (Amino acids are the building blocks of proteins.) A type of RNA called transfer RNA (tRNA) assembles the protein, one amino acid at a time. Protein assembly continues until the ribosome encounters a –stop” codon (a sequence of three bases that does not code for an amino acid).

The flow of information from DNA to RNA to proteins is one of the fundamental principles of molecular biology. It is so important that it is sometimes called the –central dogma.”

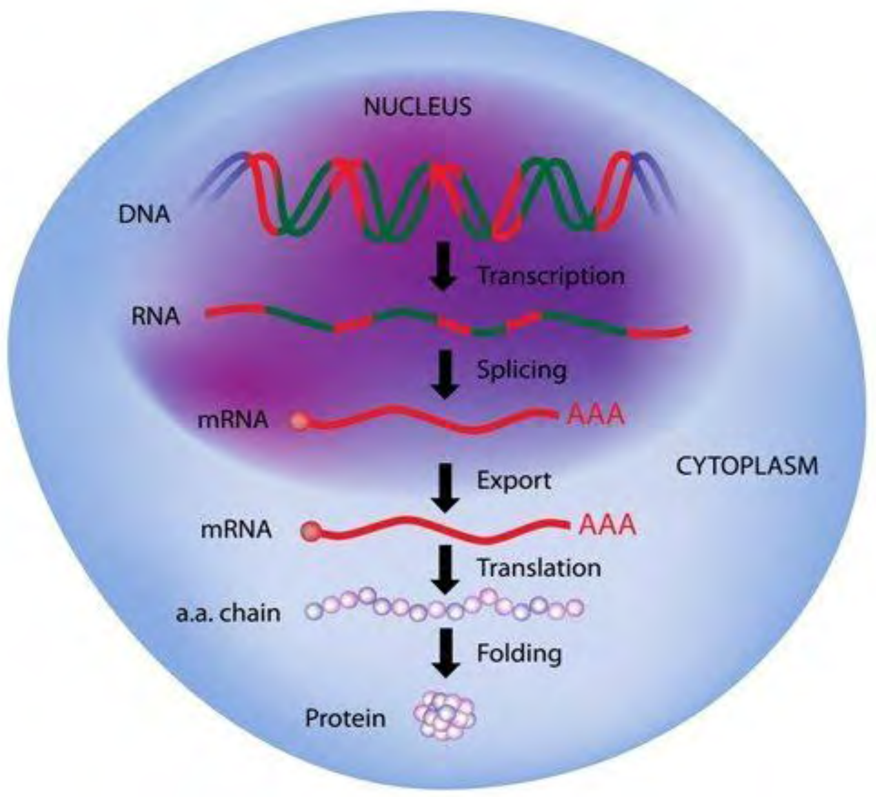


Figure 2.8: Protein synthesis inside a living cell (Source:[2])

Information encoded in genes determines which amino acids are used to assemble a protein. The sequence of amino acids is specified by a nucleotide sequence. Nucleotides are composed of nucleo bases and they are the building blocks of nucleic acids i.e. nucleotides are the monomers of nucleic acids. A linear chain of amino acid residues is called a polypeptide. A protein contains at least one long polypeptide. Short polypeptides, containing less than about 20-30 residues, are rarely considered to be proteins and are commonly called peptides, or sometimes oligopeptides.

2.2.3 Properties of protein

Proteins are the most important intracellular macro-molecules. Proteins are present in all chemical and physical activity which constitutes the life of a cell. Proteins are present in each living cell. They provide structure and protection to the body of an organism. Proteins form skin, hair, callus, cartilage, ligaments, muscles, tendons of any organism. They regulate and catalyze the body chemistry in the form of hormones, enzymes, immunoglobulins etc.

a) Some notable physical properties of protein:

1. ***Molecular weight:*** For any protein sequence, molecular weight [33]is calculated as the sum of the atomic weights of each amino acid multiplied by the number of atoms of that amino acid in the molecular formula.

2. ***Atomic composition:*** Most proteins consist of linear polymers built from series of up to 20 different L-α-amino acids. All proteinogenic amino acids possess common structural features, including an α-carbon to which an amino group, a carboxyl group, and a variable side chain are bonded. Only proline differs from this basic structure as it contains an unusual ring to the N-end amine group, which forces the CO–NH amide moiety into a fixed conformation [17] [33].

3. ***Instability index:*** The Instability index is a measure of proteins, used to determine whether it will be stable in a test tube [17] [33] [36] [37]. If the index is less than 40, then it is probably stable in the test tube. If it is greater (for example, enaptin) then it is probably not stable.

4. ***Aliphatic index:*** The aliphatic index of a protein is defined as the relative volume of its aliphatic side chains (alanine, valine, isoleucine, and leucine) [17] [33]. It may be regarded as a positive factor for increasing thermo stability in globular proteins.

b) Some notable chemical properties of protein:

1. ***Hydropathicity:*** The hydropathy index of an amino acid is a number representing the hydrophobic or hydrophilic properties of its side chain [17] [33]. The larger the number is, the more hydrophobic the amino acid. The most hydrophobic amino acids are isoleucine (4.5) and valine (4.2).

The most hydrophilic ones are arginine (-4.5) and lysine (-3.9). This is very important in protein structure; hydrophobic amino acids tend to be internal (with regard to the 3 dimensional shape of protein) while hydrophilic amino acids are more commonly found towards the protein surface.

2. ***Extinction coefficient:*** The extinction coefficient of a protein [17] [33] at a particular wavelength is the sum of the extinction coefficients of all the chromophores that absorb at that wavelength. For many proteins, the only chromophores are the aromatic amino acid side chains and disulfide bonds, which absorb in the UV range. The extinction coefficient of such proteins can be calculated based on the amino acid composition, but this is only valid when the protein is completely denatured in 8M guanidine hydrochloride. Some proteins have bound ligands or prosthetic groups, however, that also contribute to the absorbance (heme, FAD, etc.).

3. ***Half-life:*** The biological half-life of a biological substance (e.g.: protein) [17] [33] is the time it takes for half to be removed by biological processes when the rate of removal is roughly exponential. It is often denoted by the abbreviation $t_{½}$. Examples include metabolites, drugs, and signaling molecules.

2.2.4 Extracting properties of protein

Amino acid sequences and the features extracted from them is used to predict many protein as well as cell properties, such as - subcellular and subnuclear localization, cell functions, solubility and many more. Such extraction methods use different classification algorithms.

The sequence of a protein contains valuable information about its characteristics [1]. Various sequence-based prediction methods have proved this fact by classifying proteins according to specific properties [16]. This has given rise to various relevant and frequently used bioinformatics tools, offered by a growing number of easily accessible websites and web services [17] [18].

Sequence-based protein classifiers assign class labels to proteins based on a set of features those capture some property of sequences. This process requires three steps:

1. ***Feature extraction:*** to map protein sequences to points in a feature space
2. ***Construction of a classifier:*** to optimally separate protein classes in its feature space, using a set of proteins with known class labels
3. ***Prediction of new proteins:*** the trained classifier can be applied to predict class labels for new proteins

Although online as well as offline software tools are available both for feature extraction and classifier construction, their applications are not always straightforward. They require users to install various packages and to convert data into different formats, which are not applicable for many cases. These problems lead to lack of use of this software for sequence-based classification of protein, and requires emergence of user-friendly resources.

2.3 Subcellular localization of protein

Subcellular localization of protein determines the access of proteins to interacting partners and the post-translational modification machinery and enables the integration of proteins into functional biological networks.

2.3.1 What it is

Cells (shown in figure 2.9) are organized into membrane-covered compartments [14], which are characterized by specific sets of proteins. Functional activities of proteins are linked to their subcellular locations and complex molecular interactions [20]. Therefore, subcellular localization of protein is essential for protein function

[16] [19] [4]. It is a mean of achieving functional diversity of protein, as well as economizing protein design and synthesis [4][38][39] [40].

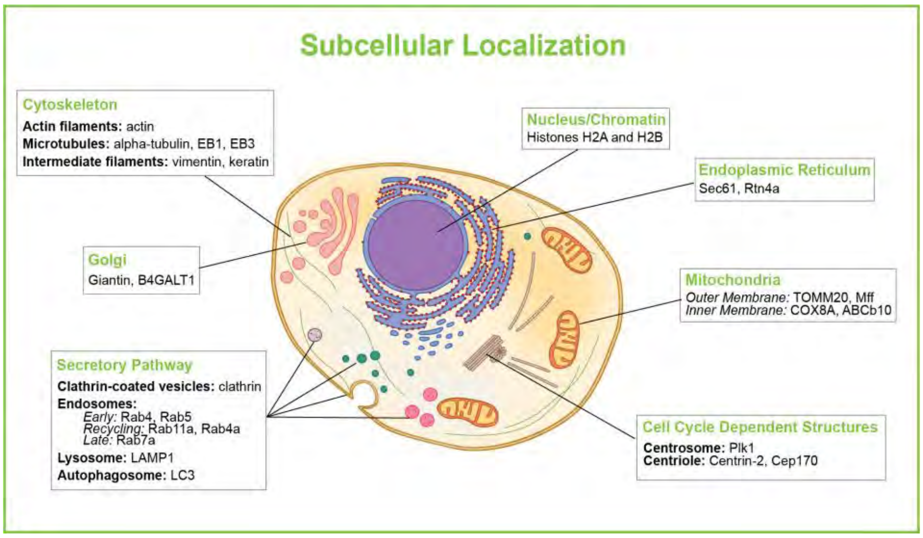


Figure 2.9: Different parts of a cell (Source:[40])

2.3.2 Result of improper localization of protein

To gain access to appropriate molecular interaction partners, protein must be at the right place at the right moment. Therefore, improper protein localization is a prominent feature of a broad range of medical conditions and diseases. Protein mislocalization could be caused by alterations, as:

1. mutations within signal sequences
2. changes in post-translational modifications or expression level of the cargo protein itself
3. by deregulation of the protein trafficking machinery

Improper protein localization can cause several medical conditions and diseases [38][39]. They have already been discussed in the previous sections.

2.3.3 Machine learning for subcellular localization of protein

Machine learning is a method of programming computers where, instead of designing the algorithm to explicitly perform a given task, the machine is programmed to learn from an incomplete set of examples. There are several different machine learning paradigms, such as the naive Bayes rule, artificial neural networks, genetic algorithms, and decision tree learning [41]. A similar area of work is called Pattern Recognition, which is a method of data classification. It is the study of methods and algorithms for putting data objects into categories [41] [42] [43]. While classical pattern recognition techniques are rooted in statistics and decision theory, the machine learning paradigm is commonly used to design practical systems. Figure 2.10 shows the process flow of pattern recognition.

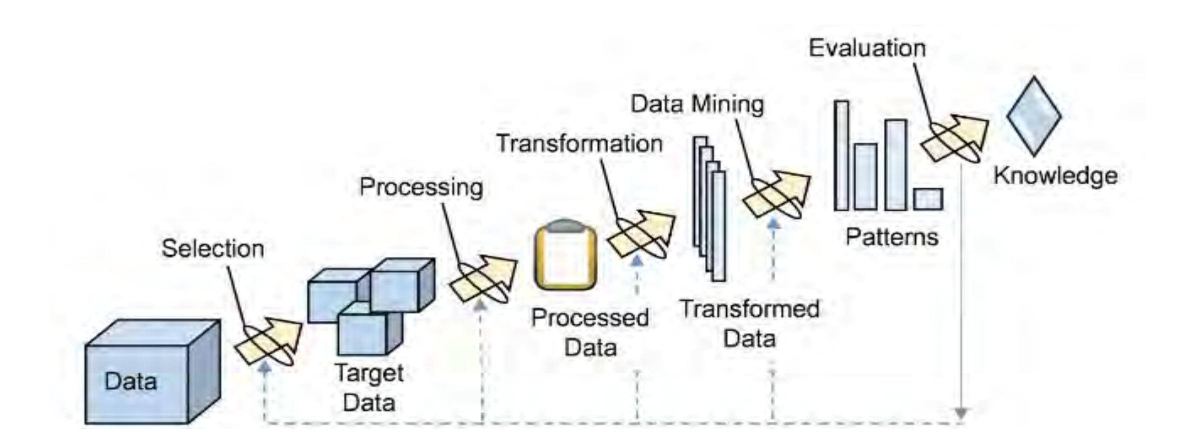


Figure 2.10: Process flow of pattern recognition (Source:[30])

Localization of proteins inside a cell is a budding topic of research in both of bioinformatics and machine learning fields [1]. Most of the popular methods for protein subcellular localization use images, or some use protein sequences with so many complex features as input data [7] [8] [9] [10][38]. Thus they require more time and money for feature extraction and implementation of their algorithm. For this purpose, this thesis work focuses on using very basic properties of protein; which are easy to extract, easy to use as features and results in a less complex classification algorithm.

In the past, machine learning techniques such as the hidden markov model (HMM) [12], neural network [13], K-nearest neighbor [14], and support vector machine (SVM) [19] [15] [20] [21] have been used for subcellular and subnuclear localizations, where several variants of SVM (e.g., linear [19] or non-linear [20] [15]) are applied in most algorithms. However, considering the inclusion of large number of features [20] [15] and complexity of the algorithms [19] [20] [21] with respect to the classification accuracy, it requires further improvement in classifier design and feature selection for better performance in labeling protein classes. Hence, the objective of this thesis is to develop an algorithm which is simple and can be implemented with low complexity compared to the existing algorithms.

2.4 Support Vector Machine (SVM) in depth

Support Vector Machines [44] [45] are based on the concept of decision planes that define decision boundaries. A decision plane is one that separates between a set of objects having different class memberships. A schematic example is shown in figure 2.11 below. In this example, the objects belong either to class GREEN or RED. The separating line defines a boundary on the right side of which all objects are GREEN and to the left of which all objects are RED. Any new object (white circle) falling to the right will be labeled/classified, as GREEN (or classified as RED if it falls to the left of the separating line).

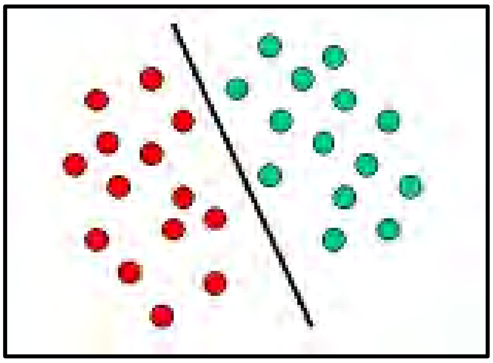


Figure 2.11: Linear decision plane in SVM (Source:[45])

Figure 2.11 is a classic example of a linear classifier, i.e., a classifier that separates a set of objects into their respective groups (GREEN and RED in this case) with a line. Most classification tasks, however, are not that simple, and often more complex structures are needed in order to make an optimal separation, i.e., correctly classify new objects (test cases) on the basis of the examples that are available (train cases). This situation is depicted in figure 2.12 below. Compared to the previous schematic, it is clear from figure 2.12 that a full separation of the GREEN and RED objects would require a curve (which is more complex than a line). Classification tasks based on drawing separating lines to distinguish between objects of different class memberships are known as hyperplane classifiers. Support Vector Machines are particularly suited to handle such tasks.

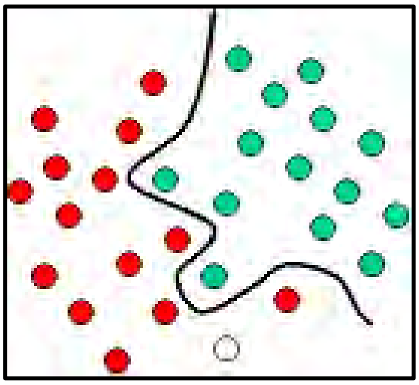


Figure 2.12: Non-linear decision plane in SVM (Source:[45])

Figure 2.13 shows the basic idea behind Support Vector Machines. Here, the original objects (left side of the schematic) are mapped, i.e., rearranged, using a set of mathematical functions, known as kernels. The process of rearranging the objects is known as mapping (transformation). Note that in this new setting, the mapped objects (right side of the schematic) is linearly separable and, thus, instead of constructing the complex curve (left schematic), all one need to do is to find an optimal line that can separate the GREEN and the RED objects.

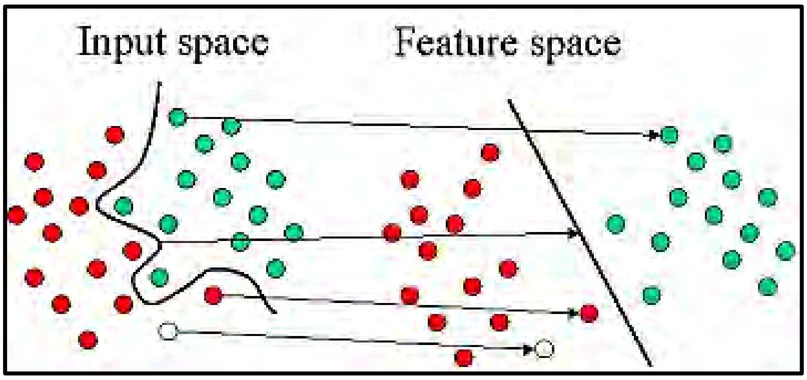


Figure 2.13: Mapping of data points from the input space to another feature space in SVM (Source:[45])

2.5 Existing applications of SVM based protein Subcellular localization

An algorithm proposed in [19] introduced new kernel functions used in an SVM learning model for the measurement of protein sequence similarity. The k-peptide vectors are first mapped by a matrix of high-scored pairs of k-peptides which are measured by BLOSUM62 scores. The new kernels are then defined on the mapped vectors. By combining these new encoding methods, they have established a multi-class classification system for the prediction of protein subnuclear localizations for the first time. The overall accuracy of prediction for 6 subnuclear localizations is about 50% (vs. random prediction 16.7%) for single localization proteins in the leave-one-out cross-validation, and 65% for an independent set of multi-localization proteins.

Another algorithm in [15] proposed an evolutionary support vector machine (ESVM) based classifier with automatic selection from a large set of physicochemical composition (PCC) features of protein to design an accurate system for predicting protein subnuclear localization. ESVM (which uses an inheritable genetic algorithm combined with SVM) can automatically determine the best number, m of PCC features and identify m out of 526 PCC features simultaneously. Using a leave-one-out cross-validation and selecting m= 33 and 28 PCC features, the system shows accuracies of 56.37% for database SNL6 and 72.82% for database SNL9.

One study in [20] introduced segmentation distribution and segmented auto-covariance feature extraction methods to explore local evolutionary-based information. It uses consensus sequence-based and semi-occurrence to extract global evolutionary-based information. Use of Support Vector Machine (SVM) as the classification technique gives them an accuracy of 86%.

A group of researchers in [22] considered a set of SVMs, and trained it to predict the subcellular location of a given protein based on its amino acid, amino acid pair, and gapped amino acid pair compositions. The predictors based on these different compositions were then combined using a voting scheme. They considered 12 subcellular locations in eukaryotic cells: chloroplast, cytoplasm, cytoskeleton, endoplasmic reticulum, extracellular medium, Golgi apparatus, lysosome, mitochondrion, nucleus, peroxisome, plasma membrane, and vacuole. Then they constructed a data set of proteins with known locations from the SWISS-PROT database. Results obtained through 5-fold cross-validation tests showed an improvement in prediction accuracy over the algorithm based on the amino acid composition only. The final accuracy is calculated to be 78-79%.

The research work in [24] proposes an efficient sequence-based method (named ProLoc-GO) by mining informative GO terms for predicting protein subcellular localization. For each protein, BLAST is used to obtain a homology with a known accession number to the protein for retrieving the GO annotation. A large number n of all annotated GO terms that have ever appeared are then obtained from a large set of training proteins. A novel genetic algorithm based method (named GO mining) combined with a classifier of support vector machine (SVM) is proposed to simultaneously identify a small number m out of the n GO terms as input features to SVM, where m <<n. For comparison, ProLoc-GO using known accession numbers of query proteins yields test accuracies of 90.6% and 85.7% for two different datasets respectively.

The effective use of GO terms in solving sequence-based prediction problems remains challenging, especially when query protein sequences have no accession number or annotated GO term. So, the study in [25] obtained homologies of query proteins with known accession numbers using BLAST to retrieve GO terms for sequence-based subnuclear localization prediction. A prediction method PGAC, which involves mining informative GO terms associated with amino acid composition features, is proposed to design a support vector machine-based classifier. PGAC yielded an LOOCV accuracy of 81.1%.

A group of researchers has developed computational methods to automatically analyze the yeast images created by the UCSF yeast GFP fusion localization project [26]. The SVM generated automated method provides an objective, quantitative and repeatable assignment of protein locations that can be applied to new collections of yeast images (e.g. for different strains or the same strain under different conditions). Results have an accuracy of 81.02% for all proteins.

To reveal the cancer-related protein mislocalization, some researchers stated in [28]about them developing an image-based multi-label subcellular location predictor, iLocator, which covers seven cellular localizations. The iLocator incorporates both global and local image descriptors and generates predictions by using an ensemble multi-label classifier. It shows an accuracy of 77.83%.

2.6 Limitations of existing methods

A number of prediction algorithms (such as above ones) have been developed based on protein images (along with complex feature extraction methods), amino acid compositions or on the N-terminal characteristics (signal peptides) of proteins. However, such approaches lead to a loss of contextual information. Moreover, where information about the physicochemical properties of amino acids has been used, the methods employed to exploit that information are less than optimal and could use the information more effectively. Also, the above methods use feature sets and protein localization procedures in such a manner that their algorithms require too much work and time to be executed.

Summary of chapter 2:

Chapter 2 describes some of the topics and techniques related to the current thesis work. They have brought to light the necessity of properly predicting the subcellular locations of proteins. The already established methods are briefly stated here, who shows the current situation and limitations in this field. In order to overcome such limitations, the newly proposed algorithm in this thesis will show a path.

CHAPTER 3
Data Collection

3.1 Experimental setup

This thesis proposes a new method of protein localization within cells. It uses some basic physical and chemical features of protein (described in next segments), extracting them from protein sequences. They create an opportunity to start a new era of working with less complex features and developing newer algorithms with less-complex computations.

Finding out the optimal feature subset is as important as selecting an appropriate algorithm. So some important feature selection techniques have been incorporated here to select optimal features, such as: t-test, filter method and wrapper method. Feature selection has a significant impact on subsequent stages of the learning.

The selection of appropriate classification techniques is one of the most important aspects for prediction issues. To find out the appropriate one for the proposed method, a number of classification algorithms (SVM with different kernels, PART [46], KNN [47]) have been experimented with and then the best one was selected.

The classifier requires optimum parameters those perform best with given feature set. So steps have been taken to find out the best value pair suited to the dataset. Use of cross-validation technique [48] here ensured random split of data into train and test set. Then performance of the proposed algorithm has been checked (results will be provided in chapter 5).

Figure 3.1 in the next section presents a block diagram for methodology of the proposed algorithm.

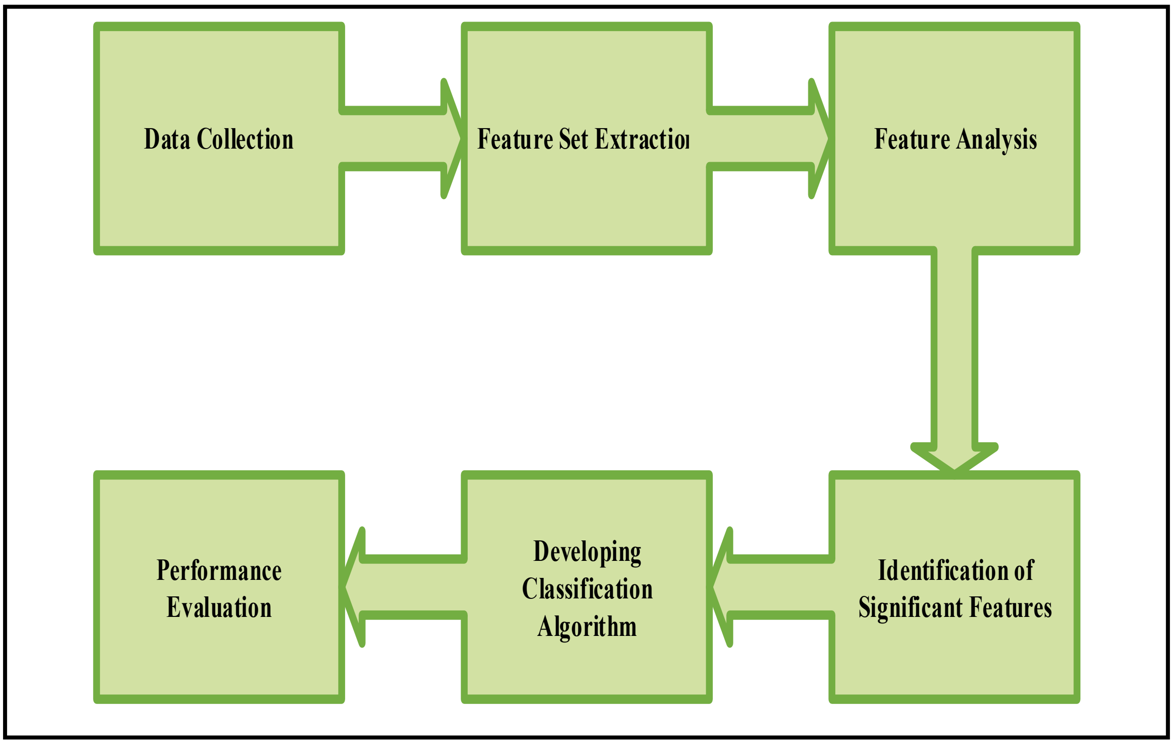


Figure 3.1: A block diagram of the proposed methodology presented in this research work

3.2 Collection of protein data

This report considers a standard protein dataset available in the web [15] for 6 different locations in a cell. The original dataset contains 504 protein sequences. All of them were annotated with singular subcellular locations. Due to the loss of several protein sequences in the online database, a total of 361 sequences were extracted out of 504 sequences. Table 3.1 presents a list of protein sequences in each of the 6 locations, which were finally included in this thesis work.

Table 3.1: Number of protein sequences in each of 6 subcellular locations

Location	**Number of sequences**
Chromatin	38
Nuclear Speckles	56
Nucleolus	101
Nucleoplasm	75
Nuclear Lamina	53
PML Body	38
Total =	**361**

3.3 Feature extraction

The features mentioned and described here are either physical or chemical properties of amino acids. They are very basic and well known properties of any protein sequence. Their extraction process is easy and inexpensive.

The high performance of the computational prediction based methods may arise from the use of a large number of complex feature set and more complex training model. To avoid such computation intensive nature of subcellular localization of proteins, the proposed algorithm aimed to utilize a generic feature set which will be easy to extract and less complex to compute with.

The values for the required features have been extracted with the help of an online service provider [49]. The features are arranged as follows:

1. The count of highest amino acid
2. The count of 2nd highest amino acid
3. The count of 3rd highest amino acid
4. Theoretical pI
5. Instability index
6. Aliphatic index
7. Molecular weight
8. Alignment score
9. Hydropathicity
10. Total negatively charged residues
11. Total positively charged residues
12. N terminal amino acid
13. The amino acid pair present in highest number
14. The amino acid pair present in 2nd highest number
15. The amino acid pair present in 3rd highest number
16. Location of the highest amino acid in the string
17. Location of the 2nd highest amino acid in the string
18. Location of the 3rd highest amino acid in the string

The description of each feature is given below:

1. ***The amino acid with highest count:*** Finding out which amino acid is present in the protein sequence in highest number. The amino acid is different for each protein location which is apparent from data distribution.

2. ***The amino acid with 2ndhighest count:*** Finding out which amino acid is present in the protein sequence in 2nd highest rank.

3. ***The amino acid with 3rd highest count:*** Finding out which amino acid is present in the protein sequence in 3rd highest rank.

   The name of the highest, 2nd highest and 3rd highest amino acid of each location is apparent from the protein sequences by visual observation of the data distribution. Then these observations were used to derive these three features from the dataset. Though these three features are easily extractable from protein sequences, no method discussed in literature have used them yet.

4. ***Theoretical pI (Isoelectric Point):*** The isoelectric point (pI, pH (I), IEP) is that pH at which a particular molecule carries no net electrical charge in the statistical mean. The pI value affects the solubility of any molecule at a given pH. These molecules have minimum solubility in water or salt solutions at the pH that corresponds to their Pi [17] [33].

   Amino acids making up proteins may be positive, negative, neutral, or polar in nature. They together give a protein its overall charge. At a pH below their pI, proteins carry a net positive charge, where at above their pI they carry a net negative charge.

   A number of algorithms have been developed for estimating isoelectric points of peptides/proteins, and most of them use Henderson–Hasselbalch equation with different pKa values. The pKa value of an amino acid is dependent on its side chain. Some recent approaches are based on a support vector machine algorithm [50] and pKa optimization against experimentally known protein/peptide isoelectric points [51]. Again, experimentally measured isoelectric point of proteins was aggregated into the databases [52] [53] . A database of isoelectric points for all proteins predicted using most of the available methods had been also developed recently [54].

5. ***Instability index:*** The instability index provides an estimate of the stability of any protein inside a test tube [17] [33]. Research on 12 unstable and 32 stable proteins has revealed [36] [37] that there are certain dipeptides, the occurrence of which is significantly different in the unstable proteins compared with those in the stable ones. This method includes a weight value of instability to each of the 400 different dipeptides (DIWV), and therefore an instability index (II) can be computed as defined in equation 3.1 below:

$$II = \left(\frac{10}{L}\right) \times \sum_{i=1}^{i=L-1} DIWV(x_i x_{i+1}) \quad \textit{... (3.1)}$$

Where L is the length of sequence, $DIWV(x_i x_{i+1})$is the instability weight value for the dipeptide (x) starting in position i.

For example: a protein with instability index smaller than 40 is predicted as stable, whereas a value above 40 tells the protein may be unstable.

6. ***Aliphatic index:*** The aliphatic index of a protein is defined as the relative volume of its aliphatic side chains (alanine, valine, isoleucine, and leucine) [17] [33]. It may be regarded as a positive factor for increasing thermo stability in globular proteins. The aliphatic index of a protein is calculated as [55] it is defined in equation 3.2:

$$Aliphatic\ Index = X_{Ala} + a \times X_{Val} + b \times (X_{Ile} + X_{Leu}) \quad \textit{... (3.2)}$$

Where X_{Ala}, X_{Val}, X_{Ile}, and X_{Leu} are mole percent (100 × mole fraction) of alanine, valine, isoleucine and leucine respectively. The coefficients a and b are the relativevolume of valine side chain (a = 2.9) and of *Leu/Ile* side chains (b = 3.9) to the side chain of alanine.

7. ***Molecular weight:*** Molecular weight or molecular mass is the mass of a molecule. For any protein sequence, it is calculated as the sum of the atomic weights of each amino acid multiplied by the number of atoms of that amino acid in the molecular formula [17] [33]. Here it has been measured in kilodaltons.

8. ***Alignment score of protein sequence:*** It shows how much an unseen protein sequence is aligned with a reference sequence. Use of alignment score in classification of DNA and RNA is common [56][57][58][59], but no algorithm in literature has used it yet for classifying proteins. This thesis work incorporates SSAADR (Smart Sequence Alignment Algorithm applying DNA Replication) [56] algorithm, for a faster and more accurate alignment of the amino acid sequences than using other popular algorithms. SSAADR algorithm does both global and local alignment using the concepts of Needleman-Wunsch (NW) [57] and Smith-Waterman algorithms (SW) [58]. In SSAADR algorithm, the concept of the NW is used for the trace-back and SW is used for main iteration. Initialization of first row and first column of matrix M (i, j) is done with zeros. Trace back starts from the last cell (M (m X n)) and ends at the first cell (M (0, 0)). As Smith-Waterman performs faster than Needleman-Wunsch, the concept of SW is used for filling the matrix. For trace back purpose, the concept of NW is used. SSAADR overcomes some problems of popular NW and SW algorithms. By implementing the SSAADR algorithm the execution time is reduced and the efficiency is greater than other algorithms. SSAADR algorithm based on taking the advantage of dynamic algorithms, works to get the optimal solution for the sequences alignment, and also take the advantage of the heuristic algorithm that it is to decrease the execution time for the sequence comparisons.

9. **Hydropathicity:** It is calculated as GRAVY (Grand Average of Hydropathy). The GRAVY value for a peptide or protein is calculated as the sum of hydropathy values [60] of all the amino acids, divided by the number of residues in the sequence.

10. **Total negatively charged residues:** It is the total count of the amino acids- Aspartic acid (Asp) and Glutamic acid (Glu).

11. **Total positively charged residues:** It is the total count of the amino acids- Arginine (Arg) and Lysine (Lys).

12. **N terminal amino acid:** The N-terminus (also known as the amino-terminus, NH2-terminus, N-terminal end or amine-terminus) is the start of a protein or polypeptide referring to the free amine group (-NH2) located at the end of a polypeptide. Normally the amine group is bonded to another carboxylic group in a protein to make it a chain, but since the end of a protein has only 1 out of 2 areas chained, the free amine group is referred to the N-terminus. By convention, peptide sequences are written N-terminus to C-terminus, left to right in LTR languages. This correlates the translation direction to the text direction (because when a protein is translated from messenger RNA, it is created from N-terminus to C-terminus - amino acids are added to the carbonyl end).

13. **The amino acid pair present in highest number:** Finding out which pair of amino acid is present in the protein sequence in highest number. This pair of amino acid is different for each location which is apparent from data distribution.

14. **The amino acid pair present in 2nd highest number:** Finding out which pair of amino acid is present in the protein sequence in 2nd highest rank.

15. **The amino acid pair present in 3rd highest number:** Finding out which pair of amino acid is present in the protein sequence in 3rd highest rank.

The name of the highest, 2nd highest and 3rd highest amino acid pair of each protein location is apparent from the protein sequences by visual observation of the data distribution. Then these observations were used to derive these three features from the dataset. Though these three features are easily extractable from protein sequences, no method discussed in literature have used them yet.

16. **Location of the highest amino acid in the string:** Finding out the highest amino acid (from the 1st feature) is present in which location of the protein sequence.

17. **Location of the 2nd highest amino acid in the string:** Finding out the 2nd highest amino acid (from the 2nd feature) is present in which location of the protein sequence.

18. **Location of the 3rd highest amino acid in the string:** Finding out the 3rd highest amino acid (from the 3rd feature) is present in which location of the protein sequence.

For extracting value of feature no. 16, 17 and 18, the protein sequences have been divided into 3 parts. Then the specific amino acid has been searched for in these 3 parts. The target locations are named as: 1st part, middle part and last part.

3.4 Analysis of features

The data distribution tells how the data points are actually related to each other, either within the same location or with points of different locations. Figure 3.2, figure 3.3, figure 3.4, figure 3.5, figure 3.6 and figure 3.7 depict that the 18 features for this dataset can be proved significant when they are considered together. Observation from the 18 features presented in these three figures says that no feature is strong enough to differentiate among the 6 subcellular locations alone. So they need to be used altogether in a feature set.

3.5 Feature selection

Two types of methods have been used during this thesis work for selecting significant features. They are statistical significance test (t-test and filter method) and wrapper methods [61] [62] [63]. Pair wise t- test tells whether the variation between two groups of data is "significant" or not. Filter method is used to build a significant feature set, before developing the actual data model. On the other hand, wrapper method uses the feature set to check its performance on the developed model.

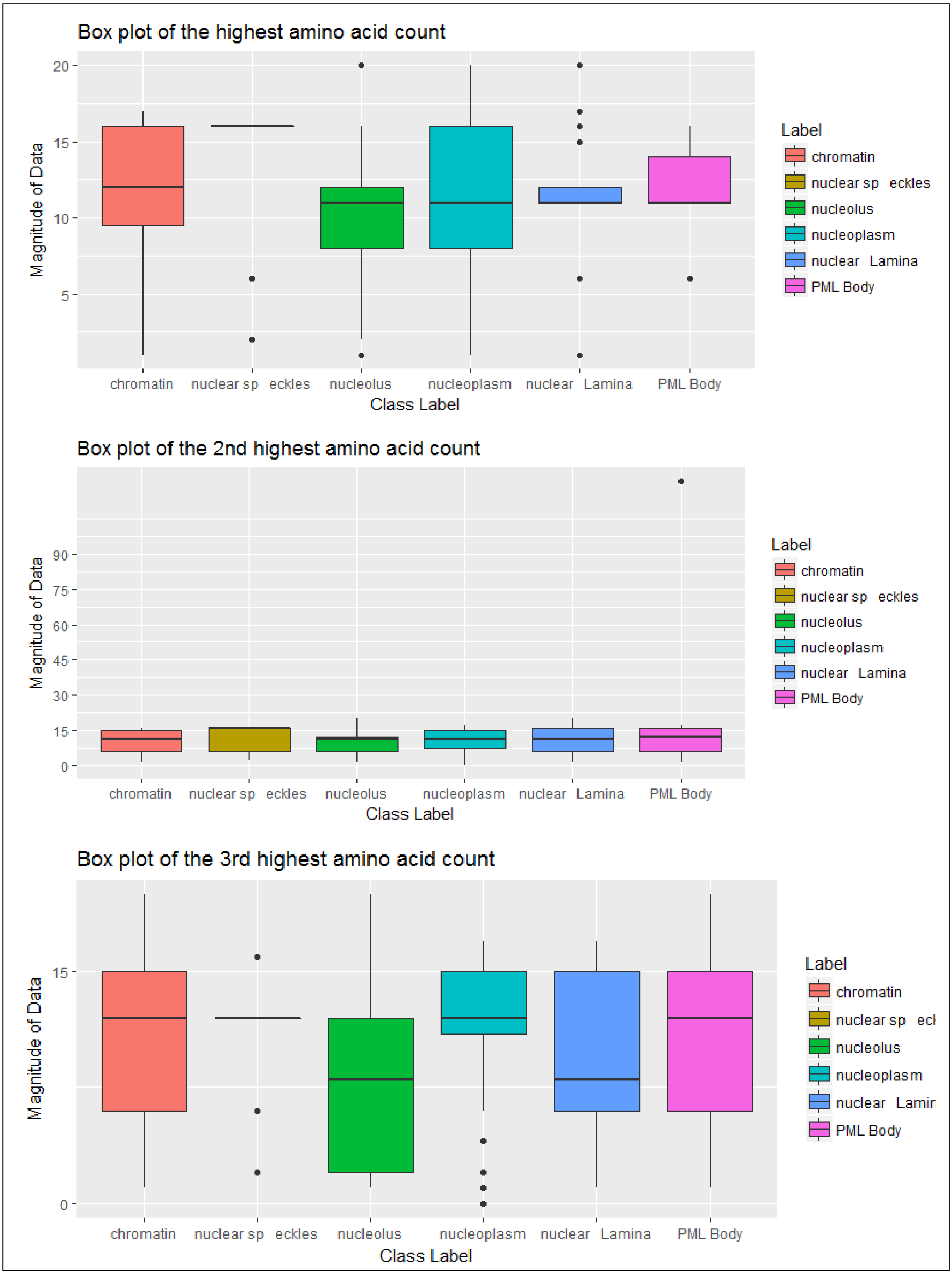
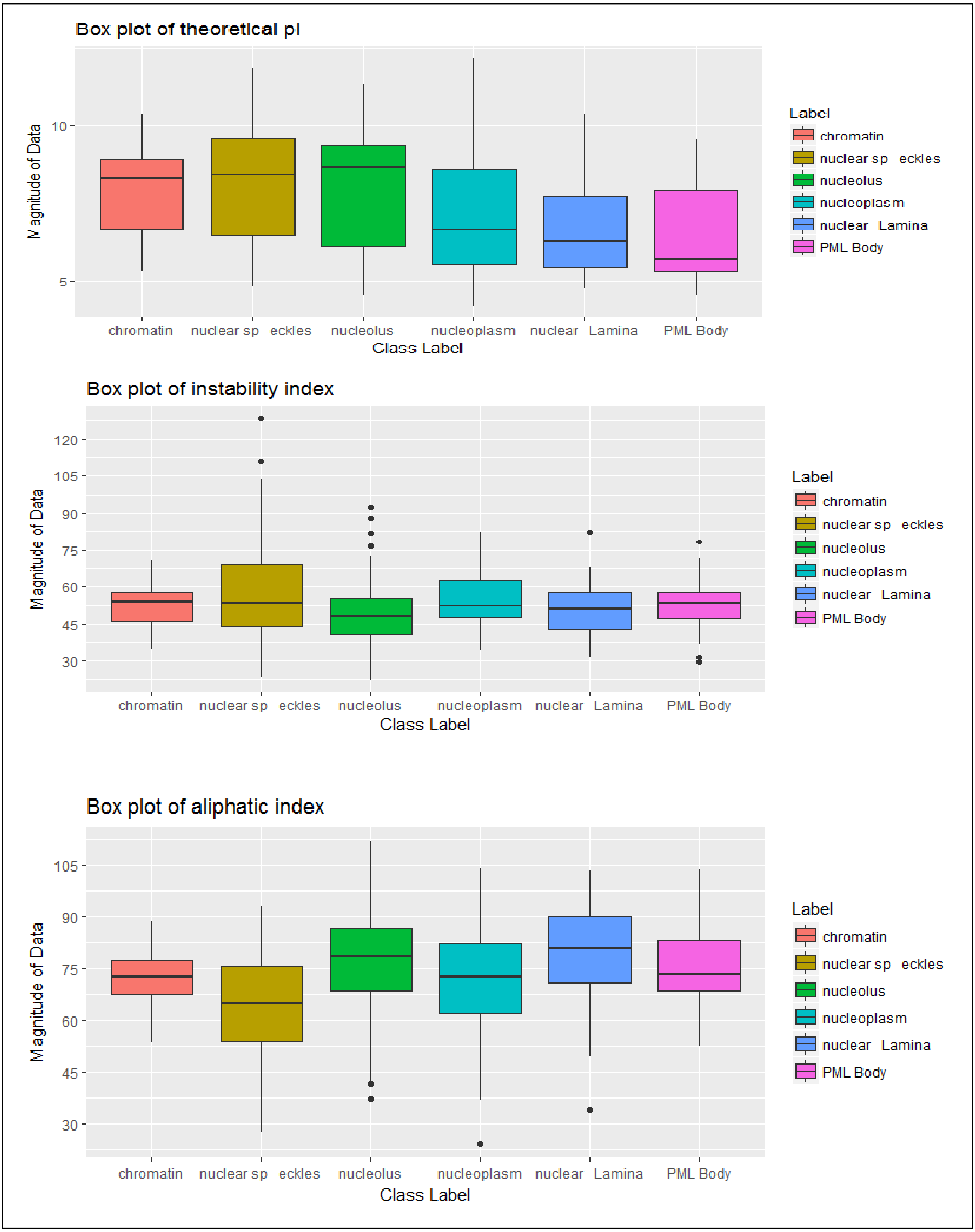


Figure 3.2: Box plots showing the data distribution of feature no. 1, 2 and 3 in 6 different locations



Box plot of theoretical pI

Box plot of instability index

Box plot of aliphatic index

Figure 3.3: Box plots showing the data distribution of feature no. 4, 5 and 6 in 6 different locations

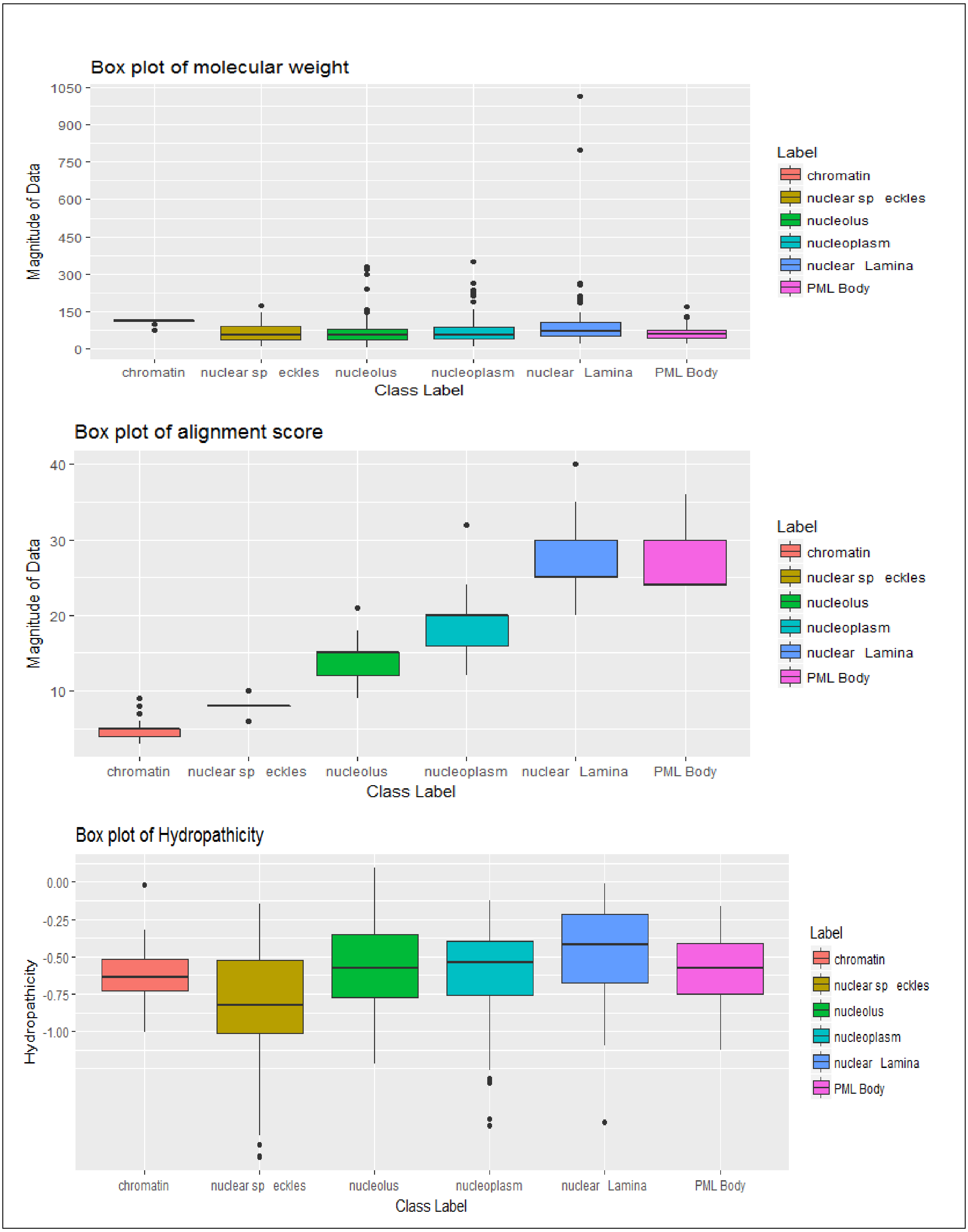


Figure 3.4:Box plots showing the data distribution of feature no. 7, 8 and 9 in 6 different locations

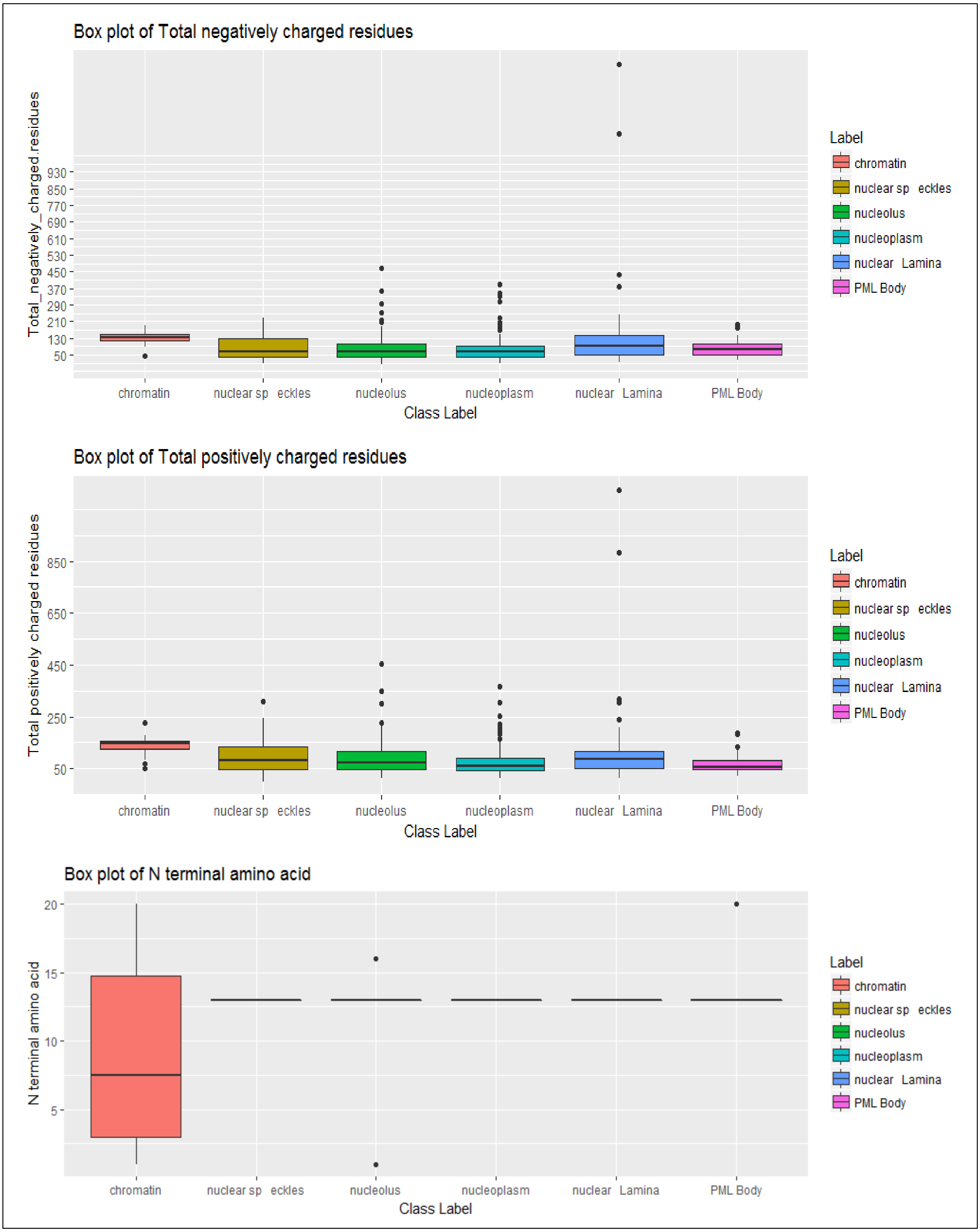


Figure 3.5:Box plots showing the data distribution of feature no. 10, 11 and 12 in 6 different locations

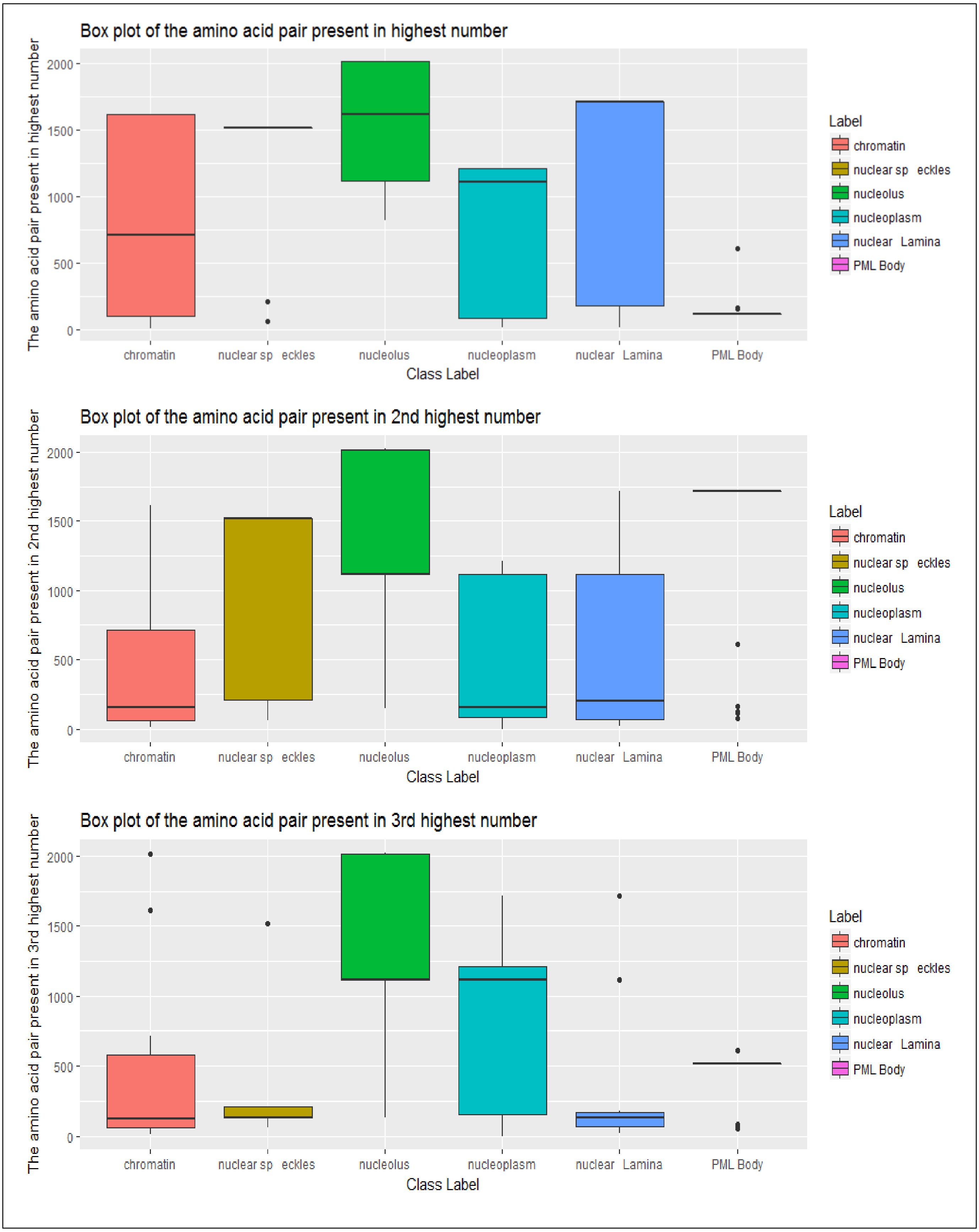


Figure 3.6:Box plots showing the data distribution of feature no. 13, 14 and 15 in 6 different locations

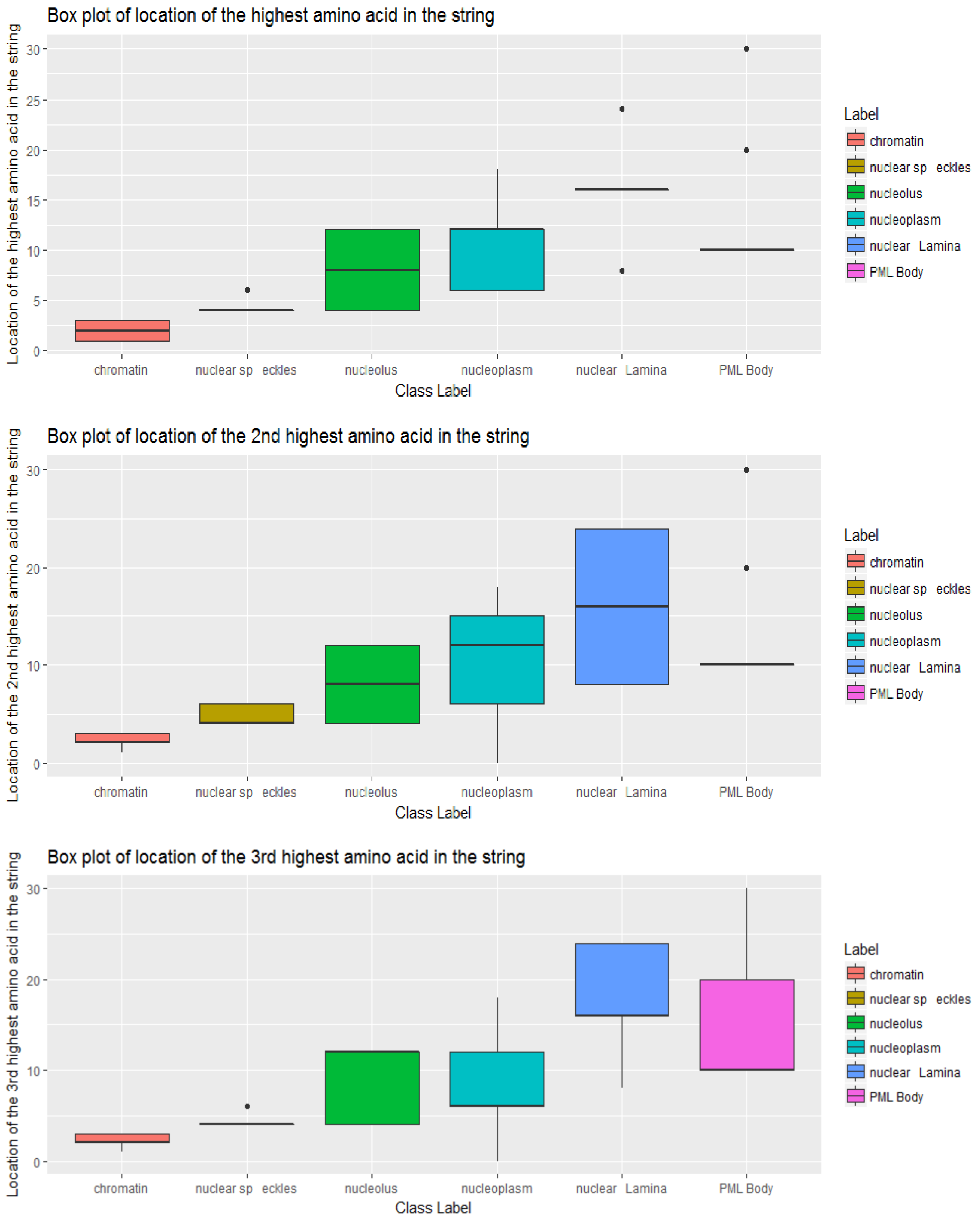


Figure 3.7:Box plots showing the data distribution of feature no. 16, 17 and 18 in 6 different locations

3.5.1 Statistical significance test

In this thesis work, a total of 18 physicochemical properties were extracted from each protein sequence. Statistical significance of the extracted features needs to be determined by the p-value of a pairwise t-test (for $p<0.05$) [61] [62] and one-way ANOVA (Analysis of Variance, a filter method) [64] [65]. Pairwise t-test looks for a significant p-value between two groups of data. ANOVA tests the significance of group differences between two or more groups. It puts all the data into one number (F) and gives one p-value to check the significance of total dataset. Online tools have been used for these tests [61] [62] [64]. Results of these tests are provided in the result section (chapter 5).

3.5.2 Wrapper Method test

Wrapper method [65] selects best features for prediction. It evaluates more than one data model using different procedures to find combination where there is maximum performance of the model. Some common examples of wrapper methods are forward feature selection, backward feature elimination and embedded feature selection. A forward feature selection and a backward feature elimination is required to be adopted to find out the importance of each feature of the dataset of this thesis work. The result of this technique on the proposed model will be discussed in the result section (chapter 5), after development of the classification method.

Summary of chapter 3:

Chapter 3 discusses on how the features have been extracted and needs to be selected according to their significance. The final dataset has been used to simulate the classification algorithm which will be described in the next chapter.

CHAPTER 4

Development of the proposed algorithm

4.1 Overall methodology

A block diagram of the proposed algorithm is presented in figure 4.1. The techniques of data collection, feature extraction and feature selection have already been discussed in the previous chapter. After selecting significant features, the final feature set have been constructed. Now, a classification algorithm needs to be developed which will maximize the accuracy of predicting subcellular locations of proteins. For this purpose, a number of classification algorithms is required to be tested with for this dataset. Their description and working procedure will be discussed in this chapter. Their performance will be presented in chapter 5 (result section). At the end, the best classifier for this dataset will be selected for the proposed method.

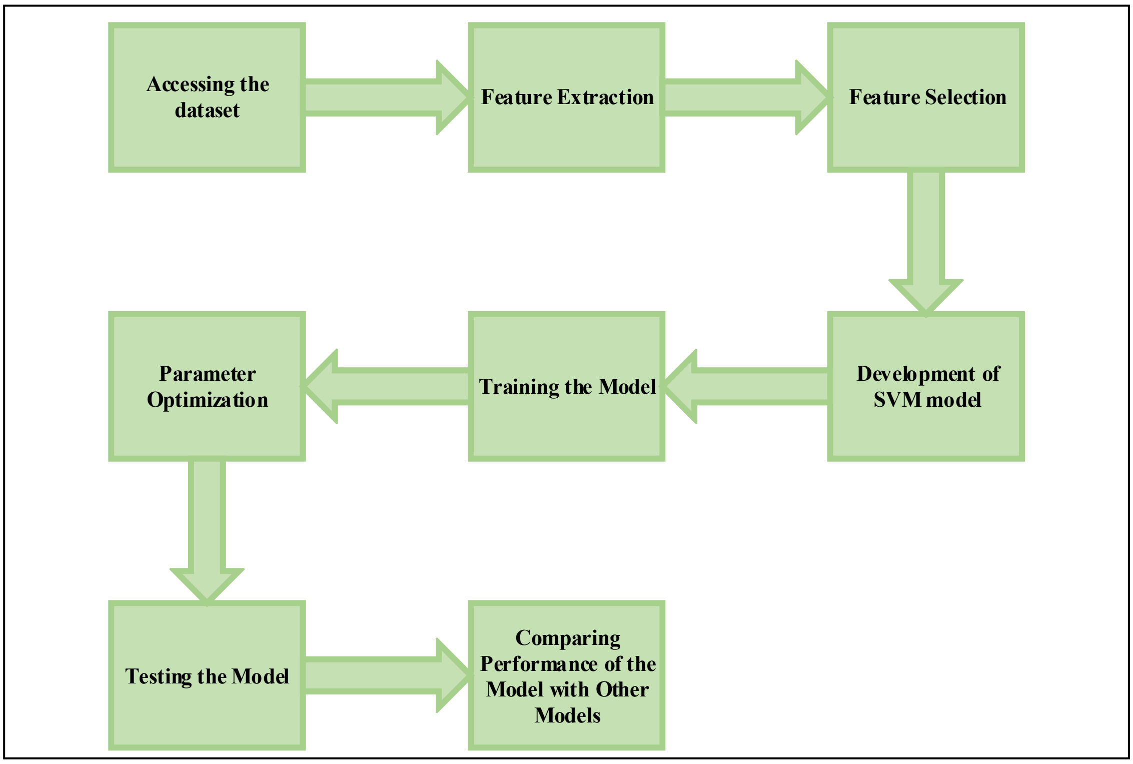


Figure 4.1: Flowchart showing different steps of the algorithm

4.2 Development of classification algorithm

In the next section, different steps related to the training model development for classification have been discussed in detail.

4.2.1 Selection of classifier

Support vector machine (SVM) [41] [44] is the most popular classifier in the field of predicting subcellular location of proteins. So it has been selected to be the first classifier to start working with. It can have different kernels depending on the nature of dataset. All types of kernels have been tested for the dataset in consideration to find out the best suited kernel. The results are recorded in table 5.1 of chapter 5. Then, some other classification algorithms have also been tested with for this dataset. Their performance outcome is in table 5.2 of chapter 5. Brief description and techniques of these classification algorithms are given here.

4.2.1.1 SVM

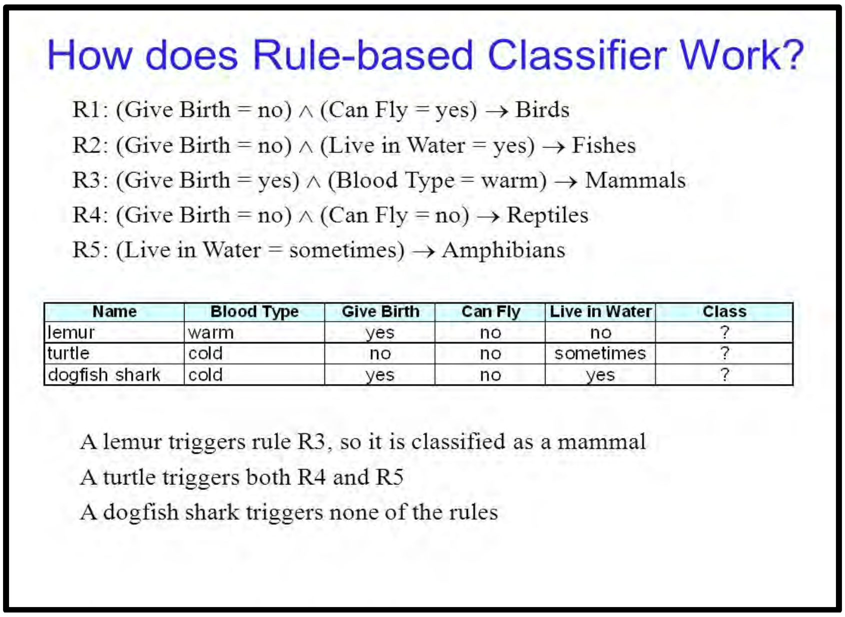
Support vector machine (SVM) [41] [44] attempt to pass a linearly separable hyperplane through a dataset in order to classify the data into two groups. This hyperplane is a linear separator for any dimension; it could be a line (2D), plane (3D), and hyperplane (4D+). If the data is not linearly separable, then a kernel trick is used. Kernels are functions that quantify similarities between observations. Common types of kernels used to separate non-linear data are - polynomial kernels, radial basis kernels, and linear kernels [66] [67] [68]. Simply, these kernels transform the data in order to pass a linear hyperplane and thus classify the data. So, the rule of thumb is to use linear SVMs (or logistic regression) for linear problems, and nonlinear kernels such as the Radial Basis Function kernel for non-linear problems. Extensions of support vector machines can be used to solve a variety of other problems, such as - multiple class SVMs using One-Versus-One Classification or One-Versus-All Classification. The chosen kernel defines the function class one is working with. The squared exponential kernel (radial basis function kernel) defines a function space that is a lot larger than that of the linear kernel or the polynomial kernel. A linear kernel allows the users to use linear functions, which are really impoverished. As the order of the polynomial kernel increases, the size of the function class increases. An n^{th} order polynomial kernel gives all analytic functions whose derivatives of order (n+1) are constant, and hence all derivatives of and above order (n+2) are zero. The rbf kernel gives access to all analytic functions (that is, all infinitely differentiable functions). So in some sense the rbf kernel can be viewed as powerful as an infinite order polynomial kernel. Technically if users use squared exponential kernel, then the method is nonparametric. And if the kernel is polynomial, the model is parametric. In a way nonparametric model means that the complexity of the model is potentially infinite, its complexity can grow with the data. If the users give it more and more data, it will be able to represent more and more complex relationships. In contrast, a parametric model's size is fixed. So after a certain point this model will be saturated, and giving it more and more data won't help. So asymptotically assuming users have unlimited data and very weak assumptions about the problem, a nonparametric method is always better. So, the rbf kernel is generally more flexible than the linear or polynomial kernels, as in it the users can model a whole lot more functions with its function space.

4.2.1.2 PART

Rules are a good way of representing information or bits of knowledge. A rule-based classifier uses a set of IF-THEN rules for classification (figure 4.2).
An IF-THEN rule is an expression of the form:

IF *condition* THEN *conclusion.*

PART [46] is a popular rule-based classifier which does not incorporate global optimization. It is a simple, surprisingly effective method for learning decision lists based on the repeated generation of partial decision trees in a separate-and-conquer manner. Despite this simplicity, PART produces rule sets that compare favorably with those generated by C4.5 [69] and C5.0 [69], and are more accurate than those produced by RIPPER [70] method.



How does Rule-based Classifier Work?

R1: (Give Birth = no) ∧ (Can Fly = yes) → Birds
R2: (Give Birth = no) ∧ (Live in Water = yes) → Fishes
R3: (Give Birth = yes) ∧ (Blood Type = warm) → Mammals
R4: (Give Birth = no) ∧ (Can Fly = no) → Reptiles
R5: (Live in Water = sometimes) → Amphibians

Name	Blood Type	Give Birth	Can Fly	Live in Water	Class
lemur	warm	yes	no	no	?
turtle	cold	no	no	sometimes	?
dogfish shark	cold	yes	no	yes	?

A lemur triggers rule R3, so it is classified as a mammal
A turtle triggers both R4 and R5
A dogfish shark triggers none of the rules

Figure 4.2: Working procedure of a rule-based classifier

4.2.1.3 KNN

K nearest neighbors (KNN) [47][71] is a simple algorithm that stores all available cases and classifies new cases based on a similarity measure (e.g., distance functions). KNN is a non-parametric, lazy learning algorithm. Its purpose is to use a database in which the data points are separated into several classes to predict the classification of a new sample point. How closely out-of-sample features resemble the training set determines how accurately the algorithm classifies a given data point [47]. When KNN is used for classification - the output is a class membership (predicts a class—a discrete value). An object is classified by a majority vote of its neighbors, with the object being assigned to the class most common among its k nearest neighbors.

Suppose there is a dataset with n classified examples. Each classified example acts as a point in the feature space. A way to calculate the k-nearest neighbors for unclassified examples would be to find the k already classified examples that are closest to the unclassified data. Once the k neighbors have been identified, a majority class vote will take place among them to classify the new instances.

Figure 4.3 shows a graphical representation of KNN classification. The test sample (green circle) should be classified either to the first class of blue squares or to the second class of red triangles. If k=3 (solid line circle), it is assigned to the second class because there are 2 triangles and only 1 square inside the inner circle.

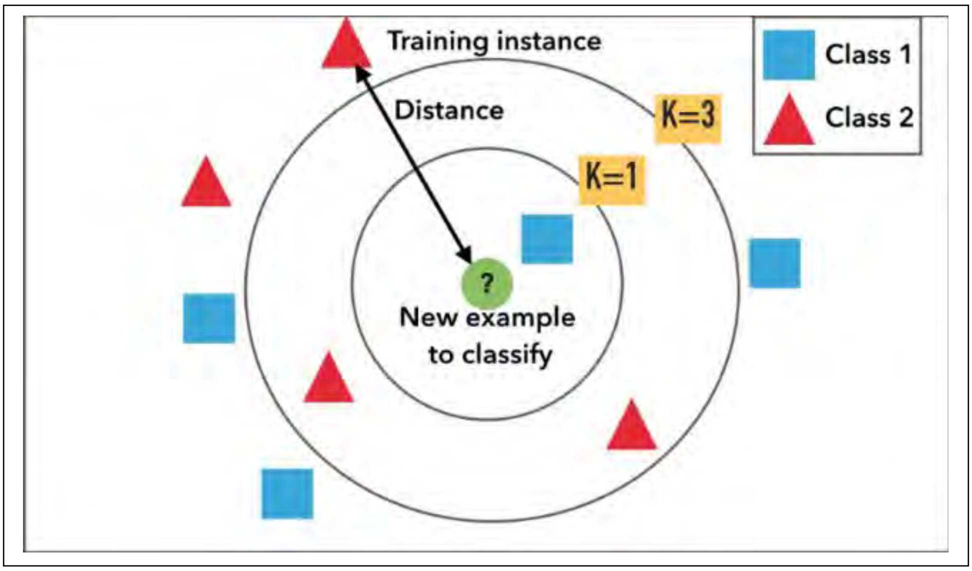


Figure 4.3: Example of KNN *classification method (Source: [71])*

4.2.2 Parameters of SVM based classifier

As the previous chapter have discussed, most of the algorithms in literature use radial basis function (rbf) kernel of SVM for protein subcellular localization. Mathematical details of the SVM with rbf kernel are shown in equation 4.1:

$$f(x) = \sum \alpha_i \, y_i e^{(-\frac{\|x_i - x_j\|^2}{2\sigma^2})} + b \ldots (4.1)$$

Where α_i is Lagrangian multiplier, $\|x_i - x_j\|^2$ is the squared Euclidean distance between the two feature vectors and b is the bias. The σ is the sigma parameter of SVM which indicates the distance between closest points of different classes. The cost function C is defined as: $C \geq \alpha_i \geq 0$.

A brief discussion on these parameters is given below. The range of values for these parameters, the selected values for the proposed algorithm and reasons behind selecting them have been discussed in the next segments.

Cost (C): Cost of constraints violation (default: 1). It is the ‗C'-constant of the regularization term in the Lagrange formulation. The *C* parameter is a regularization/slack parameter. The tuning parameter C which is claimed to be "the price of the misclassification" is exactly the weight for penalizing the "soft margin". Too much "misclassification" cannot be allowed. Smaller values of C force the weights to be small. The larger it gets, the allowed range of weights gets wider. Resultantly, larger *C* values increase the penalty for misclassification and thus reduce the classification error rate on the training data (which leads to over-fitting).

Sigma (sigma): In case of a probabilistic regression model, sigma is the scale parameter [72] of the hypothesized (zero-mean) Laplace distribution estimated by maximum likelihood. Sigma is, as usually defined in a Gaussian distribution, is standard deviation. It determines the width for Gaussian distribution.

The behavior of the model is very sensitive to the sigma parameter. If sigma is too small, the radius of the area of influence of the support vectors only includes the support vector itself and no amount of regularization with C will be able to prevent over fitting. When sigma is very large, the model is too constrained and cannot capture the complexity or ―shape" of the data. The region of influence of any selected support vector would include the whole training set. The resulting model will behave similarly to a linear model with a set of hyperplanes that separate the centers of high density of any pair of two classes.

4.2.3 Optimization of the SVM model parameters

Some problems always arise when researchers need to choose the right sigma and C values for rbf SVMs [73] [74]. Choosing an optimum sigma and C value [75] is very essential for accuracy maximization. For that a common practice is to do a grid search using different value pairs of sigma and C, which helps to find the optimal value pair. Here optimal pair of sigma and C does not mean the minimum value pair. Rather it should be the pair which has comparatively smaller error rate, higher sigma value and lower C value.

Some notable properties between C and Sigma are:

a) With too low value of C (<1) error rate increases with higher value of sigma

b) Large values of C counter balance the bias introduced by large sigma

c) Very small value of sigma may result in good training data error rate, but won‘t be useful in the case of test data recognition error rate

d) With large value of sigma the Gaussian kernel of rbf becomes almost linear

e) A stable(where error rate is almost non fluctuating) region of sigma and C values should be searched (from graph if possible) for better test data recognition accuracy

f) Number of SVs (support vectors) is not a reliable way of determining the ‗goodness‘ of the classifier

4.3 Implementation environment

For the implementation of this algorithm, the RStudio 1.0.44 has been used. MATLAB 2013 has been used for simulation of the MATLAB code for counting highest, 2nd highest, 3rd highest amino acid, total negatively charged residues, total positively charged residues, The amino acid pair present in highest number, The amino acid pair present in 2nd highest number, The amino acid pair present in 3rd highest number, Location of the highest amino acid in the string, Location of the 2nd highest amino acid in the string, Location of the 3rd highest amino acid in the string of the protein sequences. Code Blocks 13.12 (Integrated Development Environment for implementation and simulation of C codes) has been used to simulate the code of SSAADR algorithm during sequence alignment.

4.4 Training and testing of different classifier models

The dataset has randomly been split into train and test set during simulation of the algorithm. A 5-fold cross validation technique [48] with 20 repeats in each iteration has been adopted here. It randomly divides the dataset into 5 sets, uses 4 set to train the model and the remaining set tests the performance of the models, thus it ensures bias-free performance outcome of the models. They also need to be tested with cross validation techniques of more than 5-folds. But too much fold is not feasible for a small dataset considered here, and shows lower performance. So to get optimum result, 5-fold cross validation has been selected for this work. After selecting the classifier, the training samples will first be fed to the developed model for learning the features of protein sequences. After that, the test set will be used to evaluate the trained model. The steps required for the simulation of the algorithm in R are shown in the next flowchart (figure 4.4).

4.5 Performance evaluation metrics

As the dataset has 6 different locations of a cell, each location was labeled as a class. An evaluation of the proposed algorithm has been made to check the classification performance by computing the confusion matrix, considering the class labels of the training dataset as the reference and the predicted class labels as the outcome of the proposed algorithm. The output of SVM should remain in any of the four categories in the confusion matrix: True Positive (TP), True Negative (TN), False Positive (FP) and False Negative (FN). Then, these values were used to calculate the sensitivity, specificity and accuracy of the classification.

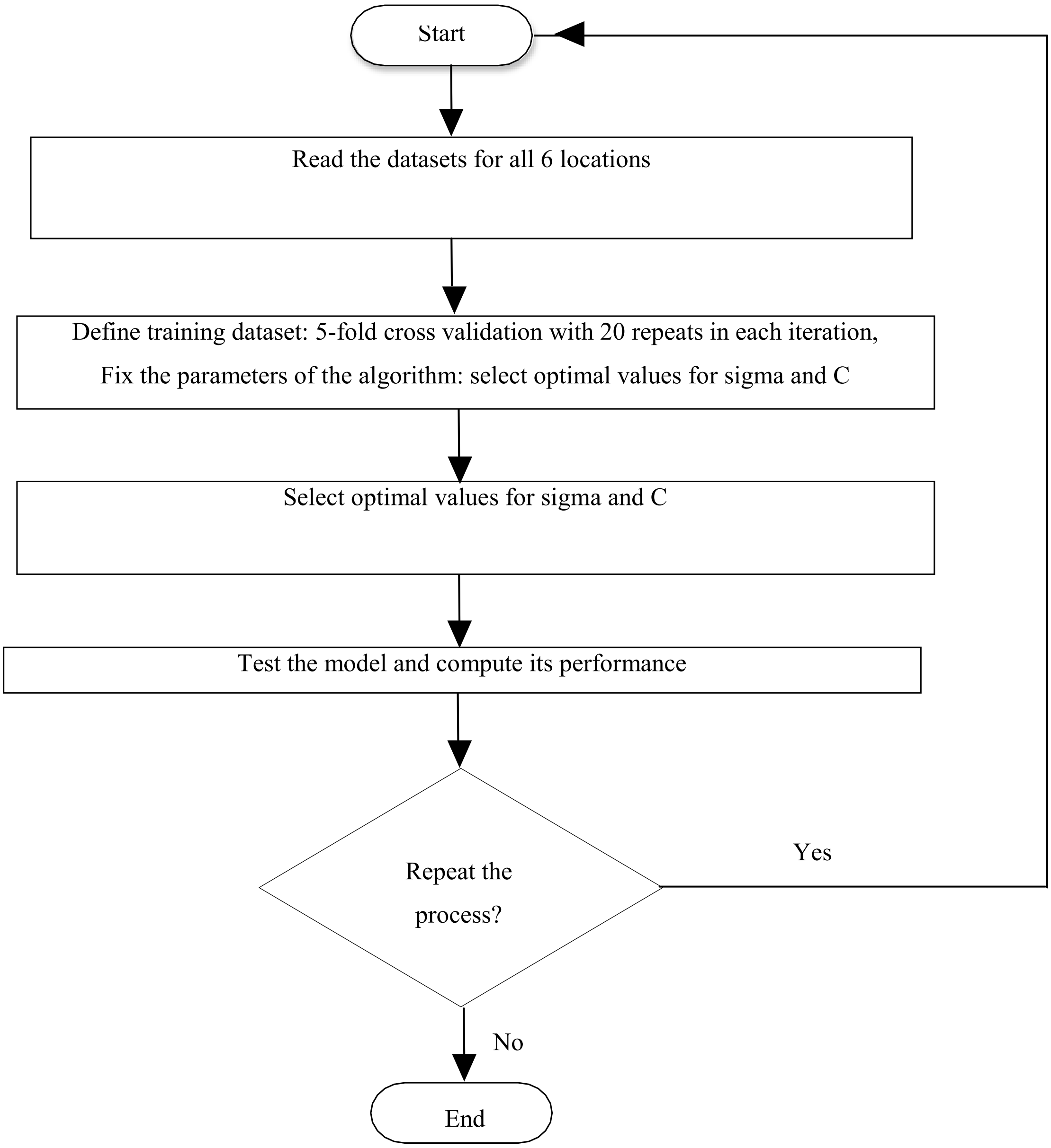

Start
Read the datasets for all 6 locations
Define training dataset: 5-fold cross validation with 20 repeats in each iteration,
Fix the parameters of the algorithm: select optimal values for sigma and C
Select optimal values for sigma and C
Test the model and compute its performance
Repeat the process?
Yes
No
End

Figure 4.4: Steps for simulation of the algorithm

From a confusion matrix of true and hypothesized class the required values can be calculated as:

P = Total number of samples classified as positive

N = Total number of samples classified as negative

True Positive (TP) = Correctly classified samples in the positive class

True Negative (TN) = Correctly classified samples in the negative class

False Positive (FP) = Samples of the Negative class incorrectly classified as positive

False Negative (FN) = Samples of the Positive class incorrectly classified as negative

And the measures used to evaluate the method stated in this report are as follows:

Sensitivity: The sensitivity is the ratio of the correctly classified samples in the positive class over the entire set of positive instances. It is calculated as:

$$\text{Sensitivity} = \frac{TP}{P} \ldots(4.2)$$

Specificity: The specificity is the ratio of the correctly classified samples in the negative class over the entire set of negative instances. It is calculated as:

$$\text{Specificity} = \frac{TN}{N} \ldots(4.3)$$

Overall accuracy: Overall accuracy is the percentages of the samples that are correctly classified as either positive or negative classes; that is, the sum of the true positives plus true negatives divided by the total number of individuals tested. It is calculated as:

$$\text{Overall accuracy} = \frac{(TP+TN)}{(P+N)} \ldots(4.4)$$

Summary of chapter 4:

Chapter 4 discusses on different aspects of selecting the classification algorithms. The one that outperforms others for the dataset considered here will be selected and used to develop the proposed algorithm. The selected classifier and results of the developed model will be discussed in chapter 5.

CHAPTER 5

Experimental results

5.1 Performance evaluation of the classification algorithms

Different classification algorithms have been discussed in the previous chapter. Of them, SVM has gained the prime interest for its superior performance in literature. Now all these algorithms need to be tested with the dataset into consideration. Their performance on this dataset will be discussed in the next segments.

5.1.1 Classifier 1: SVM

Table 5.1 lists the performance of various SVM kernels when applied to the protein dataset. It is evident from this table that SVM with rbf kernel (with optimum parameter set) is appropriate for the dataset into consideration.

Table 5.1: Comparing performance of different SVM kernels on the same dataset

Method	Parameters	Optimum result
svmRadial (SVM)	Sigma (sigma), Cost (C)	for sigma<0.7,C=1, 89%<accuracy< 95%
svmLinear (SVM)	Cost (C)	for C=1, accuracy <78%
svmPoly (SVM)	Polynomial Degree (degree), Scale (scale), Cost (C)	for degree=3, scale=0, C=1, accuracy <30%

5.1.2 Classifier 2 and 3: KNN and PART

In order to prove the superiority of proposed SVM based method with other well-known classification techniques, a rule based classification technique and k-nearest neighbor methods have been applied to the protein dataset. The results are listed in table 5.2, and they indicate that the proposed SVM based technique with rbf kernel outperforms the other methods.

Table 5.2: Comparison of different classification methods on the same dataset

Method	Parameters	Optimum result
svmRadial (SVM)	Sigma(sigma), Cost (C)	for sigma<0.7,C=1, 89%<accuracy< 95%

PART (rule based classifier)	Confidence threshold (threshold),
	Confidence threshold (pruned)
		For
		threshold = 0.01,
		pruned = yes, accuracy=86%

knn (k nearest neighbor)	Number of neighbors (k)	For k=2 accuracy~= 70%

5.1.3 Selection of classifier and the optimum parameter values for the proposed algorithm

When compared to other classification techniques, SVM with rbf kernel has proved itself to be the best. The results in table 5.1 and 5.2 indicate that the SVM based classifier with rbf kernel outperforms the other methods for the dataset in consideration. So it has been selected to be used in the proposed algorithm.

The regularization parameter, C, which is known as the cost of SVM is chosen to trade-off between norm (complexity/smoothness/capacity) and loss (error penalization). Hence, if cost C increases the effect of the regularization decreases and the SVM tends to over fit the data. In order to avoid the over fitting of the decision boundaries of SVM, the cost C and σ parameter have been optimized while maximizing classification accuracies during the current work.

Numerous iterations of SVM model parameters against the classification outcome of the proposed algorithm have been helpful to select the optimum range of the parameters to be: cost (C) = 1 and sigma < 0.7. The result shows improvement with an increase in cost, but it will also increase the complexity of computations incorporating over fitting of data points. So in order to minimize complexity and error rate, the limits for SVM parameters have been chosen to be as such. Here a grid search method was adopted to find out the optimal value pair. Also, a 5-fold cross validation technique [48] with 20 repeats in each iteration has been incorporated.

Effects of different sigma and C parameters on the proposed algorithm (of trained model with training dataset) with SVM classifier incorporating rbf kernel is shown in figure 5.1, figure 5.2 and figure 5.3.

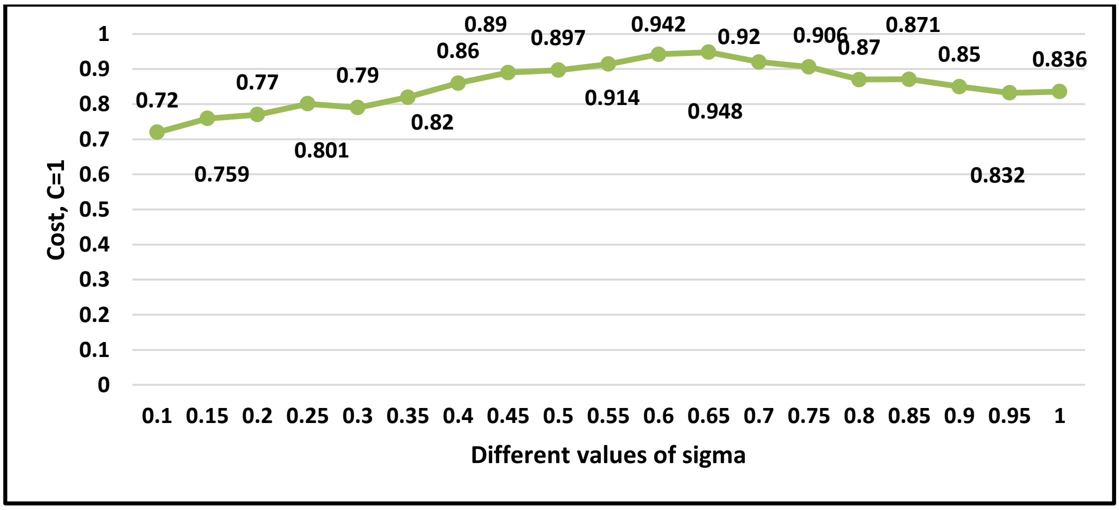


Figure 5.1: Change in accuracy of the trained model due to various values of sigma parameter, while keeping a constant value of C=1

Figure 5.1 depicts that, at sigma=0.6 and C=1 the model achieves the highest accuracy (considering the plotted data points). Increase in sigma lowers that accuracy (accuracy becomes inconsistent too), telling that higher values of sigma are over fitting the model. So its performance is inconsistent after the point of sigma=0.6. As the higher values of sigma are not giving any satisfactory result, the optimum sigma should be 0.6 (considering data points of figure 5.1).

But this optimum sigma value shows different accuracy while experimenting with different values of C (figure 5.2 and figure 5.3). The accuracy is higher around the point of sigma=0.6. But the inconsistent accuracies over these graphs are signs of over fitting that needs to be avoided.

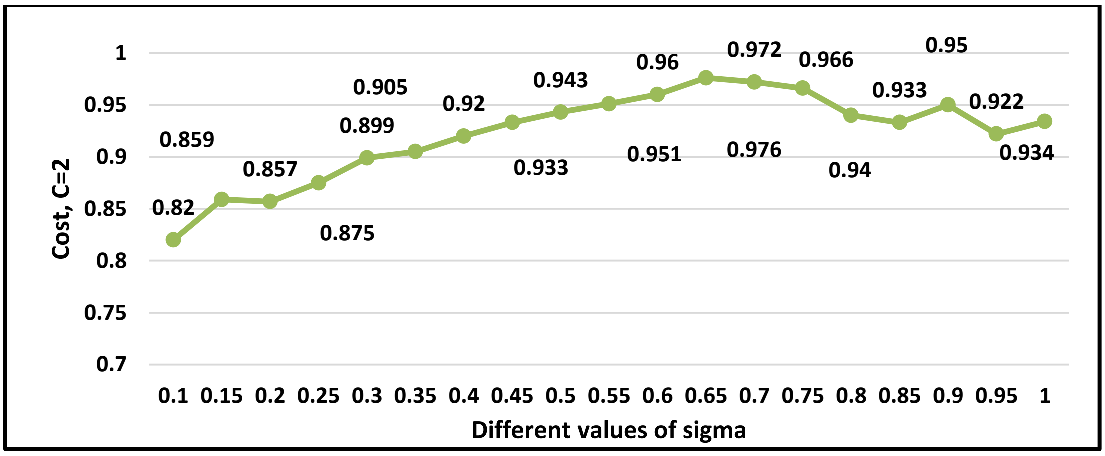


Figure 5.2: Change in accuracy of the trained model due to various values of sigma parameter, while keeping a constant value of C=2

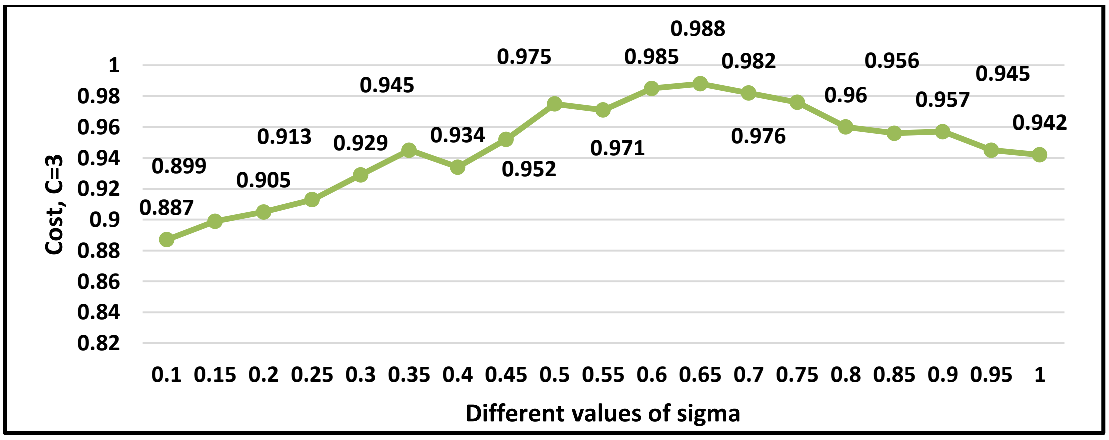


Figure 5.3: Change in accuracy of the trained model due to various values of sigma parameter, while keeping a constant value of C=3

Actually, higher values of sigma try to capture the properties of every data point from the training set which leads to over fitting of the model. Again, high values of C try to minimize the amount of misclassification and in turn, over fit the model. So both C and sigma are important for SVM classifier with rbf kernel and should be optimized carefully. Thus, the choice of sigma<0.7 and C=1 as optimum parameter set for the considered dataset of this thesis work is justified.

Figure 5.4 shows the accuracy the proposed algorithm achieved after testing for different sigma and cost pairs. It is to be noted from figure 5.4 that for sigma<0.7, the accuracy can be improved from 90% to 94%. However to increase this accuracy, the sigma and cost parameters needs to be risen which increases the possibility of over fitting of the training model, which has already been discussed. The drop in accuracy at sigma-0.7 and C=1 tells that model has already reached its highest accuracy point, and now its performance may be inconsistent due to various possible reasons discussed above.

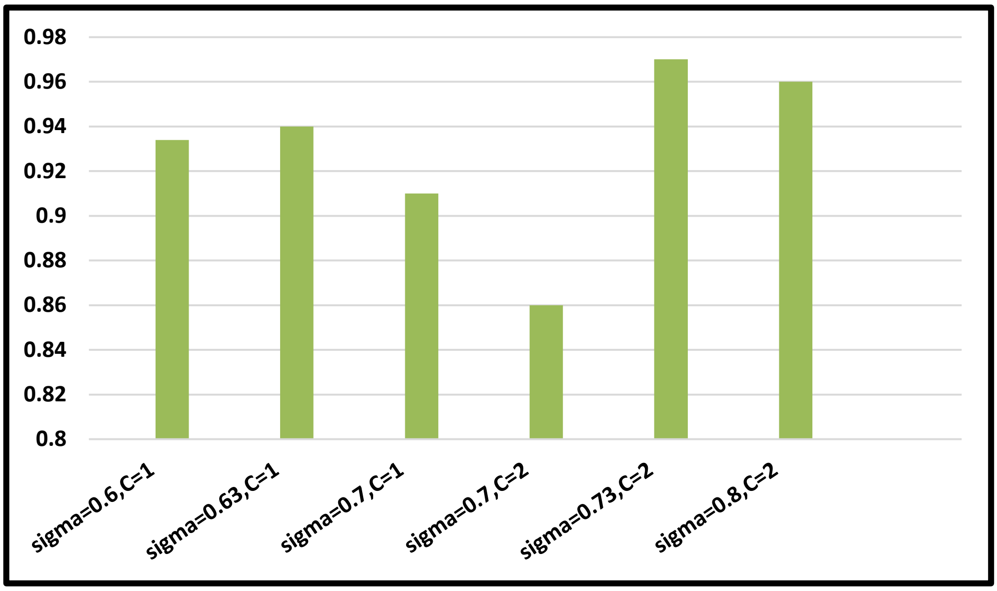


Figure 5.4: Accuracy of the proposed algorithm after testing with different sigma and cost pair

5.2 Selection of features

Various feature selection techniques (both statistical and wrapper methods) have been used here to find out the most significant feature set for the dataset into consideration. The statistical tests include a pair wise t-test and one-way ANOVA, which work on the feature set without considering the actual dataset. The wrapper methods are forward feature selection and backward feature elimination test, which work on the feature set taking into account the actual data model. The details of these techniques are stated below.

5.2.1 Pair wise t-test and ANOVA

The p-values found from a pair wise t-test have been put down into table 5.3. The values in red color are not <0.05, hence the corresponding feature pairs are not significant. Because the significance level for this test was chosen to be less than 0.05.

Table 5.3: Result of pair wise t-test

Features↓	**Subcellular locations↓→**	**Chromatin**	**Nuclear Speckles**	**Nucleolus**	**Nucleoplasm**	**Nuclear Lamina**	**PML Body**
Highest amino acid count	Chromatin	X	0.0001	0.5545	0.0003	0.3723	0.0091
	Nuclear Speckles	0.0001	X	0.0001	0.0009	0.0007	0.0048
	Nucleolus	0.5545	0.0001	X	0.1132	0.0916	0.0382
	Nucleoplasm	0.0003	0.0009	0.1132	X	0.9437	0.6277
	Nuclear Lamina	0.3723	0.0007	0.0916	0.9437	X	0.6215
	PML Body	0.0091	0.0048	0.0382	0.6277	0.6215	X
2nd highest amino acid count	Chromatin	X	0.0196	0.8277	0.7492	0.9776	0.3115
	Nuclear speckles	0.0196	X	0.0033	0.0140	0.0166	0.8173

Features↓	Subcellular locations↓→	Chromatin	Nuclear Speckles	Nucleolus	Nucleoplasm	Nuclear Lamina	PML Body
	Nucleolus	0.8277	0.0033	X	0.5168	0.7881	0.1016
	Nucleoplasm	0.7492	0.0140	0.5168	X	0.7641	0.2119
	Nuclear Lamina	0.9776	0.0166	0.7881	0.7641	X	0.2476
	PML Body	0.3115	0.8173	0.1016	0.2119	0.2476	X
3rd highest amino acid count	Chromatin	X	0.5530	0.1473	0.4012	0.5225	1.00
	Nuclear Speckles	0.5530	X	0.0144	0.7655	0.1542	0.5528
	Nucleolus	0.1473	0.0144	X	0.0035	0.3912	0.1472
	Nucleoplasm	0.4012	0.7655	0.0035	X	0.0832	0.4011
	Nuclear Lamina	0.5225	0.1542	0.3912	0.0832	X	0.5223
	PML Body	1.00	0.5528	0.1472	0.4011	0.5223	X
Theoretical pI	Chromatin	X	0.6665	0.8160	0.0043	0.0007	0.0001
	Nuclear Speckles	0.6665	X	0.8005	0.0008	0.0002	0.0001
	Nucleolus	0.8160	0.8005	X	0.0003	0.0001	0.0001
	Nucleoplasm	0.0043	0.0008	0.0003	X	0.4713	0.1205
	Nuclear Lamina	0.0007	0.0002	0.0001	0.4713	X	0.3726
	PML Body	0.0001	0.0001	0.0001	0.1205	0.3726	X
Instability index	Chromatin	X	0.0904	0.0869	0.2651	0.4417	0.9980
	Nuclear Speckles	0.0904	X	0.0003	0.1804	0.0173	0.0955

Features↓	Subcellular locations↓→	Chromatin	Nuclear Speckles	Nucleolus	Nucleoplasm	Nuclear Lamina	PML Body
	Nucleolus	0.0869	0.0904	X	0.0009	0.2858	0.0958
	Nucleoplasm	0.2651	0.1804	0.0009	X	0.0476	0.2817
	Nuclear Lamina	0.4417	0.0173	0.2858	0.0476	X	0.4671
	PML Body	0.9980	0.0955	0.0958	0.2817	0.4671	X
Aliphatic index	Chromatin	X	0.0001	0.5545	0.0003	0.8044	0.0945
	Nuclear Speckles	0.0001	X	0.0001	0.0041	0.0001	0.0001
	Nucleolus	0.5545	0.0001	X	0.0223	0.3090	0.7617
	Nucleoplasm	0.0003	0.0041	0.0223	X	0.0041	0.1119
	Nuclear Lamina	0.8044	0.0001	0.3090	0.0041	X	0.2433
	PML Body	0.0945	0.0001	0.7617	0.1119	0.2433	X
Molecular weight	Chromatin	X	0.0001	0.0001	0.0004	0.8403	00001
	Nuclear Speckles	0.0001	X	0.7044	0.4106	0.0312	0.8000
	Nucleolus	0.0001	0.7044	X	0.6170	0.0115	0.5950
	Nucleoplasm	0.0004	0.4106	0.6170	X	0.0460	0.3633
	Nuclear Lamina	0.8403	0.0312	0.0115	0.0460	X	0.0613
	PML Body	00001	0.8000	0.5950	0.3633	0.0613	X
Alignment	Chromatin	X	0.0001	0.0001	0.0001	0.0001	0.0001

Features↓	Subcellular locations↓→	Chromatin	Nuclear Speckles	Nucleolus	Nucleoplasm	Nuclear Lamina	PML Body
score	Nuclear Speckles	0.0001	X	0.0001	0.0001	0.0001	0.0001
	Nucleolus	0.0001	0.0001	X	0.0001	0.0001	0.0001
	Nucleoplasm	0.0001	0.0001	0.0001	X	0.0001	0.0001
	Nuclear Lamina	0.0001	0.0001	0.0001	0.0001	X	0.0001
	PML Body	0.0001	0.0001	0.0001	0.0001	0.0001	X
Hydropathic -ity	Chromatin	X	0.0015	0.1648	0.4003	0.0092	0.1691
	Nuclear Speckles	0.0015	X	0.0001	0.0002	0.0001	0.0003
	Nucleolus	0.1648	0.0001	X	0.2230	0.0384	0.4591
	Nucleoplasm	0.4003	0.0002	0.2230	X	0.0149	0.3021
	Nuclear Lamina	0.0092	0.0001	0.0384	0.0149	X	0.0585
	PML Body	0.1691	0.0003	0.4591	0.3021	0.0585	X
Total negatively charged residues	Chromatin	X	0.0001	0.0001	0.0008	0.3402	0.0001
	Nuclear Speckles	0.0001	X	0.4453	0.3868	0.0324	0.3283
	Nucleolus	0.0001	0.4453	X	0.4306	0.0106	0.3075

Features↓	Subcellular locations↓→	Chromatin	Nuclear Speckles	Nucleolus	Nucleoplasm	Nuclear Lamina	PML Body
	Nucleoplasm	0.0008	0.3868	0.4306	X	0.0262	0.2701
	Nuclear Lamina	0.3402	0.0324	0.0106	0.0262	X	0.0481
	PML Body	0.0001	0.3283	0.3075	0.2701	0.0481	X
Total positively charged residues	Chromatin	X	0.0001	0.0000	0.0001	0.3744	0.0001
	Nuclear Speckles	0.0001	X	0.4697	0.1820	0.0738	0.0262
	Nucleolus	0.0000	0.4697	X	0.1835	0.0312	0.0437
	Nucleoplasm	0.0001	0.1820	0.1835	X	0.0201	0.1747
	Nuclear Lamina	0.3744	0.0738	0.0312	0.0201	X	0.0280
	PML Body	0.0001	0.0262	0.0437	0.1747	0.0280	X
N terminal amino acid	Chromatin	X	0.0001	0.0001	0.0001	0.0001	0.0001
	Nuclear Speckles	0.0001	X	0.0001	0.0001	0.0001	0.0001
	Nucleolus	0.0001	0.0001	X	0.0001	0.0001	0.0001
	Nucleoplasm	0.0001	0.0001	0.0001	X	0.0001	0.0001

Features↓	**Subcellular locations↓→**	**Chromatin**	**Nuclear Speckles**	**Nucleolus**	**Nucleoplasm**	**Nuclear Lamina**	**PML Body**
	Nuclear Lamina	0.0001	0.0001	0.0001	0.0001	X	0.0001
	PML Body	0.0001	0.0001	0.0001	0.0001	0.0001	X
The amino acid pair present in highest number	Chromatin	X	0.0001	0.0001	0.4645	0.0073	0.0001
	Nuclear Speckles	0.0001	X	0.0008	0.0001	0.0586	0.0001
	Nucleolus	0.0001	0.0008	X	0.0001	0.0001	0.0001
	Nucleoplasm	0.4645	0.0001	0.0001	X	0.0006	0.0001
	Nuclear Lamina	0.0073	0.0586	0.0001	0.0006	X	0.0001
	PML Body	0.0001	0.0001	0.0001	0.0001	0.5	X
The amino acid pair present in 2nd highest number	Chromatin	X	0.0001	0.0001	0.0963	0.0005	0.0001
	Nuclear Speckles	0.0001	X	0.02251	0.0001	0.000346	0.0001
	Nucleolus	0.0001	0.0225	X	0.0001	0.0001	0.3902
	Nucleoplasm	0.0963	0.0001	0.0001	X	0.0026	0.0001
	Nuclear Lamina	0.0005	0.0003	0.0001	0.0026	X	0.0001

Features↓	Subcellular locations↓→	Chromatin	Nuclear Speckles	Nucleolus	Nucleoplasm	Nuclear Lamina	PML Body
	PML Body	0.0001	0.0359	0.3902	0.0001	0.0001	X
The amino acid pair present in 3rd highest number	Chromatin	X	0.2539	0.0001	0.0010	0.1907	0.3433
	Nuclear Speckles	0.2539	X	0.0001	0.0010	0.1907	0.3433
	Nucleolus	0.0001	0.0001	X	0.0001	0.0001	0.0001
	Nucleoplasm	0.0010	0.0001	0.0001	X	0.0001	0.0003
	Nuclear Lamina	0.1907	0.4043	0.0001	0.0001	X	0.0386
	PML Body	0.3433	0.0825	0.0001	0.0003	0.0386	X
Location of the highest amino acid in the string	Chromatin	X	0.0001	0.0001	0.0001	0.0001	0.0001
	Nuclear Speckles	0.0001	X	0.0001	0.0001	0.0001	0.0001
	Nucleolus	0.0001	0.0001	X	0.0001	0.0001	0.0001
	Nucleoplasm	0.0001	0.0001	0.0001	X	0.0001	0.0001
	Nuclear Lamina	0.0001	0.0001	0.0001	0.0001	X	0.0001
	PML Body	0.0001	0.0001	0.0001	0.0001	0.0001	X
Location of	Chromatin	X	0.0001	0.0001	0.0001	0.0001	0.0001

Features↓	**Subcellular locations↓→**	**Chromatin**	**Nuclear Speckles**	**Nucleolus**	**Nucleoplasm**	**Nuclear Lamina**	**PML Body**
the 2nd highest amino acid in the string	Nuclear Speckles	0.0001	X	0.0001	0.0001	0.0001	0.0001
	Nucleolus	0.0001	0.0001	X	0.0001	0.0001	0.0001
	Nucleoplasm	0.0001	0.0001	0.0001	X	0.0001	0.0001
	Nuclear Lamina	0.0001	0.0001	0.0001	0.0001	X	0.0001
	PML Body	0.0001	0.0001	0.0001	0.0001	0.0001	X
Location of the 3rd highest amino acid in the string	Chromatin	X	0.0001	0.0001	0.0001	0.0001	0.0001
	Nuclear Speckles	0.0001	X	0.0001	0.0001	0.0001	0.0001
	Nucleolus	0.0001	0.0001	X	0.0001	0.0001	0.0001
	Nucleoplasm	0.0001	0.0001	0.0001	X	0.0001	0.0001
	Nuclear Lamina	0.0001	0.0001	0.0001	0.0001	X	0.0001
	PML Body	0.0001	0.0001	0.0001	0.0001	0.0001	X

A ―multiple t-test” has been incorporated here that compares among several groups of data. But the problem arises when the number of groups grows, and the number of needed pair comparisons grows quickly with it. Then the process gets complicated (as shown in table 5.3). Also, results are insignificant as only two groups of data is never enough to show significant comparison result, when the total dataset consists of so many features.

But this result differs a lot when a one-way ANOVA (with significance level 0.05) has been applied on this dataset, which significantly proves that the features altogether make strong influence in classifying proteins.

One-way ANOVA looks at differences among groups of dataset on some variable of interest. For this reason, a one-way ANOVA test has been conducted on the protein dataset of this thesis work in order to make comparison among 18 features of 6 different locations/groups. The result of this test is shown in the following table (table 5.4). Here, p value<0.05 and F value> 2.239 (F critical, F_c=2.239 for this dataset) depicts that all considered features are significant for the classification of the protein dataset.

Table 5.4: Result of one-way ANOVA test

Features	**Locations**	**Mean**	**F Value**	**p Value**
Highest amino acid count	Location 1	11.263	**6.211**	**0.000015**
	Location 2	14.071		
	Location 3	9.931		
	Location 4	11.053		
	Location 5	11.113		
	Location 6	1.5		

Features	Locations	Mean	F Value	p Value
2nd highest amino acid count	Location 1	10.158	**1.848**	**0.1028**
	Location 2	12.75		
	Location 3	9.931		
	Location 4	10.467		
	Location 5	10.189		
	Location 6	13.368		
3rd highest amino acid	Location 1	10.158	**2.36**	**0.0398**
	Location 2	10.75		
	Location 3	8.594		
	Location 4	10.987		
	Location 5	9.415		
	Location 6	10.158		
Theoretical pI	Location 1	7.877	**9.0076**	**0.0015**
	Location 2	8.038		
	Location 3	7.956		
	Location 4	6.398		
	Location 5	6.723		

Features	Locations	Mean	F Value	p Value
	Location 6	6.417		
Instability Index	Location 1	52.644	**4.746**	**0.0003**
	Location 2	58.89		
	Location 3	48.944		
	Location 4	54.996		
	Location 5	51.045		
	Location 6	52.639		
Aliphatic Index	Location 1	71.938	**8.968**	**0.001**
	Location 2	63.263		
	Location 3	76.47		
	Location 4	71.309		
	Location 5	79.011		
	Location 6	75.688		
Molecular Weight	Location 1	112.536	**4.562**	**0.0005**
	Location 2	67.482		
	Location 3	70.77		
	Location 4	75.313		
	Location 5	118.072		

Features	Locations	Mean	F Value	p Value
	Location 6	65.509		
Alignment Score	Location 1	4.947	**51.514**	**0.00001**
	Location 2	7.821		
	Location 3	14.149		
	Location 4	19.36		
	Location 5	27.83		
	Location 6	26.211		
Hydropathicity	Location 1	-0.6232	**8.339**	**0.0075**
	Location 2	-0.8457		
	Location 3	-0.5726		
	Location 4	-0.6088		
	Location 5	-0.4779		
	Location 6	-0.5781		
Total negatively charged residues	Location 1	129.7105	**3.456**	**0.0046**
	Location 2	84.0536		
	Location 3	85.5941		
	Location 4	87.5867		

Features	Locations	Mean	F Value	p Value
	Location 5	146.0189		
	Location 6	79.3421		
Total positively charged residues	Location 1	139.3684	**4.326**	**0.0008**
	Location 2	91.0179		
	Location 3	90.1782		
	Location 4	80.7067		
	Location 5	129.4717		
	Location 6	69.6579		
N terminal amino acid	Location 1	8.5	**19.554**	**0.0075**
	Location 2	13		
	Location 3	12.1089		
	Location 4	13		
	Location 5	13		
	Location 6	13.1842		
The amino acid pair present in highest number	Location 1	776.1316	**50.803**	**0.007**
	Location 2	1327.0179		
	Location 3	1564.6139		

Features	Locations	Mean	F Value	p Value
	Location 4	786.6		
	Location 5	1146.8491		
	Location 6	168.5789		
The amino acid pair present in 2nd highest number	Location 1	296.9737	**39.158**	**0.0007**
	Location 2	1147.8393		
	Location 3	1354.2871		
	Location 4	413.8667		
	Location 5	709.8491		
	Location 6	1387.2105		
The amino acid pair present in 3rd highest number	Location 1	432.2368	**47.372**	**0.00075**
	Location 2	353.7857		
	Location 3	1346.1683		
	Location 4	783.08		
	Location 5	330.717		
	Location 6	474.7105		
Location of the highest amino acid in the string	Location 1	2	**98.127**	**0.007**
	Location 2	4.2857		
	Location 3	7.8812		

Features	Locations	Mean	F Value	p Value
	Location 4	11.04		
	Location 5	16		
	Location 6	11.5789		
Location of the 2nd highest amino acid in the string	Location 1	2.2368	**82.794**	**0.0008**
	Location 2	4.5357		
	Location 3	7.2079		
	Location 4	11.84		
	Location 5	17.0566		
	Location 6	12.3684		
Location of the 3rd highest amino acid in the string	Location 1	2.1316	**87.753**	**0.0005**
	Location 2	4.4643		
	Location 3	8.5941		
	Location 4	9.92		
	Location 5	18.1132		
	Location 6	13.9474		

5.2.2 Wrapper method

A forward feature selection test has been adapted here to find out the significant features of the dataset. This method starts working with an empty feature set and keeps including one feature at a time into it to find out the model accuracy. Thus the dataset has been changed from containing 0 (zero) feature to a total of 18 features. The model accuracies found during this change in the dataset have been shown in the next line graph (figure 5.5).

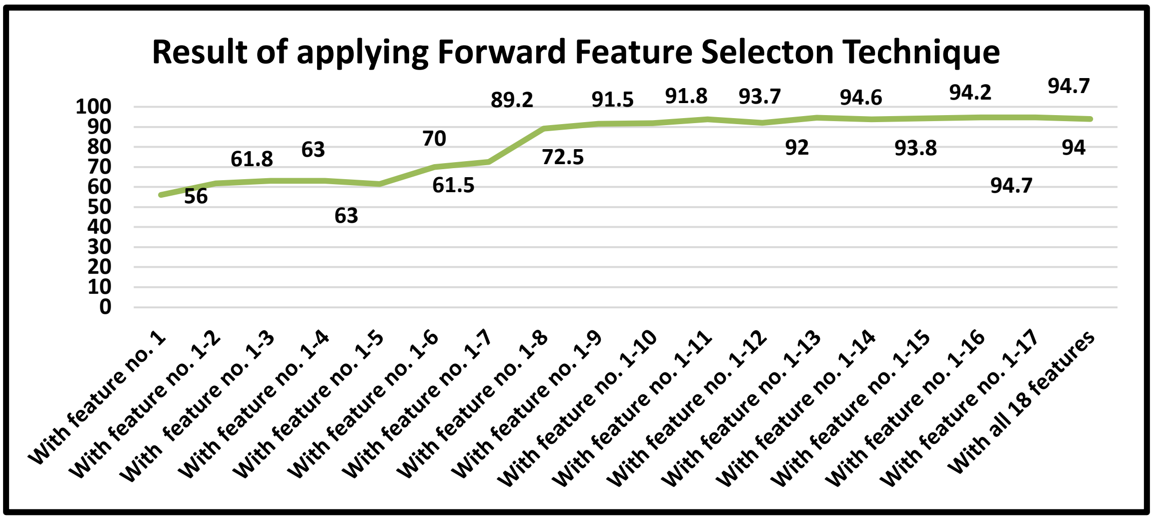


Figure 5.5: Accuracy after applying Forward Feature Selection on dataset

After that, backward feature elimination has been adapted to check the importance of each feature of the dataset. During this technique one feature has been removed from the dataset at a time and then used the dataset to check the model accuracy. Thus the dataset has been changed from containing a total of 18 features to only 1 feature. The model accuracies found during this change in the dataset have been shown in the next line graph (figure 5.6).

It is clear from the above graphs (figure 5.5 and figure 5.6) that all 18 features together are significant for this dataset and classification model. Removing any of them lowers the accuracy of the method. So, the final dataset will consist of all 18 features.

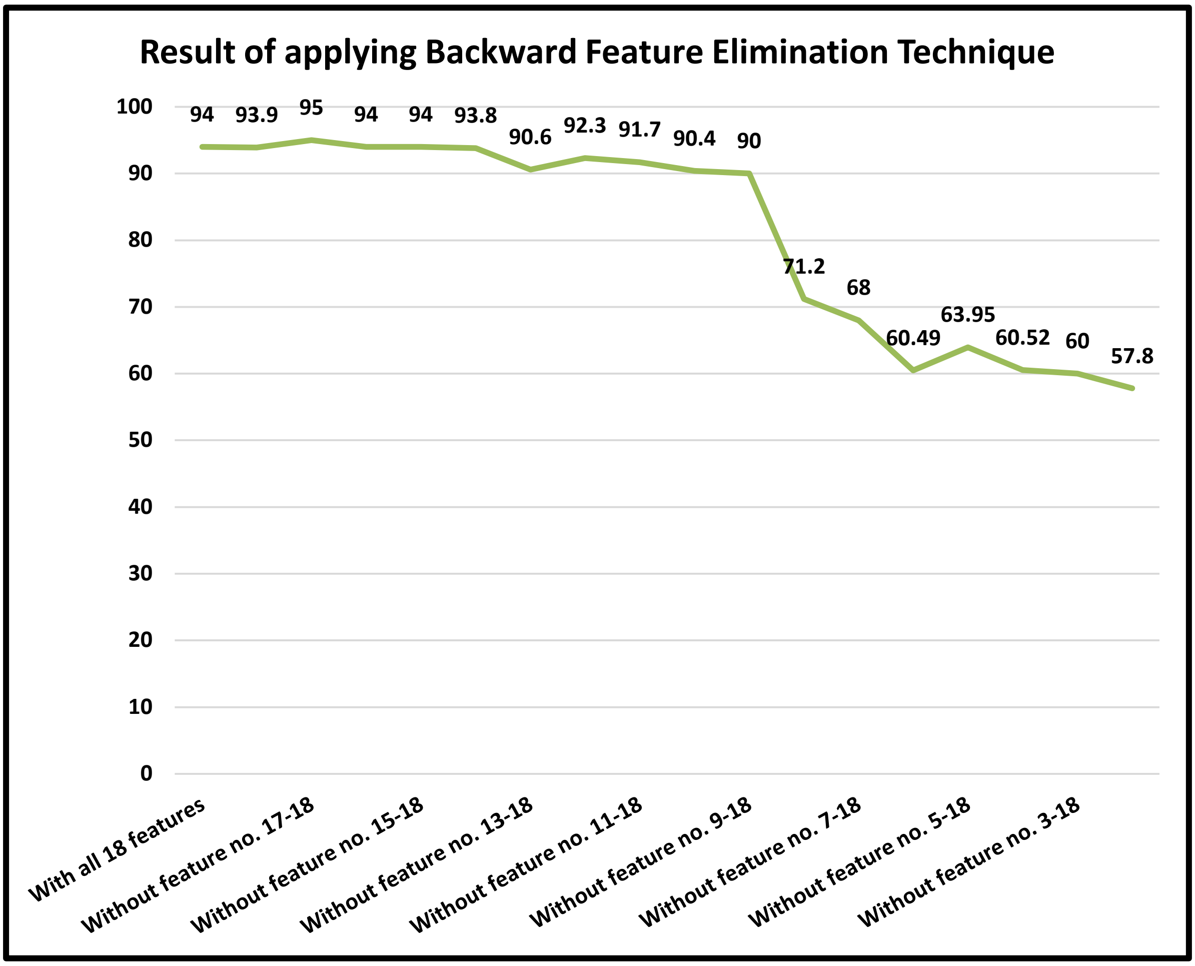


Figure 5.6: Accuracy after applying Backward Feature Elimination on dataset

Count of the highest, 2nd highest and 3rd highest amino acid for each subcellular location are three completely new features introduced in this thesis work. Closer look at figure 5.5and figure 5.6 shows that, without using these features the classifier gets 60.52% accuracy. After incorporating them in the dataset its accuracy is increased by 30%, which proves high significance of these novel features. Also, removing alignment scores from the feature set reduces the overall accuracy by almost 20%. Though this feature is also a new one in protein classification and has been introduced here only, they have proven to be strong predictors for DNA and RNA classification tasks [56][57][58]. Again, the amino acid pair present in highest number, the amino acid pair present in 2nd highest number, the amino acid pair present in 3rd highest number, location of the highest amino acid in the string, location of the 2nd highest amino acid in the string, location of the 3rd highest amino acid in the string are the novel features to be used as predictors for subcellular localization of protein. Including them in the feature set improves the accuracy by 3% and makes it 94% (approx.).

Thus, the final dataset contains all 18 features and the proposed model shows a final accuracy of 94%.

5.3 Performance of the proposed algorithm after feature selection

The final dataset consists of 18 features of 6 different locations. No feature could be removed from it as per the results of all the feature selection process. Figure 5.7 shows the graph the true positive rates and false positive rates computed after several iterations of 5 fold cross-validations with 20 repetitions.

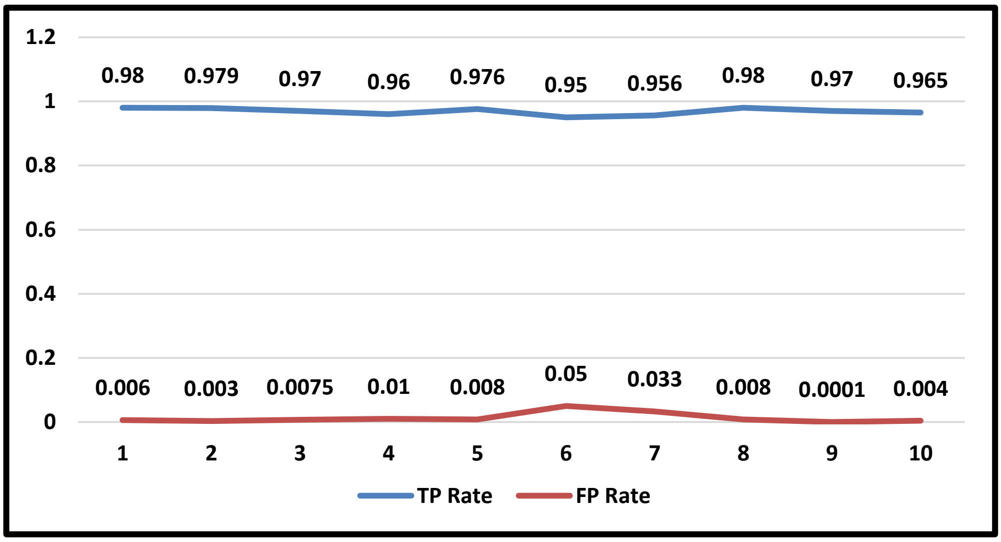


Figure 5.7: True Positive Rates and False Positive Rates for sigma=0.6 and C=1

Here, the values have been selected as:σ = 0.6 and C = 1. Classification performance for this particular parameter set was: sensitivity=0.97, specificity=1 and accuracy=0.94.

5.4 Comparison of the proposed method with similar existing ones

A comparison among the accuracies of the proposed method and other existing methods in literature is presented in table 5.5, and the same is shown using bar diagram later in figure 5.8.

In method 2 [19], new kernel functions used in SVM learning model were introduced for the measurement of sequence similarity. The authors use mapping of k-peptide vectors of protein sequences as a method of searching subcellular or subnuclear locations of the proteins. The k-peptide vectors were first mapped by a matrix of high-scored pairs of k-peptides which are measured by BLOSUM62 scores. The kernels, measuring the similarity for sequences, were then defined on the mapped vectors. By combining all these new encoding methods, a multi-class classification system for the prediction of protein subnuclear localizations is established for the first time. But it has the accuracy of around 50% only. The physicochemical feature (along with complex encoding methods) did not help the system to get an accurate class label for the proteins, and so no further improvement work was attempted for this system.

In method 3 [15], the authors developed an evolutionary support vector machine (ESVM) based classifier with automatic selection from a large set of physicochemical composition (PCC) features of protein to design an accurate system for predicting protein subnuclear localization. ESVM (which uses an inheritable genetic algorithm, IGA, combined with SVM) can automatically determine the best number, m of PCC features and identify m out of 526 PCC features simultaneously. It shows an accuracy of 56.37% only, which is not enough to provide accurate classification results.

Method 4 in [20] presents segmentation distribution and segmented auto-covariance feature extraction technique to explore local evolutionary-based information. It uses consensus sequence-based and semi-occurrence to extract global evolutionary-based information. Use of SVM as the classification technique gives the system an accuracy of 86%, which enhance Gram-positive and Gram-negative subcellular localization prediction accuracies by up to 6.4% (considering previous work). But the proposed techniques require vast calculations and complex computations, which can be solved by the proposed method of this work.

According to method 5 in [22], a set of SVMs can be trained to predict the subcellular location of a given protein based on its amino acid, amino acid pair, and gapped amino acid pair compositions. The predictors based on these different compositions were then combined using a voting scheme. 5-fold cross-validation tests show its accuracy to be 78-79%. Again, huge amount of calculations needed for different amino acid compositions increases computational cost and complexity.

Table 5.5: Comparison of the proposed method with several similar methods

	Methods	**Accuracy**
(1)	Proposed method in this thesis work	90-94%
(2)	An SVM-based system for predicting protein subnuclear Localizations **[19]**	50% (for single localization)
(3)	ProLoc: Prediction of protein subnuclear localization using SVM with automatic selection from physicochemical composition features **[15]**	56.37%

	Methods	Accuracy
(4)	Gram-positive and Gram-negative protein subcellular localization by incorporating evolutionary-based descriptors into Chou's general PseAAC **[20]**	79% - 86%
(5)	Prediction of protein subcellular locations by support vector machines using compositions of amino acids and amino acid pairs **[22]**	78-79%

It is to be noted from figure 5.8 that the proposed method (method 1) outperforms the average of accuracy of similar other methods in method 2 [19], method 3 [15], method 4 [20] and method 5 [22].

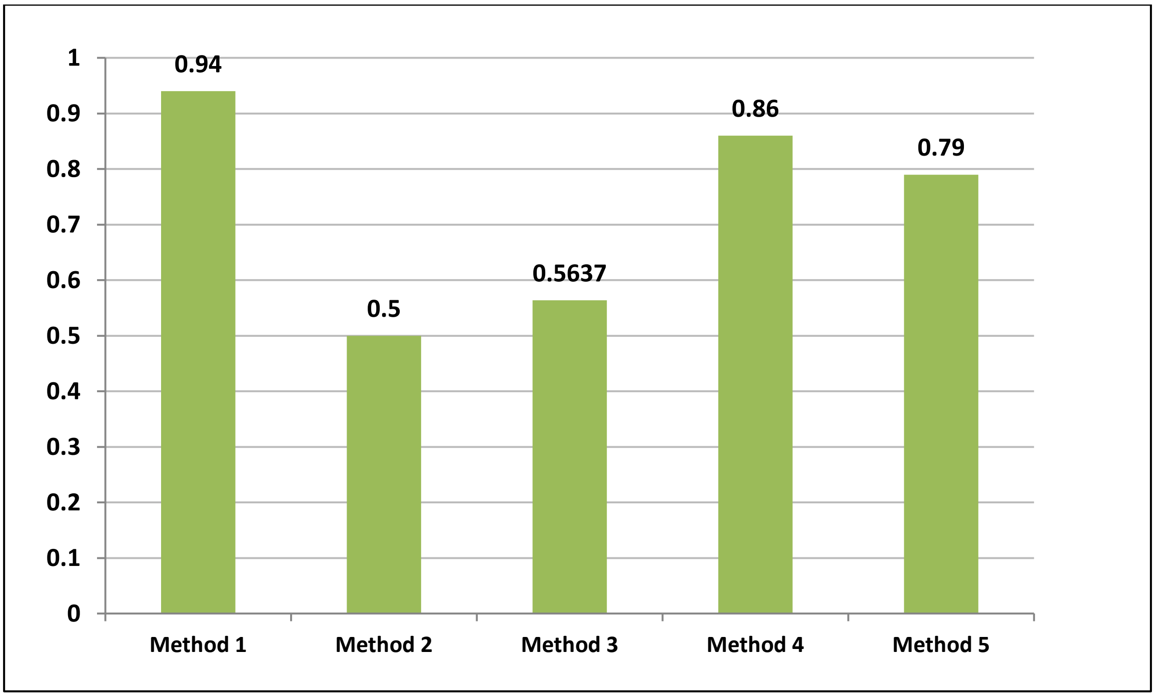


Figure 5.8: Bar diagram showing comparison of accuracy of 5 different methods. Method 1 is the proposed method in this thesis, Method 2 is the SVM-based system for predicting protein subnuclear localizations[19], Method 3 is the SVM based technique with automatic feature selection [15], Method 4 is the Gram-positive and Gram-negative protein subcellular localization by incorporating evolutionary SVM [20] and Method 5 is the SVM based prediction of protein using compositions of amino acids and amino acid pairs[22].

The reason of why it is better to use the proposed method (mentioned as method 1 in figure 5.8) for experimental purposes has already been discussed in the previous chapters. When it is compared to some existing methods (in figure 5.8), it shows higher accuracy than the others. So not in terms of easy access and lower complexity of the features only, the proposed method outperforms the others in resultant accuracy too.

5.5 Discussion

Protein subcellular localization has been an active area of research due to the important role it plays in indicating, if not determining, protein function. A number of efforts have previously used amino acid compositions as well as limited sequence order information in order to predict protein localization. In this thesis work, a novel approach has been made to determine protein subcellular localization based on its multiple physicochemical properties. The above report brings to light a pretty simple method for subcellular localization of protein. While comparing the proposed method with few similar established methods, the proposed one proves itself to have better accuracy with simpler algorithm. The thesis work has been performed according to the following steps:

a) It works with a feature set for subcellular localization of protein sequences found within 6 different areas of a cell. The most common features were extracted from the protein sequences with a view to developing a less complex algorithm. Measures have been taken to validate the feature set.

b) It is to be noted that the first three features, which are completely new for such classification tasks, have shown great impact on the final results. Removing them from the dataset decreases the overall accuracy by 30%, which makes a huge difference in performance. Also, removing alignment scores from the feature set reduces the overall accuracy by almost 20%. Though this feature has not yet been used in other algorithms, it is a proven strong predictor for DNA and RNA classification. Again, adding the last 6 features in the feature set gives the accuracy level a rise and makes it 94% finally.

c) It includes SVM to provide the unknown proteins a class label (proper subcellular location of proteins in this case). It randomly splits the dataset into train and test sets to get unbiased result

d) It checks for optimal values of SVM parameters (sigma and C), and sets them as : sigma<0.7 and C=1

e) The final result is calculated after incorporating a 5-fold cross validation with 20 repeats in each iteration. The overall accuracy of the proposed method is 94%, which is better than similar existing methods.

Summary of chapter 5:

Chapter 5 records the results of the proposed algorithm considering different aspects. Though it performs better than other algorithms, it has some limitations. There is scope for further improvement too. The limitations of the existing algorithm and scopes of improvement will be discussed in the next chapter.

CHAPTER 6
Conclusion and Future Works

6.1 Limitations

Despite its better performance than other similar algorithms, few limitations of the proposed algorithm have been found out during the thesis work. They will help to find out the scopes for further development of the algorithm. A list of these limitations is given below:

a) Using a small sized dataset, along with few features only

b) Not trying out platforms other than R, which could have led towards new discoveries by providing more control over the classifier parameters

c) Some other popular classifiers, such as deep learning with artificial neural networks, could have been tested with

d) Using the method accuracy of other algorithms reported by their authors in the literature, while comparing them with the newly proposed algorithm in this thesis work; as those algorithms could not be implemented due to lack of related information in the literature

6.2 Conclusion

In this thesis work, a novel approach has been made to predict protein subcellular localization based on its multiple physicochemical properties those finally help to develop a simple and less complex algorithm. It uses SVM classifier with rbf kernel and achieves an overall accuracy of 94%. It starts working with a new set of features, which can be easily extracted from protein sequences. It introduces some novel features to be used for the first time as predictors in subcellular localization of proteins; those show impressive results and improve performance of the proposed algorithm. Thus the proposed algorithm proves to be better than other similar algorithms doing protein subcellular localization.

6.3 Future works

There are several areas of future work that can be served to improve the algorithm functionality and provide additional evaluation of its performance. They are:

a) Further improvements can best be obtained by preparing data sets of higher quality. It should be possible to increase the number of data entries from updated databases.

b) Adding new subcellular locations or defining finer classifications will also be important in practical applications of gene annotations and functional predictions. At the same time it would become necessary to consider protein groups that inherently belong to multiple locations, such as those that move between cytoplasm and nucleus under different conditions

c) Other common (physicochemical or not) features of protein, which are easily accessible and require less computations, can be tested with

d) It has been examined with, in a fairly comprehensive way, various SVM kernels and parameters as well as other classification techniques, and it is unlikely that a significant improvement will be obtained by changing training methods alone. So in order to further improve the efficiency and techniques of the algorithm, possible modification in each and every step of the algorithm will be taken into consideration

References

[1] K. Nakai, "Protein Sorting Signals and Prediction of Subcellular Localization," in *Advances in Protein Chemistry*, 2000, pp. 277-344.

[2] " Biology Q&As," [Online]. Available: https://www.biology-questions-and-answers.com/protein-synthesis.html.

[3] J. S. Valastyan and S. Lindquist, "Mechanisms of protein-folding diseases at a glance," *Disease Models & Mechanisms,* vol. 7, no. 1, pp. 9-14, 2014.

[4] G. S. Butler and C. M. Overall, "Proteomic identification of multitasking proteins in unexpected locations complicates drug targeting," *Nature reviews Drug Discovery,* vol. 8, no. 12, pp. 935-948, 2009.

[5] J.-E. Kim, Y. H. Hong, J. Y. Kim, G. S. Jeon, J. . H. Jung, B.-N. Yoon, S.-Y. Son, K.-W. Lee, J.-I. Kim and J.-J. Sung, "Altered nucleocytoplasmic proteome and transcriptome distributions in an in vitro model of amyotrophic lateral sclerosis," *PLOS ONE,* 2017.

[6] X. Wang and S. Li, "Protein mislocalization: Mechanisms, functions and clinical applications in cancer," *Biochimica et Biophysica Acta (BBA) - Reviews on Cancer,* pp. 13-25, 2014.

[7] B. Yu, S. Li, C. Chen, J. Xu, W. Qiu, X. Wu and R. Chen, "Prediction subcellular localization of Gram-negative bacterial proteins by support vector machine using wavelet denoising and Chou's pseudo amino acid composition," *Chemometrics and Intelligent Laboratory Systems,* vol. 167, pp. 102-112, 2017.

[8] X. Cheng, . X. Xiao and K.-C. Chou, "pLoc-mEuk: Predict subcellular localization of multi-label eukaryotic proteins by extracting the key GO information into general PseAAC," *Molecular BioSystems,* no. 9, 2017.

[9] S. Hua and Z. Sun, "Support vector machine approach for protein subcellular localization prediction," *Bioinformatics,* vol. 17, no. 1, pp. 721-728, 2001.

[10] X. Xiao, S. Shao, Y. Ding, Z. Huang and K.-C. Chou, "Using cellular

automata images and pseudo amino acid composition to predict protein subcellular location," *Amino Acids,* vol. 30, no. 1, pp. 49-54, 2006.

[11] G. Dellaire, R. Farrall and W. A. Bickmore, "The Nuclear Protein Database (NPD): sub-nuclear localisation and functional annotation of the nuclear proteome," *Nucleic Acids Research,* vol. 31, no. 1, pp. 328-330, 2003.

[12] T. H. Lin, R. F. Murphy and Z. Bar-Joseph , "Discriminative motif finding for predicting protein subcellular localization," *IEEE/ACM Transaction of Computational Biology and Bioinformatics,* vol. 8, no. 2, pp. 441-451, 2011.

[13] R. N. Kalate, S. S. Tambe and B. D. Kulkarni, "Artificial neural networks for prediction of mycobacterial promoter sequences," *Computational Biology and Chemistry,* vol. 27, no. 6, pp. 555-564, 2003.

[14] X. Xiao, Z.-C. Wu and K.-C. Chou, "A multi-label classifier for predicting the subcellular localization of gram-negative bacterial proteins with both single and multiple sites," *PLoS ONE,* vol. 6, 2011.

[15] W. Huang, C. Tung, H. Huang, S. Hwang and S. Ho, "ProLoc: Prediction of protein subnuclear localization using SVM with automatic selection from physicochemical composition features," *BioSystems,* vol. 90, no. 2, pp. 573-581, 2007.

[16] L. Jensen, R. Gupta, N. Blom, D. Devos, J. Tamames, C. Kesmir, H. Nielsen, H. Staerfeldt, K. Rapacki, C. Workman, C. Andersen, S. Knudsen, A. Krogh, A. Valencia and S. Brunak, "Prediction of human protein function from post-translational modifications and localization features," *Journal of molecular biology,* vol. 319, no. 5, pp. 1257-1265, 2002.

[17] E. Gasteiger, C. Hoogland, A. Gattiker, S. Duvaud, M. R. Wilkins, R. D. Appel and A. Bairoch, "Protein Identification and Analysis Tools on the ExPASy Server," in *The Prteomics Protocols Handbook*, Humana Press, 2005, pp. 571-607.

[18] D. Chen, X. Tian, B. Zhou and J. Gao, "ProFold: Protein Fold Classification with Additional Structural Features and a Novel Ensemble Classifier," *BioMed Research International,* 2016.

[19] Z. Lei and Y. Dai, "An SVM-based system for predicting protein subnuclear localizations," *BMC Bioinformatics,* vol. 6, no. 1, pp. 291-298, 2005.

[20] A. Dehzangi, R. Heffernan, A. Sharma, J. Lyons, K. Paliwal and A. Sattar, "Gram-positive and Gram-negative protein subcellular localization by incorporating evolutionary-based descriptors into Chou's general PseAAC," *Journal of theoretical biology,* vol. 364, pp. 284-294, 2015.

[21] J. Chen , H. Xu , P. A. He, Q. Dai and Y. Yao, "A multiple information fusion method for predicting subcellular locations of two different types of bacterial protein simultaneously," *Biosystems,* vol. 139, pp. 37-45, 2016.

[22] K.-J. Park and M. Kanehisa, "Prediction of protein subcellular locations by support vector machines using compositions of amino acids and amino acid pairs," *Bioinformatics,* vol. 19, no. 13, pp. 1656-1663, 2003.

[23] A. A. Polyansky, M. Hlevnjak and . B. Zagrovic, "Analogue encoding of physicochemical properties of proteins in their cognate messenger RNAs," *Nature Communications,* 2013.

[24] W.-L. Huang, C.-W. Tung, S.-W. Ho, S.-F. Hwang and S.-Y. Ho, "ProLoc-GO: Utilizing informative Gene Ontology terms for sequence-based prediction of protein subcellular localization," *BMC Bioinformatics,* vol. 9, no. 80, 2008.

[25] W.-L. Huang, C.-W. Tung, H.-L. Huang and S.-Y. Ho, "Predicting protein subnuclear localization using GO-amino-acid composition," *BioSystems,* vol. 98, pp. 73-79, 2009.

[26] S.-C. Chen, T. Zhao, G. J. Gordon and R. F. Murphy, "Automated image analysis of protein localization in budding yeast," *Bioinformatics,* vol. 23, 2007.

[27] H.-B. Shen and K.-C. Chou, "Predicting protein subnuclear location with optimized evidence-theoretic K-nearest classifier and pseudo amino acid composition," *Biochemical and Biophysical Research Communications,* vol. 337, no. 3, pp. 752-756, 2005.

[28] Y.-Y. Xu, F. Yang, Y. Zhang and H.-B. Shen, "An image-based multi-label

human protein subcellular localization predictor (iLocator) reveals protein mislocalizations in cancer tissues," *Bioinformatics,* vol. 29, no. 16, pp. 2032-2040, 2013.

[29] P. Shah, "Insights into Machine Learning," 8 1 2018. [Online]. Available: https://opensourceforu.com/2018/01/insights-machine-learning/.

[30] P. Hassani, "An Insight into 26 Big Data Analytic Techniques: Part 2," 30 11 2016. [Online]. Available: https://blogs.systweak.com/2016/11/an-insight-into-26-big-data-analytic-techniques-part-2/.

[31] Priyadharshini, "Machine Learning: What it is and Why it Matters," 18 3 2018. [Online]. Available: https://www.simplilearn.com/what-is-machine-learning-and-why-it-matters-article.

[32] R. van Loon, "Machine Learning Explained: Understanding Supervised, Unsupervised, and Reinforcement Learning," 6 1 2018. [Online]. Available: https://www.datasciencecentral.com/profiles/blogs/machine-learning-explained-understanding-supervised-unsupervised.

[33] R. M. Leimgruber, "Extraction and Solubilization of Proteins for Proteomic Studies," in *The Proteomics Protocols Handbook*, Humana Press, 2005.

[34] "Note on Amino acids, Proteins Lipids or Fats and Steroids (Biomolecules)," [Online]. Available: https://www.kullabs.com/classes/subjects/units/lessons/notes/note-detail/2428.

[35] G. M. Cooper, "Protein Synthesis, Processing, and Regulation," in *The Cell: A Molecular Approach*.

[36] K. Guruprasad, B. Reddy and M. Pandit, "Correlation between stability of a protein and its dipeptide composition: a novel approach for predicting in vivo stability of a protein from its primary sequence," *Protein Eng.,* vol. 4, no. 2, pp. 155-161, 1990.

[37] R. Singh, P. R. Singh and D. Pal Kaur, "Improved Protein Function Classification Using Support Vector Machine," *International Journal of Computer Science and Information Technologies,* vol. 6, no. 2, pp. 964-968,

2015.

[38] S. Wan and M.-W. Mak, Machine Learning for Protein Subcellular Localization Prediction, De Gruyter, 2015.

[39] M.-C. Hung and W. Link, "Protein localization in disease and therapy," *Journal of Cell Science,* vol. 124, pp. 3381-3392, 2011.

[40] S. Bachle, "addgene," 22 6 2017. [Online]. Available: https://blog.addgene.org/plasmids-101-visualizing-subcellular-structures-organelles.

[41] C. M. Bishop, Pattern Recognition and Machine Learning, Springer.

[42] J. Gu, Z. Wang, J. Kuen, L. Ma, A. Shahroudy, B. Shuai, T. Liu, X. Wang, G. Wang, J. Cai and T. Chen, "Recent advances in convolutional neural networks," *Pattern Recognition,* vol. 77, pp. 354-377, 2018.

[43] L. He, X. Ren, Q. Gao, X. Zhao, B. Yao and Y. Chao, "The connected-component labeling problem: A review of state-of-the-art algorithms," *Pattern Recognition,* vol. 70, pp. 25-43, 2017.

[44] T. Hill and P. Lewicki, "Support Vector Machines," in *Electronic Statistics Textbook*, StatSoft Inc..

[45] "Support Vector Machines," in *Electronic Statistics Textbook*, 1995.

[46] E. Frank and I. H. Witten, "Generating Accurate Rule Sets Without Global Optimization," *ICML,* 1998.

[47] V. Losing, B. Hammer and H. Wersing, "KNN Classifier with Self Adjusting Memory for Heterogeneous Concept Drift," in *IEEE 16th International Conference on Data Mining (ICDM)*, 2016.

[48] Y. Bengio and Y. Grandvalet, "No Unbiased Estimator of the Variance of K-Fold Cross-Validation," *Journal of Machine Learning Research,* vol. 5, pp. 1089-1104, 2004.

[49] "ExPASy - ProtParam tool," [Online]. Available: http://web.expasy.org/protparam/.

[50] Y. Perez-Riverol, E. Audain, A. Millan, Y. Ramos, A. Sanchez, J. A.

Vizcaíno, R. Wang, M. Müller and Y. J. Machado, "Isoelectric point optimization using peptide descriptors and support vector machines," *Journal of Proteomics,* vol. 75, no. 7, pp. 2269-2274, 2012.

[51] L. P. Kozlowski, "IPC – Isoelectric Point Calculator," *Biology Direct,* vol. 11, no. 1, p. 55, 2016.

[52] C. Hoogland, K. Mostaguir, J. Sanchez, D. Hochstrasser and R. Appel, "SWISS-2DPAGE, ten years later," *Proteomics,* vol. 4, no. 8, pp. 2352-2356, 2004.

[53] E. Bunkute, C. Cummins, F. Crofts, G. Bunce, I. Nabney and D. Flower, "PIP-DB: the Protein Isoelectric Point database," *Bioinformatics2,* vol. 31, no. 2, pp. 295-296, 2015.

[54] L. P. Kozlowski, "Proteome-pI: proteome isoelectric point database," *Nucleic Acids Research,* 2017.

[55] A. Ikai, "Thermostability and aliphatic index of globular proteins," *J. Biochem.,* vol. 88, no. 6, pp. 1895-1898, 1980.

[56] P. B. Upama, J. T. Khan, Z. Yasmin, F. Zemim and N. Sakib, "A Noble Approach on Bioinformatics: Smart Sequence Alignment Algorithm applying DNA Replication (SSAADR)," *International Journal of Applied Information Systems (IJAIS),* vol. 8, no. 1, pp. 23-28, 2014.

[57] S. B. Needleman and C. D. Wunsch, "A general method applicable to the search for similarities in the amino acid sequence of two proteins," *Journal of Molecular Biology,* vol. 48, no. 3, pp. 443-453, 1970.

[58] T. F. Smith and M. S. Waterman, "Identification of common molecular subsequences.," *Journal of Molecular Biology,* vol. 147, pp. 195-197, 1981.

[59] S. A. Shehab, A. Keshk and H. Mahgoub, "Fast Dynamic Algorithm for Sequence Alignment based on Bioinformatics," *International Journal of Computer Applications,* vol. 37, no. 7, pp. 54-60, 2012.

[60] J. Kyte and R. F. Doolittle, "A simple method for displaying the hydropathic character of a protein," *Journal of Molecular Biology,* vol. 157, no. 1, pp. 105-132, 1982.

[61] J. Frost, "How to Correctly Interpret P Values," [Online]. Available: http://blog.minitab.com/blog/adventures-in-statistics-2/how-to-correctly-interpret-p-values.

[62] "Laerd Statistics," [Online]. Available: https://statistics.laerd.com/statistical-guides/one-way-anova-statistical-guide.php.

[63] S. Kaushik, "Introduction to Feature Selection methods with an example (or how to select the right variables?)," 1 12 2016. [Online]. Available: https://www.analyticsvidhya.com/blog/2016/12/introduction-to-feature-selection-methods-with-an-example-or-how-to-select-the-right-variables/.

[64] "ANOVA:ANalysis Of VAriance between groups," College of Saint Benedict & Saint John's University, [Online]. Available: http://www.physics.csbsju.edu/cgi-bin/stats/anova_pnp.

[65] K. Abdullah-Al-Mamun, "Pattern identification of movement related states in biosignals".

[66] B. Yekkehkhany, A. Safari, S. Homayouni and M. Hasanlou, "A Comparison Study of Different Kernel Functions for SVM-based Classification of," *The International Archives of the photogrammetry, Remote Sensing and Spatial Information Sciences,* 2014.

[67] Y. Liu and K. K. Parhi, "Computing RBF kernel for SVM classification using stochastic logic," *Proceedings - IEEE International Workshop on Signal Processing Systems, SiPS 2016,* pp. 327-332.

[68] M. Ring and B. M. Eskofier, "An approximation of the Gaussian RBF kernel for efficient classification with SVMs," *Pattern Recognition Letters,* vol. 84, no. C, pp. 107-113, 2016.

[69] B. Hssina, A. Merbouha, H. Ezzikouri and M. Erritali, "A comparative study of decision tree ID3 and C4.5," *International Journal of Advanced Computer Science and Applications, Special issue on Advances in Vehicular Ad Hoc Networking and Applicatios,* pp. 13-19.

[70] M. Britsch, N. Gagunashvili and M. Schmelling, "Application of the rule-growing algorithm," [Online]. Available:

https://indico.cern.ch/event/34666/contributions/813578/attachments/683858/939357/talkBritschAcat2008.pdf.pdf.

[71] A. Bronshtein, "A Quick Introduction to K-Nearest Neighbors Algorithm," 2017. [Online]. Available: https://medium.com/@adi.bronshtein/a-quick-introduction-to-k-nearest-neighbors-algorithm-62214cea29c7.

[72] E.-h. Zheng , C. Zou , J. Sun, L. Chen and P. Li, "SVM-Based Cost-sensitive Classification Algorithm with Error Cost and Class-dependent Reject Cost," in *Second International Conference on Machine Learning and Computing* , 2010.

[73] H. Cao, T. Naito and Y. Ninomiya, "Approximate RBF Kernel SVM and Its Applications in," *The 1st International Workshop on Machine Learning for Vision-based Motion Analysis - MLVMA'08,* 2008.

[74] R. Christensen, W. Johnson, A. Branscum and T. E. Hanson, Bayesian Ideas and Data Analysis: An Introduction fro Scientists and Statisticians, CRC Press, 2011.

[75] L. Han, M. J. Embrechts, B. Szymanski, K. Sternickel and A. Ross, "Sigma Tuning of Gaussian Kernels: Detection of Ischemia from Magnetocardiograms," *Computational Modeling and Simulation of Intellect: Current State and Future Perspectives,* pp. 206-223, 2011.