MULTI-VIEW SUBCELLULAR LOCALIZATION PREDICTION OF HUMAN PROTEINS

A THESIS SUBMITTED TO
THE GRADUATE SCHOOL OF NATURAL AND APPLIED SCIENCES
OF
MIDDLE EAST TECHNICAL UNIVERSITY

BY

GÖKHAN ÖZSARI

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR
THE DEGREE OF MASTER OF SCIENCE
IN
COMPUTER ENGINEERING

SEPTEMBER 2019

Approval of the thesis:

MULTI-VIEW SUBCELLULAR LOCALIZATION PREDICTION OF HUMAN PROTEINS

submitted by **GÖKHAN ÖZSARI** in partial fulfillment of the requirements for the degree of **Master of Science in Computer Engineering Department, Middle East Technical University** by,

Prof. Dr. Halil Kalıpçılar
Dean, Graduate School of **Natural and Applied Sciences** _____

Prof. Dr. Halit Oğuztüzün
Head of Department, **Computer Engineering** _____

Prof. Dr. M. Volkan Atalay
Supervisor, **Computer Engineering, METU** _____

Examining Committee Members:

Prof. Dr. Tolga Can
Computer Engineering, METU _____

Prof. Dr. M. Volkan Atalay
Computer Engineering, METU _____

Assoc. Prof. Dr. Tunca Doğan
Computer Engineering, Hacettepe University _____

Date:

I hereby declare that all information in this document has been obtained and presented in accordance with academic rules and ethical conduct. I also declare that, as required by these rules and conduct, I have fully cited and referenced all material and results that are not original to this work.

Name, Surname: Gökhan Özsarı

Signature :

ABSTRACT

MULTI-VIEW SUBCELLULAR LOCALIZATION PREDICTION OF HUMAN PROTEINS

Özsarı, Gökhan
M.S., Department of Computer Engineering
Supervisor: Prof. Dr. M. Volkan Atalay

September 2019, 92 pages

Determining the subcellular localization of proteins is crucial for understanding the functions of proteins, drug targeting, systems biology, and proteomics research. Experimental validation of subcellular localization is an expensive and challenging process. There exist several computational methods for automated prediction of protein subcellular localization; however, there is still room for better performance. Here, we propose a multi-view SVM-based approach that provides predictions for human proteins. We represent each protein sequence by multi-view features; i.e., physicochemical properties, amino acid compositions, and homology-based features. Our classification model contains seven classifiers for each localization, where each classifier provides a probabilistic result. To develop a multi-view voting classifier, we employ a weighted classifier combination method that assigns different weights to classifiers based on their discriminative strengths. We evaluated the described method on previously used datasets, as well as on our in-house dataset, called Trust dataset. Trust dataset is created by using a new subcellular localization hierarchy which merges UniProt Subcellular Location hierarchy and GO Cellular Component hierarchy by ap-

plying it on only manual experimental annotations in UniProtKB. We compared our results with five state-of-the-art methods, which are SubCons, LocTree2, CELLO2.5, MultiLoc2, and DeepLoc. Our approach outperformed the others with 59%, 61%, 68% overall Matthews correlation coefficient (MCC) scores on Trust, Golden (SubCons benchmark dataset), Golden-Trust (refined Golden dataset) datasets, respectively where SubCon's MCC scores were 43%, 53%, and 56%.

Keywords: subcellular localization, prediction, human proteins, svm, multi-view

ÖZ

İNSAN PROTEİNLERİNİN ÇOKLU GÖRÜNÜM YOLUYLA HÜCRE İÇİ YERLEŞİMLERİNİN TAHMİNİ

Özsarı, Gökhan
Yüksek Lisans, Bilgisayar Mühendisliği Bölümü
Tez Yöneticisi: Prof. Dr. M. Volkan Atalay

Eylül 2019 , 92 sayfa

Protinlerin hücre içi yerleşimlerinin belirlenmesi proteinlerin fonksiyonlarını anlamakta, ilaç belirleme çalışmalarında, sistem biyolojide ve protemik araştırmalarda büyük önem arz etmektedir. Proteinlerin hücre içi yerleşimlerini otomatik olarak tahmin eden bir çok hesaba dayalı metot vardır. Fakat hala daha iyi performans verebilecek çalışmalara ihtiyaç vardır. Biz çalışmamızda çoklu görünüme ve destek vektör makinalarına dayalı insan proteinlerinin hücre içi yerleşimlerini tahmin eden yeni bir yöntem ortaya koyuyoruz. Her bir proteini çoklu görünüm sağlayan birden fazla özellik ile ifade ediyoruz. Bu özellikler fiziko kimyasal özellikler, amino asit bileşimleri ve homoloji tabanlı özelliklerdir. Bizim sınıflandırma modelimiz her bir yerleşim için ihtimal belirten sonuç veren yedi sınıflandırma içerir. Çok görünümlü sistem geliştirme amacıyla sınıflandırıcıları ayırıcı gücüne dayalı olarak farklı ağırlıklar veren ağırlıklı sınıflandırma metodunu kullandık. Bu yöntemimizi data önce test amaçlı kullanılmış veriler üzerinde ve kendi geliştirdiğimiz yeni veri üzerinde değerlendirdik. Kendi oluşturduğumuz Trust (Güven) veri kümesini UniProtKB hücre içi hiyerarşisini ve GO (gen ontoloji) hücresel bileşenler hiyerarşisini birleştirerek oluş-

turduğumuz özgün hücre içi yerleşim hiyerarşisini otomatik olmayan şerhlere sahip proteinlere uygulayarak elde ettik. Sonuçlarımızı en gelişkin beş metot olan SubCons, LocTree2, CELLO2.5, MultiLoc2 ve DeepLoc yöntemleri ile karşılaştırdık. Bizim ortaya koyduğumuz yaklaşım test için kullandığımız üç veri kümesi olan Trust (biz oluşturduk), Golden(Subcons'un veri kümesi) ve Golden-Trust (iyileştirdiğimiz Golden veri kümesi) üzerinde sırasıyla 59%, 61%, 68% ortalama Matthews correlation coefficient (MCC) skorları elde ederken diğer beş metotta bize en yakın ortalama skorlara ulaşan Subcons 43%, 53%, and 56% MCC skorlarına ulaşmıştır.

Anahtar Kelimeler: hücre içi yerleşim, tahmin, insan proteinleri, destek vektör makinaları, çoklu görünüm

to the memory of my dad

ACKNOWLEDGMENTS

I would like to express my profound gratitude to my supervisor, Prof. Dr. Volkan Atalay, for his guidance, support, and patience during this thesis. I also thank Prof. Dr. Rengül Çetin-Atalay for her priceless advice and critique for the biological aspect of this work. I'm delighted working with them.

I am much obliged to Assoc. Prof. Dr. Tunca Doğan for his fruitful comments, advice, and productive criticism to enhance my study. I would also like to thank Ahmet Rifaioğlu for his valuable support to overcome the problems when I am puzzled.

I want to thank my friends, Ahmet Atakan, Anıl Çetinkaya, Alper Karamanlıoğlu, M.Çağrı Kaya, Alperen Dalkıran, Alperen Eroğlu, Murat Öztürk, and Tuğberk İşyapar. It has always been exceptional to spend time with you.

Finally, I would like to thank my family for their continuous support during my study. I'm especially grateful to my wife Tülay Özsarı and my children, Hüsnü Musab Özsarı, Hüma Elif Özsarı, and Hayme Erva Özsarı.

TABLE OF CONTENTS

LIST OF TABLES

TABLES

LIST OF FIGURES

FIGURES

LIST OF ABBREVIATIONS

SVM	Support Vector Machine
SL	Subcellular Location
UniProtKB	Universal Protein Knowledge Base
GO	Gene Ontology
CNN	Convolutional neural network
LSTM	Long-short-term memory
CC	Cellular component
NUC	Nucleus
CYT	Cytoplasm
MEM	Cell membrane
EXC	Secreted
MIT	Mitochondrion
ERE	Endoplasmic reticulum
GLG	Golgi apparatus
LYS	Lysosome
PEX	Peroxisome
MF	Molecular Function
BP	Biological Process
Mass-Spec	Mass spectrometry
C/T/D	Composition, Transition and Distribution
PSSM	Position Specific Scoring Matrix
MCC	Mathews Correlation Coefficient
TP	True positive
FN	False negative

FP	False positive
TN	True negative
PPI	Protein-protein interactions

CHAPTER 1

INTRODUCTION

Extensive genomic and proteomic studies have contributed a colossal amount of sequence data. The sequence of a protein is an essential factor in molecular and computational biology. In order to annotate protein functions, the potential roles of proteins in a cellular context, such as metabolic pathways and interaction networks, must be elucidated.

Cells can synthesize a different type of proteins that are generated to function in the target organelles or subcellular locations within cells. Therefore, the transportation of a protein to its final destination (target organelle) is required for performing its function. The failure to transport a protein has proven to be a vital issue for a variety of human diseases, such as cancer and Alzheimer's disease.

In silico (computational) methods, to predict subcellular localization of a protein, provide prior knowledge for *in vivo* and *in vitro* (experimental) studies. Therefore, various subcellular localization prediction tools have been developed in recent years.

1.1 Motivation and Problem Definition

Cells are the basic units of life consisting of organelles and having different tasks for the survival of living things. The biological functions in the cell are carried out by the proteins in the organelles. Organelles are also the location for proteins. Proteins are the result of amino acids coming together to form protein sequences. The sequence of amino acids and the structure of amino acid compounds in the sequence are the most critical factors that determine the function and localization of proteins.

Besides, if the function of a protein is unknown, the subcellular location of a protein is a very significant clue about its function. Therefore, to identify the function of a protein, it is crucial to know its subcellular localization. There are in vivo and in vitro methods to determine the subcellular localization of proteins. However, the experimental methods are expensive and time-consuming. Therefore, several computational methods are proposed in the last two decades to predict the subcellular localization of proteins; yet, there is still room for better performance. Here, we propose a multi-view Support Vector Machine (SVM)-based approach that provides predictions for the subcellular localization of human proteins.

1.2 Proposed Method

There are three major parts that we describe in this study. We first create a new subcellular location hierarchy that merges Universal Protein Knowledge Base (UniProtKB) Subcellular Location (SL) hierarchy [2] and Gene Ontology (GO) Cellular Component (CC) hierarchy [3]. We then form a dataset that is called Trust dataset since it contains only the proteins which have experimental evidence for the subcellular localization. We finally propose a multi-view classification approach that represents each protein by using multiple protein descriptors and employs weighted mean voting based on Support Vector machines (SVM). The proposed approach has four steps to predict the subcellular localization of human proteins over nine groups of subcellular locations. The proposed classification models work as follows :

Prediction Step-1. Feature extraction and selection: In this step, we represent the protein sequence by employing seven protein descriptors that form the best combination with their features. Therefore, we employ three feature extraction tools: iFeature tool [4], POSSUM tool [5], andSPMAP [6] from which forty protein descriptors are employed to select the best representative seven of them.

Prediction Step-2. Obtaining probabilistic scores: In this step, there are seven pre-trained SVMs by using the training dataset whose features are extracted by the

selected seven protein descriptors. Each SVM gives a probabilistic score that indicates the localization probability for a query protein sequence.

Prediction Step-3. Weighted-mean voting: We add-up the products of probabilistic scores and weights where each weight represents the discriminative power of the SVM based on the MCC score.

Prediction Step-4. Thresholding: The weighted-score above the pre-determined threshold indicates positive predictions, whereas the equal or below is predicted as negative.

The predictor that we propose includes nine independently constructed classification models where each model provides binary predictions for one of the nine subcellular localizations of human proteins. These locations are CYT(Cytoplasm), NUC(Nuclear), MEM(Membrane), MIT(Mitochondrion), ERE(Endoplasmic reticulum), EXC(Secreted), GLG(Golgi apparatus), LYS(Lysosome) and PEX(Peroxisome).

1.3 Improvements

In this study, we describe a multi-view subcellular localization prediction method where it outperforms existing methods with 82% overall precision and overall 63% MCC score out of three datasets Trust-Test, Golden and Golden-Trust. Our method is called as multi-view approach since we employ multiple protein descriptors to extract the features of protein sequences. The protein descriptors are selected out of 160 cases of the protein descriptors and normalization methods. Besides, due to the lack of sufficent information about the relations among subcellular location (SL) terms in UniProtKB SL hierarchy, to use UniProtKB annotations with a better SL hierarchy we employ Gene Ontology Cellular Component hierarchy. We describe a newly generated SL hierarchy which utilizes UniProtKB subcellular location terms and Gene Ontology cellular component hierarchy. Moreover, one of the most important contribution of our study is Trust dataset of human proteins. We present a carefully prepared dataset, Trust dataset, that can be used in both training and test for future

studies. Trust dataset includes protein sequences that cover almost all the subcellular locations in a human cell. The foremost difference of Trust dataset is the experimental evidence for the subcellular localization annotations of the proteins. Also, we refined the Golden dataset(benchmark dataset of Subcons) [7] by using the way we formed the Trust dataset and introduced a new dataset which is called as the Golden-Trust dataset.

CHAPTER 2

BACKGROUND INFORMATION AND RELATED WORK

This chapter presents an overview of the biological background and related work about the subcellular localization of proteins.

2.1 Biological background

In this section, we explain the biological background of our work. Cells and organelles are first defined. Proteins and types of proteins are then presented. Finally, the subcellular localization of proteins is introduced.

2.1.1 Cells and Organelles

Cells are the smallest living units. It is a closed system, can replicate themselves and are the building blocks of our body. To understand how these small organisms work, we examine the internal components of a cell. We will focus on eukaryotic cells, nucleus-containing cells, whereas prokaryotic cells are structured differently from nucleus-free cells. A cell consists of two main regions, the cytoplasm, and the nucleus. The nucleus is enclosed by a nuclear envelope and contains chromosomal DNA. The cytoplasm is a liquid matrix that surrounds the nucleus and is bound by the outer cell membrane. Organelles are the components of cells within the cytoplasm that perform functions necessary to sustain homeostasis in the cell. They are involved in various processes, such as energy production, protein, and secretory production, the destruction of toxins, and the response to external signals.

Organelles are examined either membranous or non-membranous. Membranous or-

ganelles have their plasma membranes to produce a lumen separated from the cytoplasm while non-membranous organelles are not surrounded by a plasma membrane. Most non-membranous organelles are part of the cell skeleton, the primary carrier structure of the cell, which are filaments, microtubes, and centrifuges.

The nucleus can be considered as the center of operations in the cell. There is usually one nucleus per cell, but this is not always the case, for example, skeletal muscle cells have two. The nucleus contains most of the DNA in the cell while a small amount is in the mitochondria. The nucleus sends messages to instruct, grow, share, or die. The nucleus is isolated from the rest of the cell by a membrane called the nuclear envelope. The nuclear pores in the membrane pass through small molecules and ions, while larger molecules require transport proteins to penetrate them.

The cytoplasm dwells in the cell that encloses the nucleus and consists of about 80 percent water. It includes organelles and a gelatinous fluid called cytosol. Many of the essential reactions within the cell happen in the cytoplasm.

The plasma membrane assures that each cell remains shielded from its neighbor. This membrane is mainly composed of phospholipids, which prevents water-based substances from entering the cell. The plasma membrane comprises a series of receptors that perform various tasks.

Mitochondrion, commonly known as the cell's power, help transform the energy of food into the energy (adenosine triphosphate) that can be used in the cell. However, it has also other roles such as including the storage of calcium and a role in cell death.

The endoplasmic reticulum (ER) is an extensive membrane network responsible for protein production, metabolism and lipid transport, as well as detoxification of poisons. Two types of endoplasmic reticulum are rough endoplasmic reticulum and smooth endoplasmic reticulum. The type of ER is determined according to the existence of ribosomes in the plasma membrane of the ER.

The Golgi apparatus is regarded as the post office of the cell in which the proteins are modified, classified, packaged, and labeled for secretion. Besides, it is involved in the transportations of lipids within the cell and the creation of lysosomes. The sacks or folds of the Golgi apparatus are called cisternae.

The lysosome is one of the organelles which play a significant role in waste disposal, whereas there are also other organelles to remove wastes. The lysosome contains digestive enzymes where excessive organelles, food particles, and entangled viruses or bacteria are digested.

Peroxisomes are single membrane-bound organelles that contain enzymes. Peroxisomes have two functions: they break down the fatty acids used to form membranes and used as a fuel for breathing and transfer hydrogen from the compounds to oxygen to form hydrogen peroxide and then convert the hydrogen peroxide into water.

Figure 2.1 illustrates a schematic representation of a human cell and its organelles.

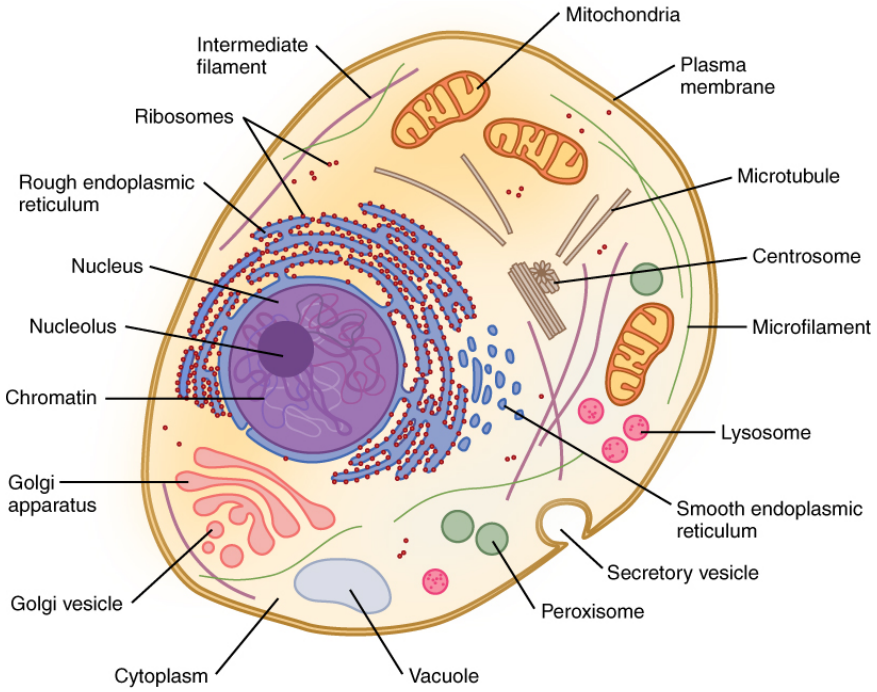

Figure 2.1: Human cell and its organelles [1]

2.1.2 Proteins

Proteins are large and complex molecules that play essential roles in the body. They perform most of the activities in the cells and are necessary for the structure, function, and regulation of body tissues and organs. Proteins comprise hundreds or thousands of small units called amino acids that link together in long chains. Amino acids are

the primary compounds of proteins that play an essential role in living organisms. Twenty different types of amino acids can be used to form proteins. Each amino acid is composed of an amino group and a carboxyl group bonded to a tetrahedral carbon, which is called alpha carbon. The amino acids vary in terms of their side chains, which are called R groups. The R group for each of the amino acids diversifies in structure, electrical charge, and polarity. The amino acid sequence determines the unique three-dimensional structure and specific function of each protein. Proteins can be listed according to their scope of functions as described below:

- Enzymes perform almost all chemical reactions within the cells. Besides, they read the genetic information stored in the DNA to generate new molecules.
- Structural component proteins provide structure and support for cells. Comprehensively they also enable the body to move.
- Antibodies bind to foreign particles such as viruses and bacteria to protect the body.
- Transport and storage proteins bind and transport small atoms and molecules in cells.
- Messenger proteins transmit signals to organize biological processes between different cells, tissues, and organs.

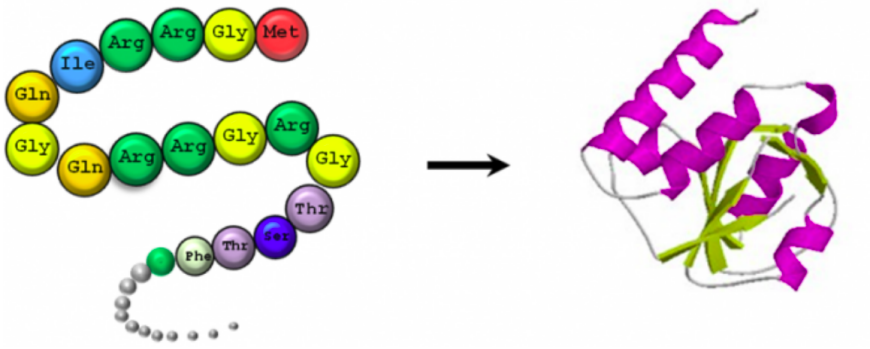

Figure 2.2: An illustration of a part of protein sequence and its three-dimensional view [8]

2.1.3 Subcellular localization of proteins

Subcellular localization of the protein is one of the essential aspects that must be characterized to comprehend cell biology. Identifying subcellular localization of a protein presents a beginning which gives lues about its functions and interaction partners for the proteins with limitations, or there is no information available. The definition of proteins of subcellular structures and organelles is a fundamental step towards the understanding of cell anatomy. Protein localization can be identified by using *in vivo* and *in vitro* methods such as mass spectrometry or immunofluorescent imaging and fluorescent protein labeling as well as by utilizing *in silico* methods (computational methods). The computational methods are described in the section of related work.

2.2 Related Work

Identifying the subcellular localization of proteins by *in vivo* and *in vitro* methods is an expensive and time-consuming process. Hence, various computational methods are proposed to predict the subcellular localization of proteins in the last two decades. These methods can be categorized as sequence-based, annotation-based, and hybrid methods. Sequence-based methods includes the following categories:

- sorting-signals based methods: PSORT [9], WoLF PSORT [10], TargetP [11] and SignalP [12].
- composition-based methods, such as amino-acid compositions [13], amino-acid pair compositions, gapped amino-acid pair compositions (GapAA) [14], and pseudo-amino-acid composition [15].
- homology-based methods: Proteome Analyst [16], PairProSVM [17], and other predictors [18].

Annotation-based methods employ the correlation between annotations of a protein and its subcellular location. The most used Gene Ontology (GO) information. GO-based predictors can be categorized into three categories:

- InterProScan [19] is a database of protein signatures [20],

- employ protein access numbers to seek the GO annotation database such as Hum-PLoc [21], Euk-mPLoc [22], and Greg-PLoc [23],
- utilizing homologous protein access numbers from BLAST [24] to seek for the GO annotation database such as ProLoc-GO [25], Cell-PLoc 2.0 [35], and mGOASVM [26].

There are also hybrid approaches that unite the characteristics of sequence-based mathods and annotation-based methods such as Cello2GO [27], SherLoc2 [28], and MultiLoc2 [29].

Table 2.1: The predictors that we employed to compare the proposed predictor.

Predictors	Year	Category	ML method	Organism	Number of SLs
CELLO2.5	2006	Sequence-based	SVM	Eukaryotes, bacteria	12
MultiLoc2	2009	Hybrid	SVM	Eukaryotes	11
LocTree2	2012	Sequence-based	SVM	Eukaryotes	18
SubCons	2017	Hybrid	Combines four predictors	Human	9
DeepLoc	2017	Sequence-based	RNN	Eukaryotes	9

We give an overview of the predictors that we employed to compare the proposed predictor, CanSLPred, as described below.

2.2.1 CELLO

CELLO [30] is a subcellular localization predictor that provides multi-class predictions for the proteins of three organisms: Gram-negative, Gram-positive, and Eukaryotes. It consists of two-layers of Support Vector Machine (SVM) that present probabilistic predictions for 12 subcellular localizations in Eukaryotes, 5 in Gram-positive and Gram-negative organisms. In the first layer of the predictor, the protein features are extracted by using four protein descriptors: amino acid composition, dipeptide composition, partitioned amino acid composition, and sequence composition based on the physicochemical properties of amino acids. Four SVMs are trained by using the features of the corresponding protein descriptors to give probabilistic predictions for all localizations where each trained SVM provides probabilistic scores for all sub-

cellular localizations. In the second layer, an SVM is employed for jury voting that combines all probabilistic predictions from four SVMs and produces a final probabilistic distribution on all localizations.

CELLO employs two datasets in training: a dataset for Gram-negative bacteria covers five subcellular locations that are extracellular, cytoplasmic, cytoplasmic membrane, periplasmic, outer membrane and a dataset for Eukaryotes includes the proteins from 12 subcellular locations: chloroplast, cytoplasmic, cytoskeleton, ER, extracellular, Golgi apparatus, lysosomal, mitochondrial, nuclear, peroxisomal, plasma membrane, vacuolar.

2.2.2 MultiLoc2

MultiLoc2 [29] is a subcellular localization predictor (the developed version of MultiLoc), which has two versions: one is LowRes (low resolution) version, the other one is HighRes (high resolution) version. LowRes (low resolution) version is for global proteins that provide predictions for five localizations whereas HighRes version predicts up to 11 subcellular localizations for eukaryotic proteins: nuclear, cytoplasmic, mitochondrial, chloroplast, extracellular, plasma membrane, peroxisomal, endoplasmic reticulum, Golgi apparatus, lysosomal and vacuolar proteins. MultiLoc2 has two layers: the first layer consists of six subpredictors: SVMTarget, SVMSA, SVMAaac, MotifSearch, PhyloLoc, and GOLoc. SVMTarget employs N-terminal targeting to predict the categories of a protein (mitochondrial, secretory pathway, or other). SVMSA is to detect the existence of a signal anchor. SVMaac consists of a set of SVMs that uses amino acid compositions of the proteins to give binary predictions for each localization. MotifSearch is to identify sequence motifs and structural domains that provide essential information for proteins. PhyLoc relies on phylogenetic profiles of the proteins in 78 fully sequenced genomes, and it employs SVMs as classifiers to predict all of the subcellular localization. GOLoc calculates Gene Ontology (GO) terms of proteins by using InterProScan [19] and forms a binary-coded vector for each protein that is generated by considering the presence of GO terms. The output of six subpredictors creates a protein profile vector which is used as input to the second-layer SVMs (one vs. one). SVMs in the second-layer provide the

probabilistic predictions for the localizations.

2.2.3 LocTree2

LocTree2 [31] is a subcellular localization predictor that provides predictions for archeal, bacterial and eukaryotic proteins in a hierarchical manner. It is a tree of 16 binary SVM classifiers which classify the eukaryotic proteins to 18 subcellular locations: chloroplast, chloroplast membrane, cytosol, endoplasmic reticulum, endoplasmic reticulum membrane, extracellular, fimbrium, Golgi apparatus, Golgi apparatus membrane, mitochondria, mitochondria membrane, nucleus, nucleus membrane, outer membrane, periplasmic space, peroxisome, peroxisome membrane, plasma membrane, plastid, vacuole, vacuole membrane.

Protein sequence profiles are created by using BLAST [24] and are employed to calculate the kernel matrices of the protein sequences. The calculated kernel matrices are used as an input to SVMs.

2.2.4 SubCons

Subcon [7] is an ensemble predictor that combines the previously proposed four predictors: CELLO2.5 [30], LocTree2 [31], MultiLoc2 [29], and SherLoc2 [28]. It has two levels in the classification of human proteins into nine subcellular locations that are nucleus, cytoplasm/cytoskeleton, mitochondria, peroxisome, ER, Golgi apparatus, lysosome, plasma membrane, secreted. In the first level, the predictions are collected from four predictors where CELLO2.5, MultiLoc2, and SherLoc2 give probabilistic predictions, and LocTree gives a binary prediction. In the second level, the predictions from four predictors are combined and used as input to the Random Forest classifier. Finally, subcellular localization prediction is provided with a probabilistic distribution on nine cellular compartments.

Moreover, a benchmark dataset of human proteins is generated by SubCons developers, which is called Golden dataset. This dataset consists of the proteins which are experimentally annotated in at least two data resources out of three: Mass Spectrom-

etry, SLHPA, and UniProt.

2.2.5 DeepLoc

DeepLoc [32] is a subcellular localization predictor that employs deep learning methods (recurrent neural networks with long-short-term memory cells, attention models, convolutional neural networks). It provides predictions for eukaryotic proteins on ten subcellular locations: nucleus, cytoplasm, extracellular, mitochondrion, cell membrane, ER, plastid, Golgi apparatus, lysosome/vacuole, peroxisome. Convolutional neural networks (CNNs) utilizes 120 filters to extract short motifs from protein sequences. Recurrent neural networks (RNNs) employs 256 long-short-term memory (LSTM) units to scan the protein sequence in both directions and outputs in 512,000 dimensions. The attention decoding layer utilizes an LSTM with 512 units through 10 decoding steps, and the attention mechanism feedforward neural network holds 256 units. The final fully connected dense layer is constituted by 512, and the two output layers have one unit for membrane-bound and ten units for the subcellular locations.

The above mentioned five predictors provide subcellular localization predictions for eukaryotic and human proteins. DeepLoc is one of the state-of-the-art methods which employs a deep-learning approach and presents other perspectives of protein sequences provided by its online version. On the other hand, SubCons is a hybrid method which combines four powerful prediction tools: LocTree2, MultiLoc2, Sherloc2, and CELLO2.5 that unite different aspects of proteins based on the sequences and GO annotations.

CHAPTER 3

DATASETS AND FEATURE EXTRACTION METHODS

In this chapter, we first explain how we create the new subcellular location hierarchy. The specifications and the steps to generate the datasets are then introduced. Further, the feature extraction tools and the information about the protein descriptors to extract the features are presented. Feature normalization methods are described at the end.

3.1 Subcellular Location Hierarchy

In our study we want to use UniProtKB annotations due to its realibility. However, UniProtKB Subcellular Location (SL) hierarchy is not sufficient to cover all relations among SL terms. Therefore, we come with an idea of integrating UniProtKB SL terms to Gene Ontology Cellular Component hierarchy and introduce a new subcellular location hierarchy that combines Universal Protein Resource Knowledge Base(UniProtKB) Subcellular Location (SL) hierarchy and Gene Ontology (GO) Cellular Component (CC) hierarchy. Our goal is to have a SL hierarchy which covers all relations (*is_a* relations) since it is vital to have a realiable hierarchy in the generation of the protein datasets. Some of the studies make use of UniProtKB Subcellular Location Hierarchy, whereas the others employ GO CC Hierarchy. To the best of our knowledge, it is the first time that these two hierarchies are combined. Two steps to create the proposed subcellular location hierarchy are described below:

Hier-1: Mapping UniProtKB SL identifiers to Gene Ontology(GO) CC terms: This step is to map UniProtKB subcellular location identifiers to G0 terms. We use a mapping of UniProtKB SL identifiers to GO CC terms which are defined in

the 'uniprotkb_sl2go' file. For instance, SL-091 the subcellular location identifier for nucleus in UniProtKB and GO:0005634 term for nucleus in GO were decided to be equivalent by the curator after comparing the definitions of both. Therefore, G0:0005634 is manually mapped to SL-091.

Hier-2: Forming the subcellular location hierarchy: To form a new subcellular location hierarchy, we consider the subcellular location terms from UniProtKB and 'is_a' relations of cellular components in GO hierarchy. For example, GO:0005634 'is_a' GO:0043231 means that GO:0043231 is a parent of GO:0005634 in GO hierarchy. We first extract GO CC hierarchy by using the document 'go-basic.obo' [33] where GO terms definitions and the hierarchical relationships with other GO terms are defined. We then replace all GO CC terms with the subcellular locations using the mapping in the document (uniprotkb_sl2go). There are 517 subcellular location identifiers in UniProtKB, but only 437 subcellular locations are mapped in 'uniprotkb_sl2go' document. Therefore, the missing subcellular locations in the mapping document are inserted into the hierarchy by using the subcellular localization hierarchy in the document 'subcell.txt' [2] which contains subcellular location terms, their definitions, and their hierarchical relationships with other subcellular locations, which are described in UniProtKB. Finally, we form the proposed subcellular location hierarchy, which Figure 3.1 depicts a part of it.

3.2 Mapping of subcellular locations

We consider the nine organelles of a human cell in our study where proteins perform their functions mainly within these organelles to sustain the life of a human cell. An organelle is a subcellular location in terms of the localization of proteins. Nine organelles that UniProtKB/SwissProt provides an adequate number of proteins for both training and test of machine learning methods are indicated as follows:

- Nucleus (NUC)
- Cytoplasm (CYT)
- Cell membrane (MEM)

- Secreted (EXC)
- Mitochondrion (MIT)
- Endoplasmic reticulum (ERE)
- Golgi apparatus (GLG)
- Lysosome (LYS)
- Peroxisome (PEX)

Moreover, the organelles consist of parts which are defined as a 'part_of' relation in both subcellular location hierarchies. For example, the mitochondrion envelope is a part of the mitochondrion. Therefore we include the mitochondrion envelope to MIT subcellular location group which can be defined as group of main organelle and its parts.

Consequently, it is inevitable to consider the parts of an organelle which form the complete structure of the organelle. Therefore, we define a mapping by considering 'part_of' relations in the two subcellular location hierarchies (UniProtKB SL hierarchy and GO CC hierarchy) and by including 'is_a' relations in the proposed subcellular location hierarchy. A schematic representation of the mapping is illustrated in Figure 3.2.

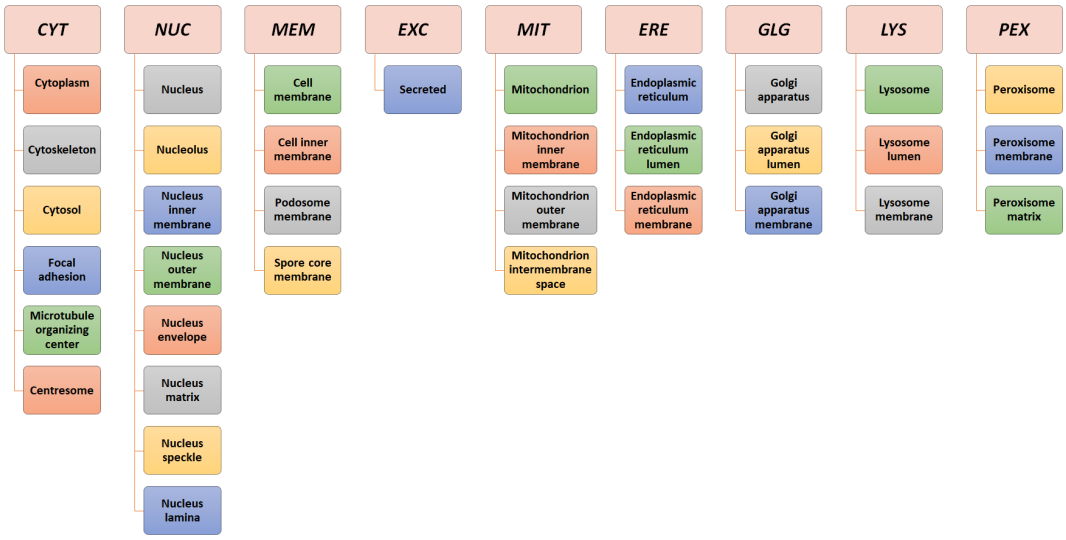

Figure 3.2: Mapping of subcellular locations formed by using 'is_a' and 'part_of' relations.

3.3 Universal Protein Resource Knowledge Base (UniProtKB)

Universal Protein Resource Knowledge Base (UniProtKB) [2] is a protein resource that provides a well-built, complete, freely available data on protein sequences and their functional annotations. UniProtKB contains the following information regarding a protein: "function(s), enzyme-specific information, biologically relevant domains, and sites, post-translational modifications, subcellular location(s), tissue specificity, developmentally specific expression, structure, interactions, splice isoform(s), diseases associated with deficiencies or abnormalities" [2]. UniProtKB has two databases, which are UniProtKB/Swiss-Prot and UniProtKB/TrEMBL. UniProtKB/Swiss-Prot includes manually annotated entries by expert curators, whereas UniProtKB/TrEMBL stores automatically annotated and unreviewed entries.

3.4 Gene Ontology (GO) database and informatics resource

The Gene Ontology (GO) [33] is a bioinformatics database that presents an ontology for the attributes of genes across all species. This ontology includes three domains that are Molecular Function (MF), Biological Process (BP), and Cellular Components

(CC). MF describes properties for the activities of genes at the molecular level. BP describes molecular events that process at living units (cells, tissues, organs). CC describes cellular components and their relationships by providing a unified representation for biological or molecular studies.

3.5 Datasets

We use three datasets for training and evaluation of our system. These datasets are Trust dataset, Golden dataset [7], Golden-Trust dataset. Trust dataset contains only manually annotated proteins from UniProtKB/Swiss-Prot [2]. Golden dataset is a benchmark dataset that is generated by SubCons developers, and Golden-Trust dataset is a refined version of Golden dataset by us. These datasets and their construction processes are explained in the following sections, and their usages are described in Chapter 4.

3.5.1 Trust dataset

Trust dataset is our in-house dataset, which is generated using manually annotated subcellular localizations in UniProtKB/Swiss-Prot and comprises nine parts that indicate independently generated datasets for nine subcellular location groups (NUC, CYT, MEM, EXC, MIT, ERE, GLG, LYS, PEX) displayed in Figure 3.2

Since we construct our method as a binary predictor for the subcellular localization prediction of proteins, Trust dataset is generated independently for all subcellular location groups in a binary form of one-SL vs. all other SL, where SL indicates the subcellular location groups. Trust dataset for each subcellular location group includes a dataset of positive class and a dataset of negative class where the dataset of positive class contains protein sequences that localize the subcellular location of positive class, and the dataset of negative class contains protein sequences for the other eight subcellular location groups. For instance, Trust dataset for subcellular locations in the group of CYT consists of two datasets which are the dataset of the positive class and the dataset of the negative class. These datasets can be defined as follows:

1. The dataset of the positive class includes protein sequences that localize at least one of the subcellular locations in CYT (positive class).

2. The dataset of the negative class comprises protein sequences that have subcellular localization annotations including at least one of the subcellular locations in the remaining eight subcellular location groups: NUC, MEM, EXC, MIT, ERE, GLG, LYS, and PEX (negative class) but not including any of the subcellular locations in CYT (positive class).

The construction steps of Trust dataset for each subcellular location group are described below:

Trust-1: Generating the subcellular location hierarchy: We create a new subcellular location hierarchy which merges UniProtKB subcellular location and Gene Ontology cellular component hierarchy. For more details, please refer to Section 3.1.

Trust-2: Obtaining protein sequences: We download all human protein sequences from UniProtKB/SwissProt that provides a well-built, comprehensive, freely available data on protein sequences and their functional annotations.

Trust-3: Filtering with UniRef50: The UniRef databases cluster the protein sequences in UniProtKB/SwissProt based on sequence similarities. There are three levels of clustering in UniRef [34] where UniRef50 is the one that contains the least similar protein sequences in comparison with UniRef90 and UniRef100. Therefore we employ the representative protein sequences from UniRef50 to reduce the redundancy which causes bias in the training process of machine learning methods.

Trust-4: Experimental evidence filter: The protein sequences that have experimentally annotated subcellular localization information are preferably taken UniProtKB/SwissProt. If the number of protein sequences is not sufficient (number of protein sequences is less than 500), we also include the protein sequences from UniProtKB/SwissProt, which have manually curated information concerning the subcellular localization.

Trust-5: Cleaning protein sequences in negative dataset: We eliminate the protein sequences from the dataset of negative class if a protein has any subcellular localization annotation with the subcellular location of positive class or its offsprings regarding GO hierarchy, UniProtKB subcellular location hierarchy, and the proposed subcellular location hierarchy.

Trust-6: Balancing the number of protein sequences in the dataset of positive class and the dataset of negative class: After completing the previous steps, we may observe that the dataset of positive class and the dataset of negative class are imbalanced. Most of the machine learning algorithms are sensitive to the imbalanced datasets, which cause bias favoring the majority class. Therefore, the dataset of positive class and the dataset of negative class are balanced by considering the cases explained below:

The first case is that the number of protein sequences in the dataset of positive class is higher than the number of protein sequences in the dataset of negative class. In this case, we randomly eliminate some of the protein sequences from the dataset of positive class.

The second case is that the number of protein sequences in the negative dataset is higher than the number of protein sequences in the positive dataset. We determine the number of protein sequences independently from each of the eight subcellular location groups in negative class that add up to the number of protein sequences in the dataset of positive class.

3.5.1.1 Multiple Localized Proteins

Trust dataset contains multiple localized proteins according to UniProtKB annotations. These protein sequences are used in both training and test process. It is also a significant difference of Trust dataset from other datasets. Since multiple SL prediction is also needed in biological studies, Trust dataset presents a benchmark dataset for testing multiple localization prediction of the proteins. The following table depicts the number of multiple localized proteins with respect to the number of locations.

Table 3.1: Number of multiple localized protein sequences in Trust dataset with respect to the number of SLs.

Number of SLs	Dataset	Number of protein sequences
at least six	Training	9
at least five	Training	81
at least four	Training	382
at least four	Test	3
at least three	Training	255
at least three	Test	39
at least two	Training	3277
at least two	Test	318

3.5.2 Golden Dataset

Golden dataset is the benchmark dataset of SubCons [7]. It comprises the protein sequences that localize the eight groups of subcellular locations: NUC, CYT, MEM, MIT, ERE, GLG, LYS, and PEX. To form Golden dataset, three data resources are used: mass spectrometry(Mass-Spec) [35], SLHPA [36] [37], and UniProtKB [2]. The proteins that have an experimental annotation of subcellular localization are retrieved from three data resources. The protein sequences are eliminated according to their homology by using BLASTClust [15]. The protein sequences that have a common subcellular localization annotation in at least two out of the three resources are selected. Eventually, Golden dataset contains a total of 1226 protein sequences that cover the eight subcellular location groups. The number of protein sequences for each subcellular location is given in Table 3.2

3.5.3 Golden-Trust Dataset

Golden-Trust dataset is a refined version of Golden dataset where the steps we follow to create Trust dataset are applied for the protein sequences in Golden dataset. Our goal in refining Golden dataset is to generate a dataset that is up-to-date con-

cerning the developments about the subcellular localization annotations in the protein sequence databases.

Table 3.2: Number of protein sequences in the datasets with respect to their subcellular location groups.

Groups of SLs	Class	Trust-All	Trust-Train	Trust-Test	Golden	Golden-Trust
CYT	positive	738	605	133	159	95
	negative	738	605	133	1067	95
NUC	positive	1599	1299	300	733	198
	negative	1599	1299	300	493	198
MEM	positive	858	685	144	47	22
	negative	858	685	133	1179	22
EXC	positive	385	311	63	NA	NA
	negative	387	312	67	NA	NA
MIT	positive	399	319	80	202	126
	negative	399	319	80	1024	126
ERE	positive	456	366	55	46	25
	negative	456	366	87	1180	25
GLG	positive	389	314	75	21	13
	negative	388	314	74	1205	13
LYS	positive	263	210	53	11	10
	negative	263	210	53	1215	10
PEX	positive	80	64	16	7	7
	negative	80	64	16	1216	7

3.6 Feature Extraction

Feature extraction is a fundamental step to develop a successful machine learning-based model. In the last two decades, various feature extraction methods and tools are developed which exploit patterns from the arrangement of amino acids in protein sequences [4] [5] [6] [38] [39] [40]. Feature extraction methods are also called as protein descriptors that are to represent protein sequences with numerical features. In our study, we employ three tools, which are iFeature [4], POSSUM [5], and SPMAP [6]. The protein descriptors in these tools provide features in different aspects that are

amino-acid composition-based, homology-based, mixture of these two, and subsequence based. We utilize 40 protein descriptors to extract protein features which 18 of them from iFeature, 21 from POSSUM and 1 from SPMAP.

3.6.1 iFeature Tool

iFeature is a Python-based toolkit that offers 48 protein descriptors to extract different numerical representations from protein and peptide sequences. iFeature also provides more functionalities that are five feature clustering algorithms, four feature selection algorithms, and three dimensionality reduction algorithms.

We employ 18 protein descriptors offered by iFeature to extract protein features that provide a numerical representation of protein sequences at different lengths. The rest of the protein descriptors in iFeature expects to have the same length of protein sequences. These 18 protein descriptors are categorized into seven groups. The first category, Amino Acid Composition, contains the descriptors which are created by producing the counts of amino acids in different ways. The descriptors in the second category, Grouped Amino Acid Composition, are generated by grouping amino acid types regarding different properties of amino acid types. The third category is Autocorrelation whose descriptors define the distribution of amino acid properties. The descriptors in fourth category C/T/D are to represent the composition, transition and distribution of amino acid patterns according to structural and physicochemical properties. The fifth category (Conjoint triad) includes the descriptors whose features are generated by considering three neighbor amino acids as a single unit. The descriptors in Quasi-sequence-order category examine the distance between the amino acid pairs in different ways. The last category Pseudo-amino acid composition contains descriptors whose features are obtained regarding "hydrophobicity values, the original hydrophilicity values and the original side chain masses of the 20 natural amino acids". The categories, the descriptors, and their properties are displayed in Table 3.3.

Table 3.3: Protein descriptors that we used from iFeature tool.

Descriptor group	Descriptor	Number of features
Amino acid composition	Amino acid composition (AAC)	20
	Composition of k-spaced amino acid pairs (CKSAAP)	2400
	Dipeptide composition (DPC)	400
Grouped amino acid composition	Grouped amino acid composition (GAAC)	5
	Composition of k-spaced amino acid group pairs (CKSAAGP)	150
	Grouped dipeptide composition (GDPC)	25
Autocorrelation	Moran (Moran)	240
	Geary (Geary)	240
	Normalized Moreau-Broto (NMBroto)	240
C/T/D	Composition (CTDC)	39
	Transition (CTDT)	39
	Distribution (CTDD)	195
Conjoint triad	Conjoint triad (CTriad)	343
	Conjoint k-spaced triad (KSCTriad)	343x(k+1)
Quasi-sequence-order	Sequence-order-coupling number (SOCNumber)	60
	Quasi-sequence-order descriptors (QSOrder)	100
Pseudo-amino acid composition	Pseudo-amino acid composition (PAAC)	50
	Amphiphilic PAAC (APAAC)	80

3.6.2 POSSUM Tool

POSSUM [5] is another Python-based toolkit that allows users to use 21 Position Specific Scoring Matrix (PSSM)-based protein descriptors to extract protein features. PSSM can be defined as a profile that is extracted from aligned protein sequences. The protein sequences are aligned by using either BLAST or PSI-BLAST. The profile indicates the number of occurrences of amino acids for each position in the alignment of protein sequences. PSSM-based feature extraction descriptors can be grouped into three categories: row transformations, column transformations, or a mixture of row and column transformations. The descriptors generated by POSSUM are categorized PSSM-based features into four groups with the other one protein descriptor that is a combination of the other descriptors in POSSUM. In the feature extraction level of our study, we use all descriptors offered by POSSUM. Table 3.4 depicts the information about the descriptors.

POSSUM [5] is another Python-based toolkit that allows users to use 21 Position Specific Scoring Matrix(PSSM)-based protein descriptors to extract protein features.

PSSM [41] can be defined as a profile that is extracted from aligned protein sequences. The protein sequences are aligned by using either BLAST or PSI-BLAST. The PSSM profile of the aligned sequences indicates the number of occurrences of amino acids for each position in the alignment of protein sequences. PSSM-based feature extraction descriptors can be grouped into three categories: row transformations, column transformations, or a mixture of row and column transformations. The descriptors generated by POSSUM are categorized PSSM-based features into four groups with one extra protein descriptor that is a combination of the other descriptors in POSSUM. In the feature extraction level of our study, we use all descriptors offered by POSSUM. Table-3.3 depicts the information about the descriptors.

Table 3.4: Protein descriptors that we use from POSSUM.

Descriptor group	**Descriptor**	**Number of features**
Row transformations	AAC-PSSM	20
	D-FPSSM	20
	smoothed-PSSM	1000
	AB-PSSM	400
	PSSM-composition	400
	RPM-PSSM	400
	S-FPSSM	400
Column transformations	DPC-PSSM	400
	k-separated-bigrams-PSSM	400
	tri-gram-PSSM	8000
	EEDP	400
	TPC	400
Mixture of row and column transformations	EDP	20
	RPSSM	110
	Pse-PSSM	40
	DP-PSSM	240
	PSSM-AC	200
	PSSM-CC	3800
Combination of above descriptors	AADP-PSSM	420
	AATP	420
	MEDP	420

3.6.3 SPMAP

SPMap [6] is a subsequence-based protein descriptor that is composed of two parts. The first part is subsequence profile map construction, which has three stages: extracting the subsequences, clustering the subsequences based on their pairwise similarities, and generating probabilistic profiles of the clusters. The steps that are followed in SPMap are as follows:

- Overlapping fixed-length subsequences are first extracted from protein sequences in the training dataset. In this study, we used 5 as a subsequence length.
- The clusters of subsequences are then created: The first subsequence in the protein sequence constitutes the representative subsequence of the first cluster. For each next subsequence, the similarity scores between the representative subsequences and the new subsequence are calculated using BLOSUM62 [42] matrix. If the similarity score of the new subsequence is higher than the predefined threshold, it is included in the most similar cluster. Otherwise, the new subsequence creates its own cluster.
- Next, position-specific scoring matrices(PSSM) are created based on the generated clusters. These PSSMs are used as a profile at the feature generation part.
- Finally, SPMap features are generated from protein sequences in two steps: Subsequences of query sequences are extracted, and the feature vectors are created based on the previously created PSSMs.

3.7 Normalization methods

Feature normalization is a process that scales features of data within a particular range. Normalization may have a significant impact on the performance of a machine learning application since most of the machine learning algorithms are quite sensitive to the distribution of features. For instance, the normalization of feature vectors before feeding to the SVM is critical since SVM assumes that the features are within a standard range.

In this section, we introduce four normalization methods that we employ in the construction of our subcellular localization prediction model.

3.7.1 Standardization (Z-score normalization)

Standardization [43] is a way of rescaling data so that they form a Gaussian-like distribution. Z-score for each sample is calculated using Equation 31.

$$\mathbf{y}'_i = \frac{\mathbf{y}_i - \mu(\mathbf{y})}{\sigma(\mathbf{y})}, \tag{31}$$

where $i = 1, 2, ..., n$, n is number of data points, $\mathbf{y}'_i$ represents normalized value, $\mathbf{y}_i$ indicates the values of protein features, $\mu(\mathbf{y})$ is the mean of $\mathbf{y}$, and $\sigma(\mathbf{y})$ is the standard deviation of $\mathbf{y}$ from the mean.

3.7.2 MinMax normalization

Min-Max normalization rescales data in the range of $[0, 1]$ where the minimum value is normalized to 0, and the maximum value is normalized to 1. Equation 32 to calculate each normalized value of a feature is

$$\mathbf{y}'_i = \frac{\mathbf{y}_i - min(\mathbf{y})}{max(\mathbf{y}) - min(\mathbf{y})}, \tag{32}$$

where $i = 1, 2, ..., n$, n is number of data points, $\mathbf{y}'_i$ represents normalized value, $\mathbf{y}_i$ indicates the values of protein features, $min(\mathbf{y})$ is minimum value of $\mathbf{y}$, and $max(\mathbf{y})$ is maximum value of the feature values ($\mathbf{y}$).

3.7.3 Power Transformation

Power transformations are useful to transform data distribution into a Gaussian-like form and to stabilize the variance of data. There are two types of this transformation which are Box-Cox transformation [44] and Yeo-Hohson transform [45]. We employed Yeo-Hohson transformation, which can be defined as follows:

$$\mathbf{y}'_i = \begin{cases} \frac{(\mathbf{y}_i+1)^\lambda}{\lambda}, & \text{if } \lambda \neq 0, \mathbf{y}_i \geq 0 \\ \log(\mathbf{y}_i + 1), & \text{if } \lambda = 0, \mathbf{y}_i \geq 0 \\ \frac{-[(-\mathbf{y}_i+1)^{(2-\lambda)}-1]}{2-\lambda}, & \text{if } \lambda \neq 2, \mathbf{y}_i < 0 \\ log(-\mathbf{y}_i + 1), & \text{if } \lambda = 2, \mathbf{y}_i < 0 \end{cases}$$

where $i = 1, 2, ..., n$, n is number of data points, $\mathbf{y}'_i$ represents normalized value, $\mathbf{y}_i$ indicates the values of protein features, and λ represents the power to which data should be raised in the range of $[0, 2]$.

3.7.4 Robust scaler normalization

Robust scalers are designed to produce statistical methods that are not affected by outliers. Commonly used robust scalers are the interquartile range(IQR) [46] and the median absolute deviation (MAD) [47]. In our study, we applied the interquartile range(IQR) that is also called as midspread, middle 50% or H-spread. IQR is a measure of statistical dispersal, and it defines quartiles to separate the dataset into four equal parts that are 25th, 50th, 75th quartiles denoted by Q1, Q2, and Q3 respectively. IQR can be defined as the subtraction of Q1 from Q3. Equation 33 to scale the data as follows:

$$\mathbf{y}'_i = \frac{\mathbf{y}_i - Q1(\mathbf{y})}{Q3(\mathbf{y}) - Q1(\mathbf{y})}, \tag{33}$$

where $i = 1, 2, ..., n$, n is number of data points, $\mathbf{y}'_i$ represents normalized value, $\mathbf{y}_i$ indicates the values of protein features, $Q1(\mathbf{y})$ is first-quartile and $Q3(\mathbf{y})$ is third-quartile in the distribution of feature values ($\mathbf{y}$).

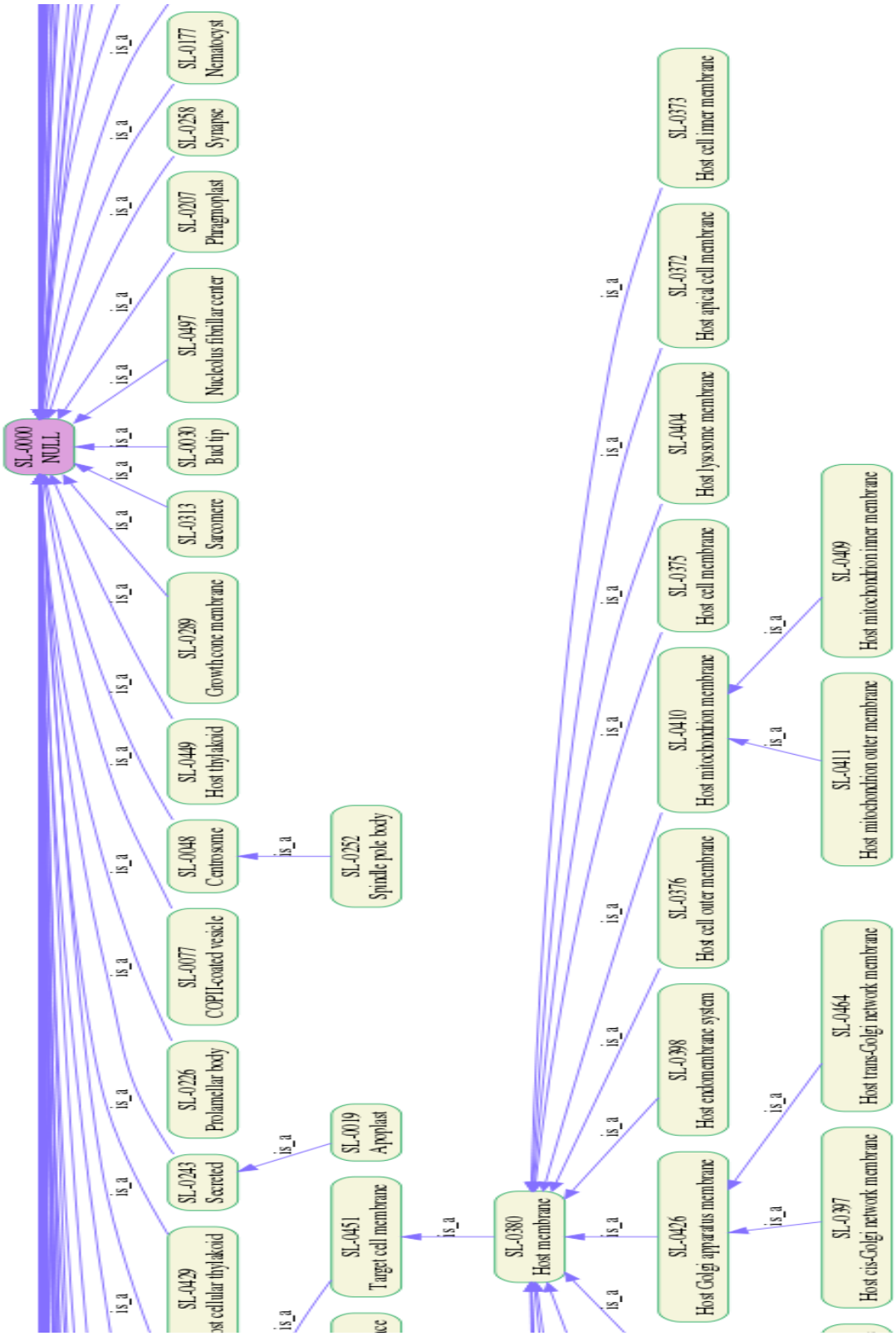

Figure 3.1: A part of the proposed subcellular location hierarchy

CHAPTER 4

PROPOSED METHOD

We propose a method to predict the subcellular localization of human proteins. The proposed method consists of nine independently constructed classification models where each model gives binary predictions for the protein sequences that localize to one of the nine subcellular location groups: CYT, NUC, MEM, MIT, ERE, EXC, GLG, LYS, and PEX. The classification models are developed by considering the subcellular localization problem as a binary classification problem where each of the models decides if a protein localizes to any of the subcellular locations in the corresponding subcellular location group (explained in Chapter 3) or not. Each classification model predicts subcellular localization of proteins by following four steps: Feature extraction and normalization, prediction by probabilistic models, weighted-mean voting, and thresholding. In the feature extraction process, 7 protein descriptors are selected out of 160 cases (40 descriptors from three tools: iFeature, POSSUM, SPMap and 4 normalization methods), which contribute the best in the combination of probabilistic prediction models. Support Vector Machine (SVM) is used to construct probabilistic prediction models, which produces probabilistic scores indicating the localization probability for a query protein sequence. A weighted score is calculated based on the obtained probabilistic scores from seven feature-based probabilistic prediction models (SVMs) by employing weighted mean voting. Binary prediction is given by applying thresholding on the weighted score.

In this chapter, the following topics are explained in detail:

- Construction of the classification models.
- Components of the constructed classification models for nine subcellular loca-

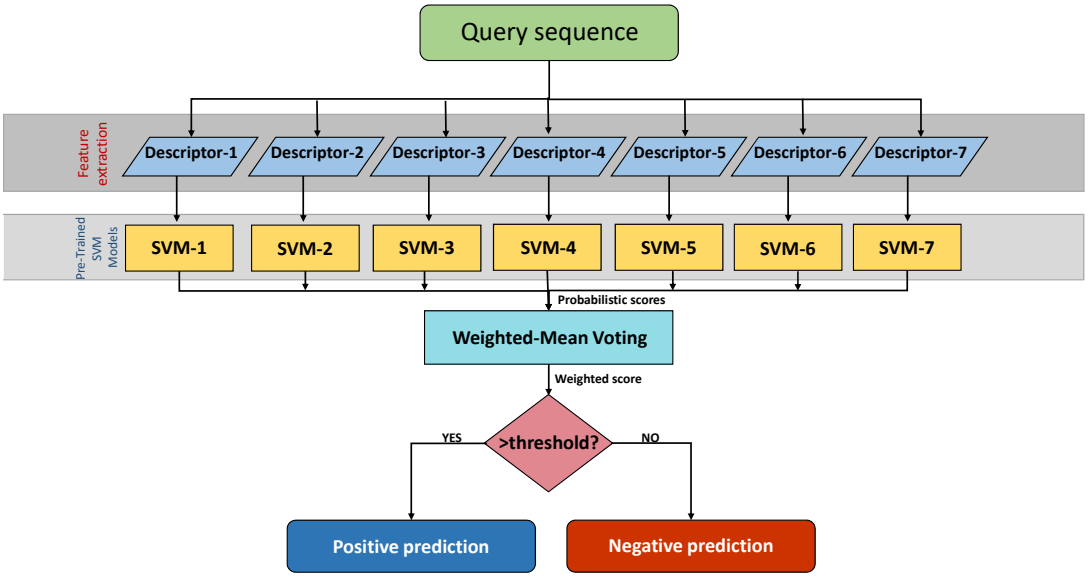

Figure 4.1: Shematic representation of the classification models for the subcellular localization prediction of human proteins.

tion groups.

- Prediction process of the constructed classfication models.
- Performance metrics to evaluate the performance of the classification models.

4.1 Construction of classification models

The methodology behind the classification models is to find the best combination of protein descriptors that represent the protein sequences with the most discriminative combination of features. Therefore, we first find the best combinations of three descriptors. We then repeatedly search two more most contributing descriptors that significantly increase classification performance to find the most discriminative combinations of five, seven, even nine descriptors. However, after finding the combinations of nine descriptors, the classification performance decreases. Hence we decided

on using the features of seven descriptors to represent protein sequences. The details to construct classification models are explained as follows:

Construction Step-1. Generating the training datasets: We generate Trust dataset and divide it into Trust-Train and Trust-Test. Trust-Train dataset is employed in the development and validation process of the classification models. Since the classification models are constructed to give binary predictions, Trust-Train dataset is designed accordingly, which consists of two datasets: a dataset of positive class and a dataset of negative class. For more details, please refer to Chapter 3.

Construction Step-2. Feature extraction: 40 protein descriptors from three feature extraction tools (iFeature, POSSUM, SPMap) are employed to extract protein features. These descriptors provide numerical representations of protein sequences, which can mainly be categorized as amino acid composition-based, homology-based, mixture of these two, and subsequence based. For more information please refer to Chapter 3

Construction Step-3. Feature normalization: The feature normalization is vital to develop accurate classification models since most of the machine learning algorithms assume that features of a dataset are within a standard range. Therefore, we utilize four different feature normalization methods which are explained in Section-3.6.

Construction Step-4. Hyperparameter optimization of SVMs: The choice of hyperparameters has a high impact on the classification performance of SVMs. Therefore a grid search is performed to find the best values for C and γ hyperparameters of SVMs for each of 160 different numerical representations of the proteins (4 normalization methods applied on the features of 40 protein descriptors) where radial basis function(RBF) is used for kernel. Protein features are formed by applying 4 normalization methods on the features of 40 protein descriptors. Accordingly, a grid search on SVM hyperparameters for all 160 cases (4 normalization methods apllied on 40 descriptors) is conducted by employing the k-fold cross-validation technique in

Trust-Train dataset, where k is used as 10. The values that we use in the grid search of hyperparameters are depicted in Table 4.1. $scale$ value of γ indicates $\frac{1}{n*var(X)}$, where n represents number of features in the protein descriptor and $var(X)$ variance of the data formed by using the corresponding protein descriptor.

Table 4.1: Hyperparameter space of SVM.

Hyperparameter	**Values**
C	$0.01, 0.1, 1, 10, 100, 1000, 10000$
γ	$scale, 100, 10, 1, 0.1, 0.01, 0.001, 0.0001, 0.00001$

Construction Step-5. Determining the weight of feature-based probabilistic prediction models: After deciding on the hyperparameter values of SVMs, we train and evaluate the performances of feature-based probabilistic prediction models (SVMs) by employing the *k*-fold cross-validation technique (k=10) in Trust-Train dataset. We obtain the probabilistic scores by feeding the protein features of the corresponding protein descriptors to SVMs for the validation datasets. Binary classification is performed on the probabilistic scores of the proteins in the validation datasets by applying thresholding for the threshold values from 0 to 1 increased by 0.01. The performance of each feature-based probabilistic prediction model is calculated for different threshold values based on the prediction results after the thresholding. Overall Mathews Correlation Coefficient(MCC) scores are calculated to evaluate the performance of each feature-based probabilistic prediction models for all validation datasets by using *k*-fold cross-validation in Trust-Train dataset. Finally, these overall MCC scores are used as the weights of the feature-based probabilistic prediction models.

Construction Step-6. Search for the best seven protein descriptors: We aim to maximize the performance of the classification model by finding the best threshold and the best combination of seven protein descriptors.

Since it is computationally costly to find the best combination of seven protein descriptors, we employ another approach which has six stages as follows:

Construction Step-6.1: Finding the best performing combination of three protein descriptors: In this stage, we aim to find the best combination of three protein descriptors by employing the *k*-fold cross-validation technique in Trust-Train dataset. Compare to the search of seven; it is less costly to search the best combinations of three for 160 cases (4 normalization methods applied on 40 descriptors), where we employed an exhaustive search for this process. 10-fold cross-validation technique in Trust-Train dataset is employed to evaluate performance for each combination of triple protein descriptors as follows:

1. The protein features from the protein sequences in Trust-Train dataset are extracted by using the protein descriptors in the triple combination. The features are scaled by employing the normalization methods that are pre-determined in the triple combination.

2. The probabilistic prediction models (SVMs) are first trained by using the training part of Trust-Train dataset, and the probabilistic scores are then obtained for the sequences in the validation part of Trust-Train dataset by feeding the protein features to SVMs.

3. The weighted-score is calculated by adding up products of seven pre-determined weights and the probabilistic scores and dividing the sum of products by the sum of the weights as diplayed in Equation 41.

4. The protein sequences are classified as a positive prediction or negative prediction by applying thresholding on the weighted-scores for all thresholds from 0 to 1 by increasing 0.01 as depicted in Equation 42.

5. The performance for the triple combination is evaluated by calculating the overall MCC (Equation 47) for all the validation datasets of the 10-fold cross-validation in Trust-Train dataset.

6. We select a hundred of the highest MCC scoring combinations of three protein descriptors.

Construction Step-6.2: Finding the best performing combination of five protein descriptors: In this stage, to find the best combinations of five protein descriptors,

we search for the most contributing two more protein descriptors to the selected combinations of three protein descriptors. The steps of this process are the same as the ones defined in Construction Step-6.1. After calculating MCC scores, we select a hundred of the highest MCC scoring combinations of five protein descriptors.

Construction Step-6.3: Finding the best performing combination of seven protein descriptors: In this stage, to find the best combinations of seven protein descriptors, we search for the most contributing two more protein descriptors to the selected combinations of five protein descriptors. The steps for this process are the same as the ones defined in Construction Step-6.1.

We apply the steps above to construct classification models for nine subcellular location groups (NUC, CYT, MEM, EXC, MIT, ERE, GLG, LYS, PEX), and indepedently determine the best combination of seven protein descriptors, the probabilistic prediction models, the weights and the threshold based on the MCC scores of 10-fold cross-validation in Trust-Train dataset for each classification model. The following section illustrates the constructed classification models and their components.

Each classification model gives binary predictions individually for nine subcellular location groups that are CYT, NUC, MEM, MIT, ERE, EXC, GLG, LYS and PEX. Figure 4.1 illustrates the schematic representation of the classification model for each of the subcellular location groups.

4.2 Classification models for nine subcellular location groups

The classification models for nine groups of subcellular locations are constructed by following the steps in Section 4.1. Each classification model has four components: the protein descriptors, the hyperparameters of SVM, the weights and the threshold. The classification models and their components are illustrated in the tables, from Table 4.2 to Table 4.10. The classification model for NUC proteins is explained in details. The other classification models and their components are given in the tables. Since the process of classification is the same as the one in NUC for other locations we just present the classification process for NUC proteins step by step.

4.2.1 Classification model to predict subcellular localization of NUC proteins:

The classification model gives binary predictions whether the proteins localize to any of the subcellular locations in the group of NUC or not. Table 4.2 depicts the employed protein descriptors, the normalization methods, the hyperparameter of SVMs, the weights, and the threshold. We explain the prediction process by illustrating the use of the components as follows:

Prediction Step-1: Feature extraction: Seven protein features are extracted from the query sequence. These descriptors are PSSM-CC, tri-gram-PSSM, DP-PSSM, tri-gram-PSSM, SPMAP, smoothed-PSSM, and Pse-PSSM. The feature extraction methods are detailed in Chapter 3.

Prediction Step-2: Feature normalization: Four normalization methods that we use are explained in Chapter 3. The protein features were normalized as follows:

1. PSSM-CC features are normalized by using Robust Scaler.
2. tri-gram-PSSM features are normalized by using Power transformation.
3. DP-PSSM features are normalized by using Power transformation.
4. tri-gram-PSSM features are normalized by using Power transformation.
5. SPMAP features are normalized by using MinMax normalization.
6. smoothed-PSSM features are normalized by using Standardization.
7. Pse-PSSM features are normalized by using Power transformation.

Prediction Step-3: Obtaining probabilistic scores: Pre-trained SVMs are fed with the protein features of the corresponding protein descriptors, and the probabilistic scores are obtained. For instance, PSSM-CC protein features are used to be fed to SVM whose C hyperparameter is 10, and γ is scale. The same process is applied to all of the normalized protein features of seven protein descriptors as shown in Table 4.2.

Prediction Step-4: Weighted-Mean Voting: The weighted score are calculated by adding the products of the weights and probabilistic scores. For instance, the weight for PSSM-CC is 0.45, and it is 0.50 for DP-PSSM. The formula to calculate weighted score is given in Equation 41

Prediction Step-5: Thresholding: The determined threshold is 0.60 (explained in Section 4.1) to give binary predictions whether query sequence localizes to NUC or not. Equation 42 illustrates the thresholding process.

Table 4.2: The components for the classification model to predict subcellular localization of NUC proteins.

Components/Models		**Model-1**	**Model-2**	**Model-3**	**Model-4**	**Model-5**	**Model-6**	**Model-7**
Protein features		PSSM-CC	tri-gram-PSSM	DP-PSSM	tri-gram-PSSM	SPMAP	smoothed-PSSM	Pse-PSSM
Normalization methods		Robust	Power	Power	Standard	MinMax	Standard	Power
SVM	**C**	10	10	10	10	100	1	1
	γ	scale	scale	0.001	scale	0.1	scale	scale
Weights		0.14	0.17	0.15	0.16	0.09	0.13	0.16
Threshold		0.6						

4.2.2 Classification model to predict subcellular localization of CYT proteins

The classification model gives binary predictions whether the proteins localize to any of the subcellular locations in the group of CYT or not. The pediction process (explained in Section 4.2.1) is followed for subcelllular location prediction of CYT proteins by employing the components illustrated in Table 4.3.

Table 4.3: The components for the classification model to predict subcellular localization of CYT proteins.

Components/Models		**Model-1**	**Model-2**	**Model-3**	**Model-4**	**Model-5**	**Model-6**	**Model-7**
Protein features		DP-PSSM	tri-gram-PSSM	Pse-PSSM	DPC-PSSM	GDPC	SPMAP	SPMAP
Normalization methods		Standard	Standard	MinMax	Robust	MinMax	Standard	Robust
SVM	**C**	1	100	10000	1000	1	1	1
	γ	scale	1e-05	0.01	1e-05	1	0.001	1e-05
Weights		0.18	0.17	0.18	0.16	0.12	0.10	0.10
Threshold		0.63						

4.2.3 Classification model to predict subcellular localization of MEM proteins

The classification model gives binary predictions whether the proteins localize to any of the subcellular locations in the group of MEM or not. The pediction process (explained in Section 4.2.1) is followed for subcelllular location prediction of MEM proteins by employing the components illustrated in Table 4.4.

Table 4.4: The components for the classification model to predict subcellular localization of MEM proteins.

Components/Models		**Model-1**	**Model-2**	**Model-3**	**Model-4**	**Model-5**	**Model-6**	**Model-7**
Protein features		SOCNumber	PSSM-CC	CKSAAP	EDP	NMBroto	CKSAAP	PSSM-AC
Normalization methods		Power	Robust	Power	Robust	Robust	MinMax	Robust
SVM	**C**	10	10	1	1000	10	10	1
	γ	0.001	0.0001	scale	scale	0.001	0.001	0.001
Weights		0.13	0.15	0.16	0.11	0.14	0.16	0.15
Threshold		0.61						

4.2.4 Classification model to predict subcellular localization of EXC proteins

The classification model gives binary predictions whether the proteins localize to any of the subcellular locations in the group of EXC or not. The pediction process (explained in Section 4.2.1) is followed for subcelllular location prediction of EXC proteins by employing the components illustrated in Table 4.5.

Table 4.5: The components for the classification model to predict subcellular localization of EXC proteins.

Components/Models		**Model-1**	**Model-2**	**Model-3**	**Model-4**	**Model-5**	**Model-6**	**Model-7**
Protein features		S-FPSSM	CTDD	PSE-PSSM	PSSM-CC	RPSSM	EEDP	SMOOTHED-PSSM
Normalization methods		Power	Standard	Robust	Standard	MinMax	Standard	Standard
SVM	**C**	10	1	1	100	100	1000	1
	γ	0.0001	scale	0.1	0.0001	scale	1e-05	scale
Weights		0.14	0.10	0.16	0.13	0.14	0.16	0.16
Threshold		0.55						

4.2.5 Classification model to predict subcellular localization of MIT proteins

The classification model gives binary predictions whether the proteins localize to any of the subcellular locations in the group of MIT or not. The pediction process (explained in Section 4.2.1) is followed for subcelllular location prediction of MIT proteins by employing the components illustrated in Table 4.6.

Table 4.6: The components for the classification model to predict subcellular localization of MIT proteins.

Components/Models		**Model-1**	**Model-2**	**Model-3**	**Model-4**	**Model-5**	**Model-6**	**Model-7**
Protein features		SMOOTHED-PSSM	PAAC	PSSM-CC	SMOOTHED-PSSM	NMBroto	CTDD	EEDP
Normalization methods		MinMax	Power	Robust	Standard	Standard	Standard	Power
SVM	**C**	1	1000	100	1	1	100	10
	γ	0.1	0.0001	1e-05	scale	scale	0.0001	0.001
Weights		0.16	0.15	0.14	0.16	0.09	0.12	0.18
Threshold		0.61						

4.2.6 Classification model to predict subcellular localization of ERE proteins

The classification model gives binary predictions whether the proteins localize to any of the subcellular locations in the group of ERE or not. The pediction process (explained in Section 4.2.1) is followed for subcelllular location prediction of ERE proteins by employing the components illustrated in Table 4.7.

Table 4.7: The components for the classification model to predict subcellular localization of ERE proteins.

Components/Models		**Model-1**	**Model-2**	**Model-3**	**Model-4**	**Model-5**	**Model-6**	**Model-7**
Protein features		MEDP	APAAC	D-FPSSM	SMOOTHED-PSSM	AAC-PSSM	PSSM-CC	SMOOTHED-PSSM
Normalization methods		Robust	Standard	MinMax	Power	Power	Robust	Standard
SVM	**C**	1	1	1	0.1	1	10	1
	γ	0.01	0.01	10	scale	0.1	scale	1e-05
Weights		0.16	0.15	0.14	0.12	0.18	0.15	0.12
Threshold		0.61						

4.2.7 Classification model to predict subcellular localization of GLG proteins

The classification model gives binary predictions whether the proteins localize to any of the subcellular locations in the group of GLG or not. The pediction process (explained in Section 4.2.1) is followed for subcelllular location prediction of GLG proteins by employing the components illustrated in Table 4.8.

Table 4.8: The components for the classification model to predict subcellular localization of GLG proteins.

Components/Models		**Model-1**	**Model-2**	**Model-3**	**Model-4**	**Model-5**	**Model-6**	**Model-7**
Protein features		PSSM-CC	SMOOTHED-PSSM	Geary	k-separated-bigrams-PSSM	RPSSM	DP-PSSM	k-separated-bigrams-PSSM
Normalization methods		Robust	MinMax	Standard	Standard	MinMax	MinMax	MinMax
SVM	**C**	10	1	10	100	10	10	10
	γ	1e-05	0.1	0.01	0.1	1	1	10
Weights		0.12	0.16	0.01	0.20	0.14	0.18	0.19
Threshold		0.62						

4.2.8 Classification model to predict subcellular localization of LYS proteins

The classification model gives binary predictions whether the proteins localize to any of the subcellular locations in the group of LYS or not. The pediction process (explained in Section 4.2.1) is followed for subcelllular location prediction of LYS proteins by employing the components illustrated in Table 4.9.

Table 4.9: The components for the classification model to predict subcellular localization of LYS proteins.

Components/Models		**Model-1**	**Model-2**	**Model-3**	**Model-4**	**Model-5**	**Model-6**	**Model-7**
Protein features		PSSM-COMPOSITION	NMBroto	SPMap	EDP	SMOOTHED-PSSM	CKSAAP	S-FPSSM
Normalization methods		Standard	Robust	Power	Robust	MinMax	Standard	Standard
SVM	**C**	10	1	1	1	10	10	100
	γ	0.001	0.01	scale	scale	0.1	0.0001	0.01
Weights		0.22	0.11	0.14	0.08	0.16	0.15	0.14
Threshold		0.59						

4.2.9 Classification model to predict subcellular localization of PEX proteins

The classification model gives binary predictions whether the proteins localize to any of the subcellular locations in the group of PEX or not. The pediction process (explained in Section 4.2.1) is followed for subcelllular location prediction of PEX proteins by employing the components illustrated in Table 4.10.

Table 4.10: The components for the classification model to predict subcellular localization of PEX proteins.

Components/Models		**Model-1**	**Model-2**	**Model-3**	**Model-4**	**Model-5**	**Model-6**	**Model-7**
Protein features		SMOOTHED-PSSM	EEDP	PSSM-CC	APAAC	SPMap	MEDP	CTDD
Normalization methods		Standard	MinMax	Power	Robust	MinMax	Standard	MinMax
SVM	**C**	10	100	1000	10	10	1000	100
	γ	0.0001	scale	1e-05	0.001	0.1	0.0001	0.1
Weights		0.16	0.15	0.10	0.10	0.22	0.18	0.09
Threshold		0.67						

4.3 Prediction by the classification models

After training the constructed classification models, the classification models give binary predictions

The prediction process of a classification model is as follows:

Prediction Step-1. Feature extraction and normalization: The selected seven protein descriptors are employed to extract the features from the query protein sequence.

Prediction Step-2. Feature normalization: The features are normalized by using the corresponding normalization method for each of the selected protein descriptors.

Prediction Step-3. Obtaining probabilistic scores: Seven probabilistic scores are obtained from the seven pre-trained probabilistic prediction models.

Prediction Step-4. Weighted-mean voting: To calculate the weighted score, we add-up the multiplication of seven probabilistic scores and pre-determined weights, and divide by the sum of weights where each weight represents the discriminative power of the pre-trained probabilistic prediction models as displayed in the following Equation 41.

$$\psi(s) = \frac{\sum_{i=1}^{7} \omega_i * \eta_i}{\sum_{i=1}^{7} \omega_i}, \tag{41}$$

where $\psi(s)$ represents the weighted-score of the query sequence (s), ω_i represents the weights, η_i indicates the probabilistic score obtained from the probabilistic prediction model.

Prediction Step-5. Thresholding: The calculated weighted-score above threshold is considered to be a positive prediction or equal or below threshold is considered to be a negative prediction as illustrated in the following Equation 42.

$$\phi(s) = \begin{cases} positive, & \text{if } \psi(s) > T \\ negative, & \text{if } \psi(s) \leq T, \end{cases} \tag{42}$$

where $\phi(s)$ indicates prediction resultof the query sequence (s), $\psi(s)$ represents the weighted-score and T is the treshold.

4.4 Performence metrics

To evaluate the performance of the classification models, we employ five commonly used metrics [48], which are accuracy, precision, recall, F1-score, and Mathews correlation coefficient(MCC). These performance measures are calculated using the confusion matrix values which true positive, false negative, true negative, and false positive:

- True positive(TP) represents the number of positive predictions whose actual-value is positive.
- False negative(FN) represents the number of negative predictions whose actual-value is positive.

- True negative(TN) represents the number of negative predictions whose actual-value is negative.
- False positive(FP) represents the number of positive predictions whose actual-value is negative.

Table 4.11: Confusion Matrix.

		Predicted	
		Positive	Negative
Actual	Positive	TP	FN
	Negative	FP	TN

The performance measures that we employ describe the different aspects of prediction performance.

- Accuracy measures the strength of a predictor in classifying all samples correctly, no matter it is positive or negative. Accuracy can be defined as:

$$accuracy = \frac{TP + TN}{TP + FN + TN + FP} \tag{43}$$

- Precision answers the question "What proportion of positive predictions was actually correct?". The formula of precision is as follows:

$$precision = \frac{TP}{TP + FP} \tag{44}$$

- Recall answers the question "What proportion of actual positives was predicted correctly?". Recall can be formulated as follows:

$$recall = \frac{TP}{TP + FN} \tag{45}$$

- F1score is a measure to balance the performance evaluation between precision and recall. It considers both false positive and false negative in the performance evaluation. It differs from accuracy because the cost of a false positive and false negative is different in the calculation of F1 score, while accuracy is the right choice if both false positive and false negative have similar cost. F1 score can be calculated as follows:

$$F1 = \frac{2 * (precision * recall)}{precision + recall} \tag{46}$$

- MCC widely applied for evaluation of the classifier when testing datasets are imbalanced. MCC score ranges between -1 and 1. 0 indicates that the performance of the predictor is the same as a random guess, while 1 indicates that the classification model has a perfect performance of correctly classifying. -1 means that the classifier gives perfect predictions if the prediction output is interpreted as the opposite. MCC formula is as follows:

$$MCC = \frac{(TP * TN - FP * FN)}{\sqrt{(TP + FP) * (TP + FN) * (TN + FP) * (TN + FN)}} \quad (47)$$

CHAPTER 5

RESULTS

In this chapter, we present performance evaluations of the proposed classification models for nine subcellular location groups, where the method of the proposed classification models is named as CanSLPred. The classification models of CanSLPred are independently constructed for nine subcellular location groups. During the construction process of the classification models, the k-fold cross-validation (k=10) technique is employed in Trust-Train dataset for each subcellular location group. Therefore, we first illustrate the 10-fold cross-validation performances of CanSLPred. We then compare the performances of CanSLPred for nine subcellular locations groups with five state-of-the-art methods: MultiLoc2 [29], LocTree2 [31], CELLO2.5 [30], SubCons [7], and DeepLoc [32]. The performances are evaluated with three datasets: Trust-Test dataset, Golden dataset, and Golden-Trust datasets (explained in Chapter 3). DeepLoc is evaluated on only Trust-Test dataset since Trust-Test dataset is generated by excluding the protein sequences in DeepLoc's training dataset whereas Golden dataset is not regenerated accordingly not to ruin the orginality of the dataset. We employ accuracy, precision, recall, F1-score, and Mathews Correlation Coefficient(MCC) as performance metrics (explained in Chapter 4).

5.1 10-fold cross-validation results in Trust-Train datasets of nine subcellular location groups

In the construction process of the classification models, we employ k-fold cross-validation technique (k=10) in Trust-Train dataset of each subcellular location

group. The 10-fold cross-validation performance of the constructed classification models is first evaluated by using Trust-Train dataset. The performance results are shown in Table 5.1 and Figure 5.1 for nine subcellular location groups. The results indicate that our method (CanSLPred) achieves 74% average accuracy, 87% average precision, 53% average recall, 65% average F1-Score, and 52% average MCC. Moreover, CanSLPred excels in the precision results for all subcellular locations with more than 79% precisions and the accuracy results indicate that CanSLPred is capable of classifying correctly 74% of human proteins in Trust-Train dataset.

Table 5.1: Performance results of CanSLPred by employing 10-fold cross-validation in Trust-Train datasets of nine subcellular location groups.

SLs/Metrics	TP	FN	FP	TN	Accuracy	Precision	Recall	F1-score	MCC
NUC	917	382	179	1120	0.78	0.84	0.71	0.77	0.58
CYT	304	305	52	553	0.71	0.85	0.50	0.63	0.45
MEM	440	245	11	674	0.81	0.97	0.64	0.77	0.66
EXC	258	53	32	280	0.86	0.89	0.82	0.85	0.73
MIT	187	132	11	308	0.78	0.94	0.58	0.72	0.59
ERE	156	210	10	356	0.70	0.94	0.43	0.57	0.48
GLG	108	206	7	307	0.66	0.93	0.34	0.50	0.41
LYS	73	137	18	192	0.63	0.79	0.36	0.47	0.32
PEX	29	35	1	63	0.72	0.80	0.41	0.53	0.44
Overall					**0.74**	**0.87**	**0.53**	**0.65**	**0.52**

5.2 Performance comparison of CanSLPred with the other five methods:

The performance of our method is compared with five state-of-the-art methods that are Multiloc2, LocTree2, Cello2.5, SubCons, and Deeploc. The prediction results for MultiLoc2 [29], LocTree2 [31], CELLO2.5 [30], and SubCons [7] are obtained from SubCons' web server, and DeepLoc [32] prediction results are obtained from its web server.

We employ three datasets (Trust-Test, Golden, Golden-Trust) to evaluate the performances of the methods mentioned above. Trust-Test is our in-house

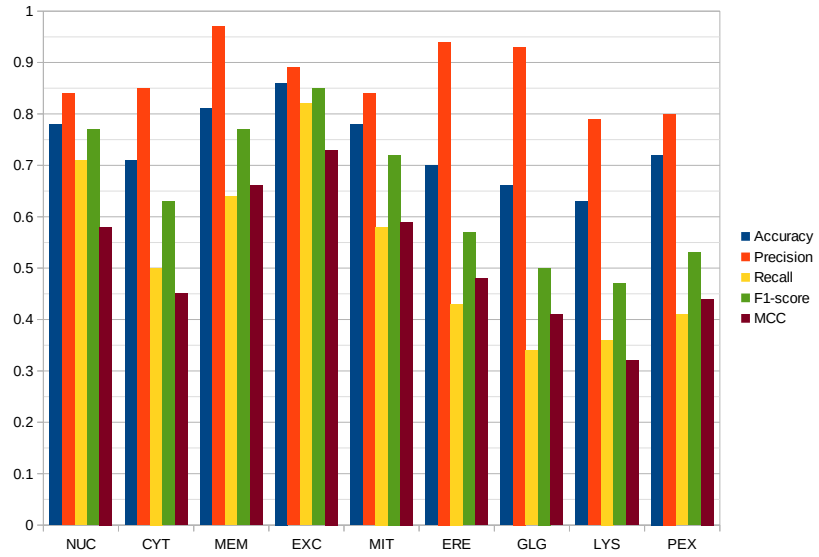

Figure 5.1: Performance results of CanSLPred by employing 10-fold cross-validation in Trust-Train dataset of nine subcellular location groups.

dataset whereas Golden dataset is the benchmark dataset of SubCons [7] and Golden-Trust is a refined version of Golden dataset. The details about the datasets are explained in Chapter 3. The classification models are independently evaluated for nine subcellular location groups as follows:

5.2.1 Performance evaluation and comparison of the methods for NUC proteins

The performance of CanSLPred is evaluated for classifying NUC proteins by employing three datasets: Trust-Test dataset, Golden dataset, Golden-Trust dataset, and is compared with the mentioned five other methods.

5.2.1.1 Performance evaluation in Trust-Test dataset of NUC

Our method (CanSLPred) outperformes the other methods with 0.55 MCC score and 0.81 precision on Trust-Test dataset since the second-highest scores

are 0.49 MCC score and 0.77 precision from SubCons Reliability. Trust-Test dataset of NUC contains 300 positive samples and 300 negative samples where our method correctly predicted the subcellular localization of 465 out of 600 proteins. Additionally, our method achieves the highest F1-score which indicates the balance between precision and recall. The performance of the predictors in Trust-Test dataset of NUC location is shared in Table 5.2. Figure 5.2 is the column chart to illustrate the performances of the methods where the methods are put in order according to their MCC scores in Trust-Test dataset of NUC.

Table 5.2: Performance results of the methods for the proteins in Trust-Test dataset of NUC.

Methods/Metrics	TP	FN	FP	TN	Accuracy	Precision	Recall	F1-score	MCC
MultiLoc2	57	243	11	286	0.58	0.84	0.19	0.31	0.24
LocTree2	202	98	75	225	0.71	0.73	0.67	0.7	0.42
CELLO2.5	188	112	75	225	0.69	0.71	0.63	0.67	0.38
SubCons Realibity	235	65	90	210	0.74	0.72	**0.78**	0.75	0.49
SubCons RF	133	167	39	261	0.66	0.77	0.44	0.56	0.35
DeepLoc	136	164	16	284	0.7	**0.89**	0.45	0.6	0.46
CanSLPred	214	86	49	251	**0.78**	0.81	0.71	**0.76**	**0.55**

5.2.1.2 Performance evaluation in Golden dataset of NUC

Table 5.3 displays the performances of the methods in Golden dataset. Golden dataset consists of 733 proteins that localize NUC subcellular location and 492 proteins for other locations. Our method (CanSLPred) achieves the highest MCC of 0.69, whereas SubCons-RF scored 0.68 regarding MCC. In the comparison of precision scores, although MultiLoc2 attaines the highest score of 0.92, it fails to predict correctly 443 of NUC proteins out of 733, which results in a very low recall of 0.40. However, our method accomplishes a balance between precision and recall with the scores of 0.87 and 0.88 respectively. Figure 5.3 is the column chart to illustrate the performances of the methods where the methods are put in order according to their MCC scores in Golden dataset.

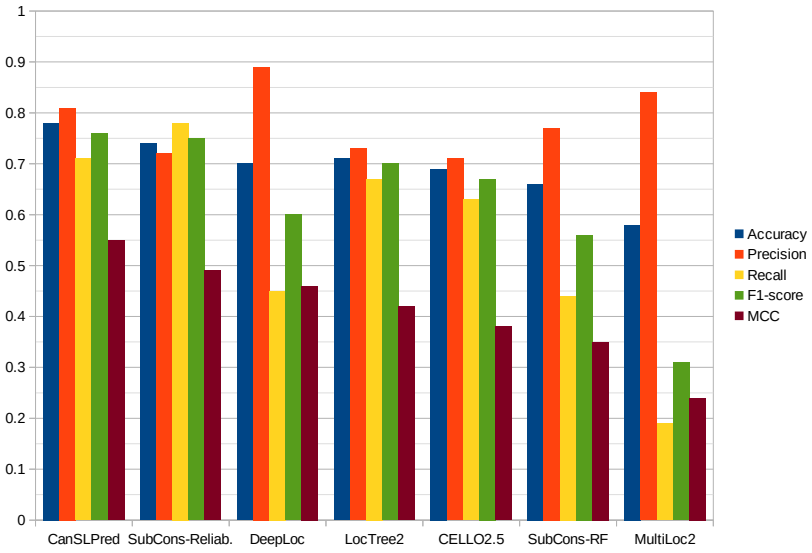

Figure 5.2: Performance results of the methods for the proteins in Trust-Test dataset of NUC.

5.2.1.3 Performance evaluation in Golden-Trust dataset of NUC

We evaluate the performances of the methods for NUC subcellular localization prediction in Golden-Trust dataset. The performances are shown in Table 5.4. SubCons-RF achieved the best MCC(0.72), whereas our method(CanSLPred) is the second MCC(0.69). Figure 5.4 is the column chart to illustrate the performances of the methods where the methods were put in order according to their MCC scores in Golden-Trust dataset.

5.2.2 Performance evaluation and comparison of CanSLPred for CYT proteins

The performance of CanSLPred is evaluated with three datasets: Trust-Test dataset, Golden dataset, and Golden-Trust dataset, and the results are compared with the other five predictors. We present CYT performances of the predictors by using three test datasets as follows:

Table 5.3: Performance results of the methods for the proteins in Golden dataset of NUC.

Methods/Metrics	TP	FN	FP	TN	Accuracy	Precision	Recall	F1-score	MCC
MultiLoc2	290	443	24	469	0.62	**0.92**	0.4	0.56	0.39
LocTree2	650	83	111	382	0.84	0.85	0.89	0.87	0.67
CELLO2.5	638	95	144	349	0.81	0.82	0.87	0.84	0.59
SubCons Realibity	705	27	194	299	0.82	0.78	**0.96**	0.86	0.63
SubCons RF	626	106	84	409	0.84	0.88	0.86	0.87	0.68
CanSLPred	647	86	97	396	**0.85**	0.87	0.88	**0.88**	**0.69**

Table 5.4: Performance results of the methods for the proteins in Golden-Trust dataset of NUC.

Methods/Metrics	TP	FN	FP	TN	Accuracy	Precision	Recall	F1-score	MCC
MultiLoc2	79	119	10	188	0.67	**0.89**	0.4	0.55	0.42
LocTree2	173	25	43	155	0.83	0.8	0.87	0.83	0.66
CELLO2.5	169	29	55	143	0.79	0.75	0.85	0.8	0.58
SubCons Realibity	190	8	72	126	0.8	0.73	**0.96**	0.83	0.63
SubCons RF	170	28	27	171	**0.86**	0.86	0.86	**0.86**	**0.72**
CanSLPred	172	26	35	163	0.85	0.83	0.87	0.85	0.69

5.2.2.1 Performance evaluation in Trust dataset of CYT

Table 5.5 illustrates the performances of the five predictors as well as CanSLPred for CYT proteins in Trust-Test dataset. CanSLPred achieves the highest MCC score of 0.45 in the classification of CYT proteins in Trust-Test dataset. CanSLPred can correctly classify 70% of proteins in Trust-Test of CYT. SubCons Reliability reaches the highest precision score of 0.93 and Multiloc2 accomplishes the highest recall and F1-score. Figure 5.5 depicts the sorted performances of the predictors concerning their MCC scores.

5.2.2.2 Performance evaluation in Golden dataset of CYT

Table 5.6 illustrates the performances of the predictors for the subcellular localization prediction of CYT proteins in Golden dataset. Golden dataset of CYT consists of 159 CYT protein sequences and 1067 protein sequences from the

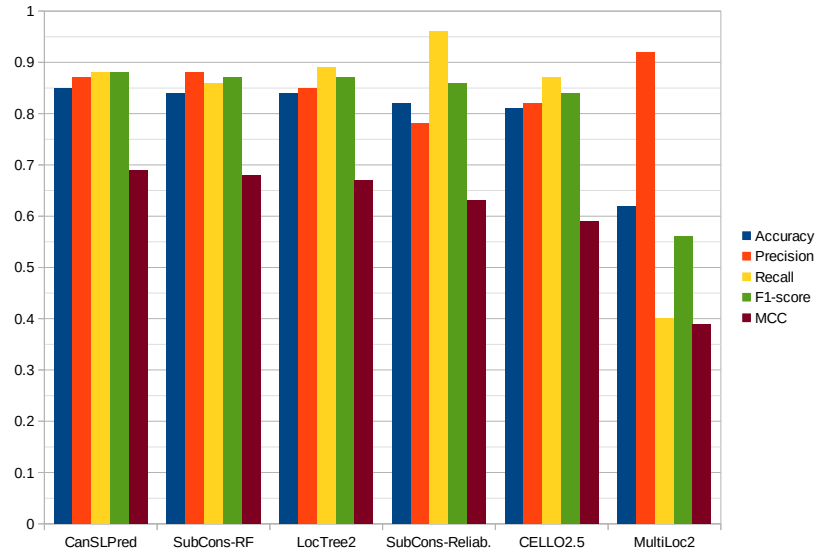

Figure 5.3: Performance results of the methods for the proteins in Golden dataset of NUC.

other seven location groups. The performance metrics: precision, recall, and F1-score are sensitive to imbalanced datasets. Therefore, the results of precision, recall, and F1-score are affected by the imbalanced Golden dataset of CYT. However, CanSLPred achieves the highest MCC score of 0.39 and the highest F1-score of 0.46, whereas SubCons Realibity reaches the highest accuracy of 0.87. Figure 5.6 depicts the sorted performances of the predictors concerning their MCC scores.

5.2.2.3 Performance evaluation in Golden-Trust dataset of CYT

Table 5.7 illustrates the performance results of the predictors. Golden-Trust dataset of CYT consists of 95 CYT protein sequences and 95 from the other location groups. The results indicate that our predictor CanSLPred achieves the highest MCC score of 0.70, whereas the second beat predictor (LocTree2) reaches 0.56. Additionally, CanSLPred can correctly classify 85% of the proteins in Golden-Trust dataset of CYT. Figure 5.7 depicts the sorted performances of the predictors concerning their MCC scores in Golden-Trust dataset.

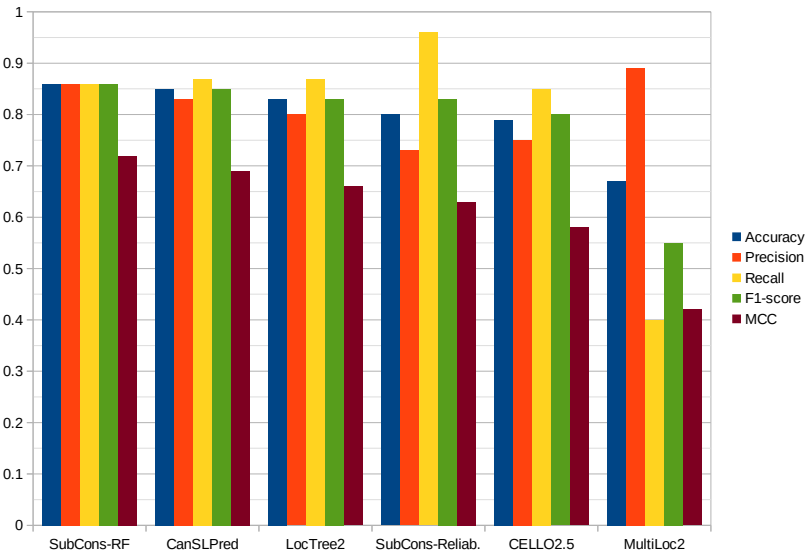

Figure 5.4: Performance results of the methods for the proteins in Golden-Trust dataset of NUC.

5.2.3 Performance evaluation and comparison of CanSLPred for MEM proteins

5.2.3.1 Performance evaluation in Trust-Test dataset of MEM

Table 5.8 illustrates the performance evaluation of the predictors for the proteins in Trust-Test of MEM. CanSLPred achieves the highest MCC score of 0.67 and can correctly predict 81% of the proteins in Trust-Test dataset of MEM. DeepLoc reaches the highest and perfect precision score of 1, whereas CanSLPred accomplishes the precision score of 0.98. Figure 5.8 depicts the sorted performances of the predictors concerning their MCC scores in Trust-Test dataset of MEM.

Table 5.5: Performance results of the methods for the proteins in Trust-Test dataset of CYT.

Methods/Metrics	TP	FN	FP	TN	Accuracy	Precision	Recall	F1-score	MCC
MultiLoc2	87	46	33	100	0.70	0.72	**0.65**	**0.68**	0.41
LocTree2	46	87	9	124	0.64	0.84	0.35	0.49	0.34
CELLO2.5	36	97	9	124	0.60	0.80	0.27	0.4	0.27
SubCons Realibity	26	107	2	131	0.59	**0.93**	0.20	0.33	0.29
SubCons RF	54	79	11	122	0.66	0.83	0.41	0.55	0.38
DeepLoc	63	70	11	122	0.70	0.85	0.47	0.61	0.44
CanSLPred	63	70	10	123	**0.70**	0.86	0.47	0.61	**0.45**

Table 5.6: Performance results of the methods for the proteins in Golden dataset of CYT.

Methods/Metrics	TP	FN	FP	TN	Accuracy	Precision	Recall	F1-score	MCC
MultiLoc2	129	30	468	599	0.59	0.22	**0.81**	0.35	0.25
LocTree2	82	77	152	915	0.81	0.35	0.52	0.42	0.32
CELLO2.5	62	97	96	971	0.84	0.39	0.39	0.39	0.30
SubCons Realibity	12	147	10	1057	**0.87**	**0.55**	0.08	0.14	0.17
SubCons RF	74	85	97	970	0.85	0.43	0.47	0.45	0.36
CanSLPred	120	39	238	829	0.77	0.34	0.75	**0.46**	**0.39**

5.2.3.2 Performance evaluation in Golden dataset of MEM

Table 5.9 illustrates the performance evaluation of the predictors for the proteins in Golden dataset of MEM. Although LocTree2 accomplishes the highest precision of 0.81, its recall value is lower than the other four predictors. CanSLPred achieves the highest MCC score of 0.65, the highest F1-score of 0.66, the highest recall of 0.68 and the highest accuracy of 0.97. Figure 5.9 depicts the sorted performances of the predictors concerning their MCC scores in Golden dataset of MEM.

5.2.3.3 Performance evaluation in Golden-Trust dataset of MEM

Table 5.10 illustrates the performance evaluation of the predictors for the proteins in Golden-Trust dataset of MEM. CanSLPred outperforms at all performance metrics for the classification of the proteins in Golden-Trust dataset

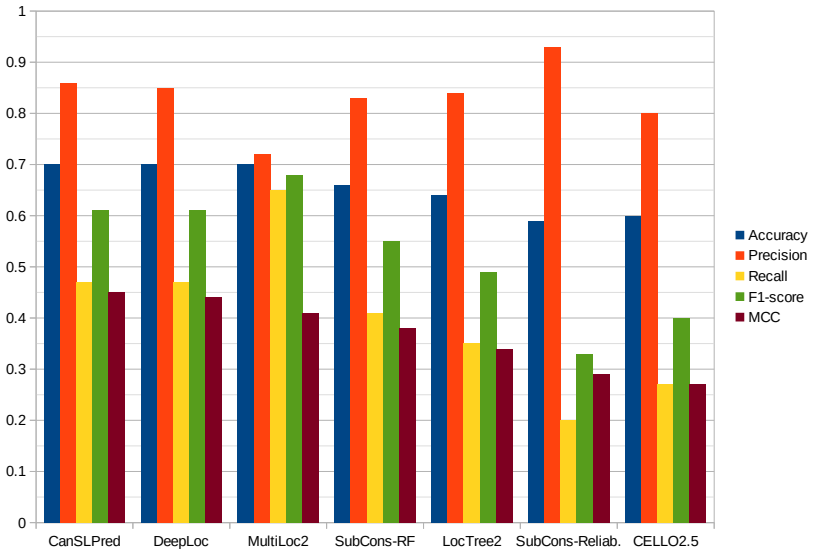

Figure 5.5: Performance results of the methods for the proteins in Trust-Test dataset of CYT.

(MEM). Figure 5.10 depicts the sorted performances of the predictors according to their MCC scores in Golden-Trust dataset of MEM.

5.2.4 Performance evaluation and comparison of CanSLPred for EXC proteins

The performances of the predictors are evaluated by using Trust-Test dataset. Golden dataset does not contain any EXC proteins. Therefore, the performances of the predictors are compared by using only Trust-Test dataset as follows:

5.2.4.1 Performance evaluation in Trust-Test dataset of EXC

Table 5.11 illustrates the performance evaluation of the predictors for the proteins in Trust-Test dataset of EXC. Although DeepLoc accomplishes a perfect score of precision, CanSLPred outshines all of the methods by achieving an MCC score of 0.86. Additionally, CanSLPred can correctly classify 93% of

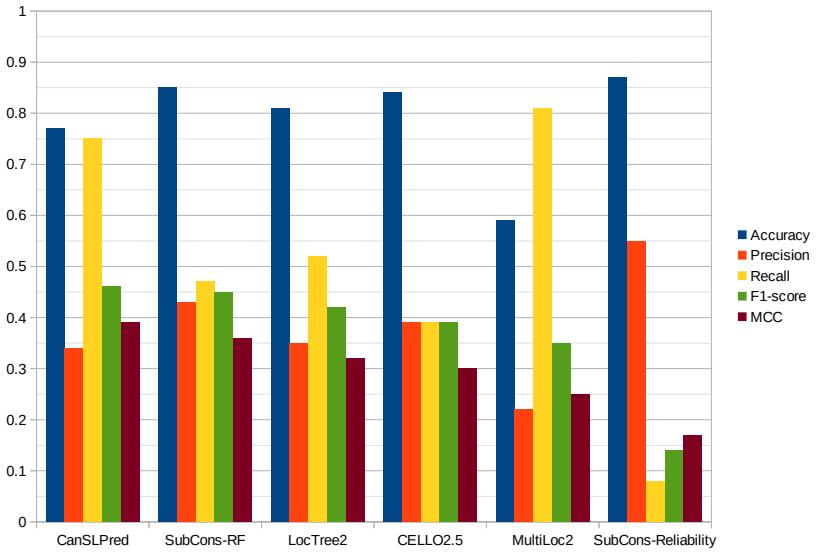

Figure 5.6: Performance results of the methods for the proteins in Golden dataset of CYT.

the proteins in Trust-Test dataset of EXC. Figure 5.11 depicts the sorted performances of the predictors according to their MCC scores in Trust-Test dataset of EXC.

5.2.5 Performance evaluation and comparison of CanSLPred for MIT proteins

The performances of the predictors are evaluated by using three datasets of MIT: Trust-Test dataset, Golden dataset, and Golden-Trust dataset, and the results are compared with the other five predictors. We present MIT performances of the predictors by using three test datasets as follows:

5.2.5.1 Performance evaluation in Trust-Test dataset of MIT

Table 5.12 illustrates the performance evaluation of the predictors for the proteins in Trust-Test dataset of MIT. The results indicate that CanSLPred and DeepLoc are capable of classifying 96% of the proteins correctly in Trust-Test

Table 5.7: Performance results of the methods for the proteins in Golden-Trust dataset of CYT.

Methods/Metrics	TP	FN	FP	TN	Accuracy	Precision	Recall	F1-score	MCC
MultiLoc2	74	17	47	48	0.66	0.62	**0.82**	0.71	0.34
LocTree2	51	44	3	92	0.75	**0.94**	0.54	0.69	0.56
CELLO2.5	36	59	9	86	0.64	0.80	0.38	0.52	0.33
SubCons Realibity	7	88	1	94	0.53	0.88	0.07	0.13	0.16
SubCons RF	47	48	6	89	0.72	0.89	0.49	0.63	0.48
CanSLPred	76	19	10	85	**0.85**	0.88	0.80	**0.84**	**0.70**

Table 5.8: Performance results of the methods for the proteins in Trust-Test dataset of MEM.

Methods/Metrics	TP	FN	FP	TN	Accuracy	Precision	Recall	F1-score	MCC
MultiLoc2	25	119	4	129	0.56	0.86	0.17	0.28	0.23
LocTree2	39	105	4	129	0.61	0.91	0.27	0.42	0.33
CELLO2.5	51	93	4	129	0.65	0.93	0.35	0.51	0.41
SubCons Realibity	78	66	5	128	0.74	0.94	0.54	0.69	0.55
SubCons RF	79	65	6	127	0.74	0.93	0.55	0.69	0.55
DeepLoc	86	58	0	133	0.79	**1**	0.60	0.75	0.64
CanSLPred	94	50	2	131	**0.81**	0.98	**0.65**	**0.78**	**0.67**

dataset of MIT. Additionally, CanSLPred reaches the highest MCC score of 0.62 whereas the closest one DeepLoc reaches 0.60 MCC score. Figure 5.12 depicts the sorted performances of the predictors according to their MCC scores in Trust-Test dataset of MIT.

5.2.5.2 Performance evaluation in Golden dataset of MIT

Table 5.13 illustrates the performance evaluation of the predictors for the proteins in Golden dataset of MIT. The results indicate that SubCons and CanSLPred are capable of classifying 94% of the proteins correctly in Golden dataset of MIT. Additionally, SubCons reaches the highest MCC score of 0.80, whereas the most imminent one is our method CanSLPred that reaches 0.79 MCC score. Figure 5.13 depicts the sorted performances of the predictors according to their MCC scores in Golden dataset of MIT.

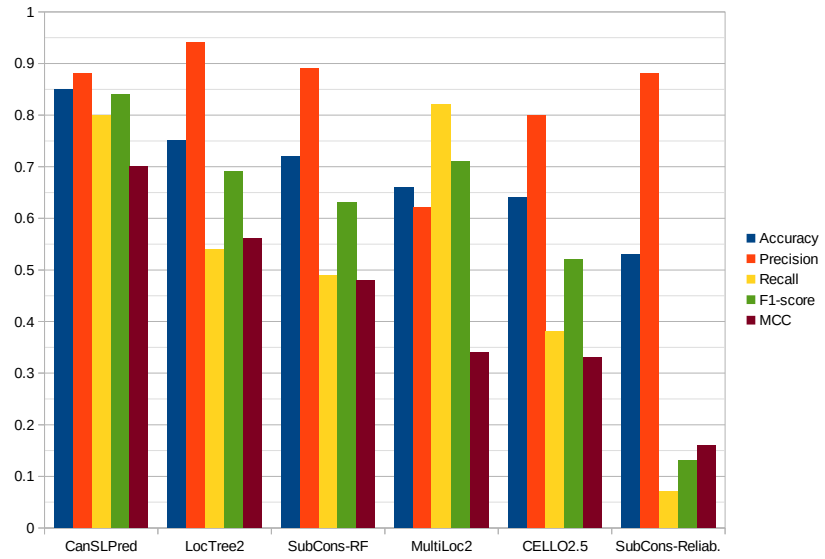

Figure 5.7: Performance results of the methods for the proteins in Golden-Trust dataset of CYT.

5.2.5.3 Performance evaluation in Golden-Trust dataset of MIT

Table 5.14 illustrates the performance evaluation of the predictors for the proteins in Golden-Trust dataset of MIT. The results indicate that CanSLPred can correctly classify 92% of the proteins in Golden-Trust dataset. Additionally, CanSLPred accomplishes the highest MCC score of 0.84 among all other predictors. Figure 5.14 depicts the sorted performances of the predictors according to their MCC scores in Golden-Trust dataset of MIT.

5.2.6 Performance evaluation and comparison of CanSLPred for ERE proteins

The performances of the predictors are evaluated by using three datasets of ERE: Trust-Test dataset, Golden dataset, and Golden-Trust dataset, and the results are compared with the other five predictors. We present ERE performances of the predictors by using three test datasets as follows:

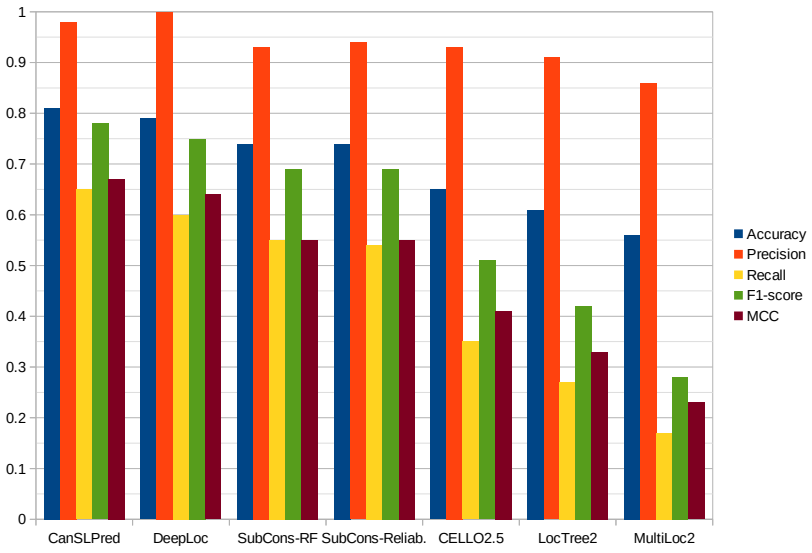

Figure 5.8: Performance results of the methods for the proteins in Trust-Test dataset of MEM.

5.2.6.1 Performance evaluation in Trust-Test dataset of ERE

Table 5.15 illustrates the performance evaluation of the predictors for the proteins in Trust-Test dataset of ERE. CanSLPred reaches the highest MCC score of 0.57 and can correctly classify 80% of the protein in Trust-Test dataset, whereas CELLO2.5 accomplishes the highest score of precision. DeepLoc is the second-best predictor which has a 0.40 MCC score in classifying ERE proteins. Figure 5.15 depicts the sorted performances of the predictors according to their MCC scores in Trust-Test dataset of ERE.

5.2.6.2 Performance evaluation in Golden dataset of ERE

Table 5.16 illustrates the performance evaluation of the predictors for the proteins in Golden dataset of ERE. SubCons significantly shows better performance than the other methods at all performance metrics but precision. Although Cello2.5 reaches the perfect precision score of 1, its recall score is the lowest. However, CanSLPred performs better than MultiLoc in terms of precision scores. Our method, CanSLPred, and MultiLoc have the second-highest

Table 5.9: Performance results of the methods for the proteins in Golden dataset of MEM.

Methods/Metrics	TP	FN	FP	TN	Accuracy	Precision	Recall	F1-score	MCC
MultiLoc2	6	41	4	1175	0.96	0.60	0.13	0.21	0.27
LocTree2	13	34	3	1176	0.97	**0.81**	0.28	0.42	0.46
CELLO2.5	18	29	18	1159	0.96	0.50	0.38	0.43	0.42
SubCons Realibity	27	20	14	1165	0.97	0.66	0.57	0.61	0.60
SubCons RF	30	17	21	1158	0.97	0.59	0.64	0.61	0.60
CanSLPred	32	15	18	1161	**0.97**	0.64	**0.68**	**0.66**	**0.65**

Table 5.10: Performance results of the methods for the proteins in Golden-Trust dataset of MEM.

Methods/Metrics	TP	FN	FP	TN	Accuracy	Precision	Recall	F1-score	MCC
MultiLoc2	5	17	0	22	0.61	1	0.23	0.37	0.36
LocTree2	6	16	0	22	0.64	1	0.27	0.43	0.40
CELLO2.5	9	13	1	21	0.68	0.90	0.41	0.56	0.43
SubCons Realibity	14	8	1	21	0.80	0.93	0.64	0.76	0.62
SubCons RF	15	7	1	21	0.82	0.94	0.68	0.79	0.66
CanSLPred	16	6	0	22	**0.86**	**1**	**0.73**	**0.84**	**0.76**

MCC score of 0.54. Figure 5.16 depicts the sorted performances of the predictors according to their MCC scores in Golden dataset of ERE.

5.2.6.3 Performance evaluation in Golden-Trust dataset of ERE

Table 5.17 illustrates the performance evaluation of the predictors for the proteins in Golden-Trust dataset of ERE. CanSLPred has the highest performance scores, but precision. Although Cello2.5 reaches the perfect precision score of 1, its recall score is the lowest (0.12). MultiLoc2 and SubCons accomplish the second-highest MCC score of 0.50, whereas CanSLPred achieves the highest MCC score of 0.60. Figure 5.17 depicts the sorted performances of the predictors according to their MCC scores in Golden-Trust dataset of ERE.

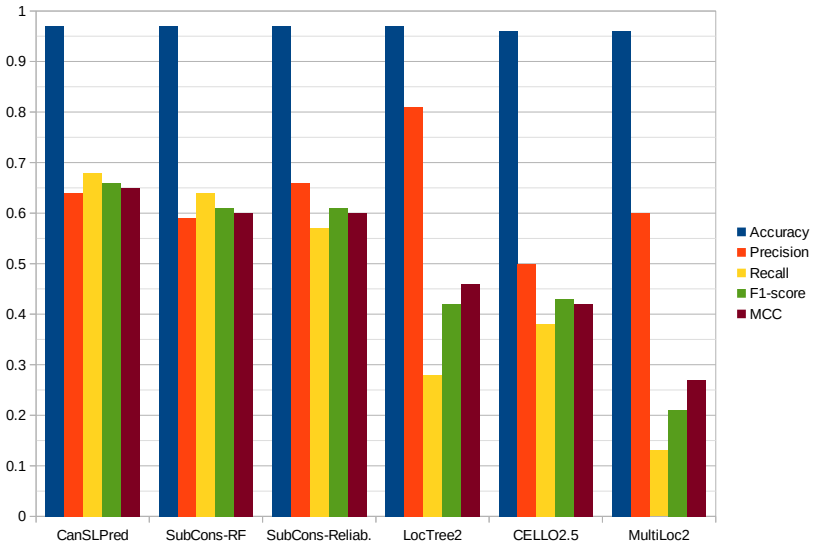

Figure 5.9: Performance results of the methods for the proteins in Golden dataset of MEM.

5.2.7 Performance evaluation and comparison of CanSLPred for GLG proteins

The performances of the predictors are evaluated by using three datasets of GLG: Trust-Test dataset, Golden dataset, and Golden-Trust dataset, and the results are compared with the other five predictors. We present GLG performances of the predictors by using three test datasets as follows:

5.2.7.1 Performance evaluation in Trust-Test dataset of GLG

Table 5.18 illustrates the performance evaluation of the predictors for the proteins in Trust-Test dataset of GLG. CanSLPred accomplishes the highest performance scores, but precision. Although MultiLoc2, CELLO2.5, and SubCons reach a perfect score of precision, they all have meager recall scores. CanSLPred is capable of classifying 65% of the proteins the dataset, whereas DeepLoc can do 56% of them. Besides, CanSLPred reaches a significantly better MCC score of 0.36 than all the other methods. Figure 5.18 depicts the sorted performances of the predictors according to their MCC scores in Trust-Test dataset of

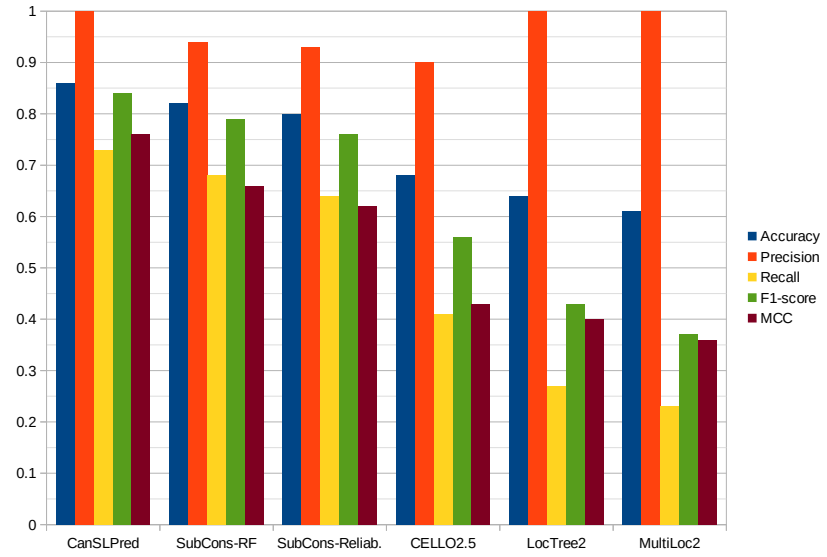

Figure 5.10: Performance results of the methods for the proteins in Golden-Trust dataset of MEM.

GLG.

5.2.7.2 Performance evaluation in Golden dataset of GLG

Table 5.19 illustrates the performance evaluation of the predictors for the proteins in Golden dataset of GLG. CanSLPred reached the highest performance scores except for precision. Although SubCons accomplishes a perfect precision score of 1, its recall is lower than our method, CanSLPred. The accuracy score indicates that CanSLPred and SubCons can correctly classify 99% of the proteins in the dataset. Figure 5.19 depicts the sorted performances of the predictors according to their MCC scores in Golden dataset of GLG.

5.2.7.3 Performance evaluation in Golden-Trust dataset of GLG

Table 5.20 illustrates the performance evaluation of the predictors for the proteins in Golden-Trust dataset of GLG. CanSLPred is well ahead of the other methods in terms of the performance scores. CanSlPred accomplishes 0.67

Table 5.11: Performance results of the methods for the proteins in Trust-Test dataset of EXC.

Methods/Metrics	TP	FN	FP	TN	Accuracy	Precision	Recall	F1-score	MCC
MultiLoc2	15	48	0	67	0.63	1	0.24	0.39	0.37
LocTree2	47	16	9	58	0.81	0.84	0.75	0.79	0.62
CELLO2.5	37	26	3	64	0.78	0.93	0.59	0.72	0.59
SubCons Realibity	37	26	0	67	0.80	1	0.59	0.74	0.65
SubCons RF	38	25	0	67	0.81	1	0.60	0.75	0.66
DeepLoc	44	19	0	67	0.85	**1**	0.70	0.82	0.74
CanSLPred	59	4	5	62	**0.93**	0.92	**0.94**	**0.93**	**0.86**

Table 5.12: Performance results of the methods for the proteins in Trust-Test dataset of MIT.

Methods/Metrics	TP	FN	FP	TN	Accuracy	Precision	Recall	F1-score	MCC
MultiLoc2	32	48	5	75	0.67	0.86	0.40	0.55	0.40
LocTree2	30	50	2	78	0.68	0.94	0.38	0.54	0.44
CELLO2.5	34	46	3	77	0.69	0.92	0.42	0.58	0.46
SubCons Realibity	39	41	6	74	0.71	0.87	0.49	0.63	0.46
SubCons RF	42	38	8	72	0.71	0.84	0.53	0.65	0.46
DeepLoc	46	34	2	78	0.78	**0.96**	0.57	0.72	0.60
CanSLPred	48	32	2	78	**0.79**	**0.96**	**0.60**	**0.74**	**0.62**

MCC score, whereas the second-highest ones (MultiLoc2, SubCons) reach 0.20. Moreover, CanSLPred can accurately classify 81% of the proteins in the dataset. Figure 5.20 depicts the sorted performances of the predictors according to their MCC scores in Golden-Trust dataset of GLG.

5.2.8 Performance evaluation and comparison of CanSLPred for LYS proteins

The performances of the predictors are evaluated by using three datasets of LYS: Trust-Test dataset, Golden dataset, and Golden-Trust dataset, and the results are compared with the other five predictors. We present LYS performances of the predictors by using three test datasets as follows:

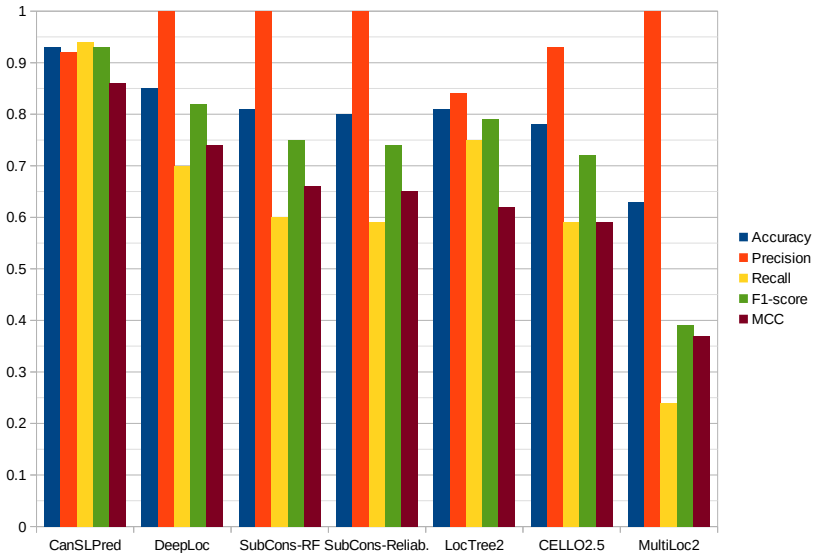

Figure 5.11: Performance results of the methods for the proteins in Trust-Test dataset of EXC.

5.2.8.1 Performance evaluation in Trust-Test dataset of LYS

Table 5.21 illustrates the performance evaluation of the predictors for the proteins in Trust-Test dataset of LYS. The results in the table indicate that CanSLPred's performance in classifying LYS proteins is significantly better than the other methods. CanSLPred accomplishes 0.59 MCC score, whereas the second-highest MCC score is 0.32 by CELLO2.5. Although SubCons and DeepLoc achieve a perfect precision score, their recall scores are inferior. Moreover, CanSLPred can accurately classify 76% of the proteins in the dataset. Figure 5.21 depicts the sorted performances of the predictors according to their MCC scores in Trust-Test dataset of LYS.

5.2.8.2 Performance evaluation in Golden dataset of LYS

Table 5.22 illustrates the performance evaluation of the predictors for the proteins in Golden dataset of LYS. CanSLPred performs significantly better than the other predictors at all performance metrics except precision. CanSLPred achieves 0.67 MCC score and can accurately classify 99% of the proteins in the

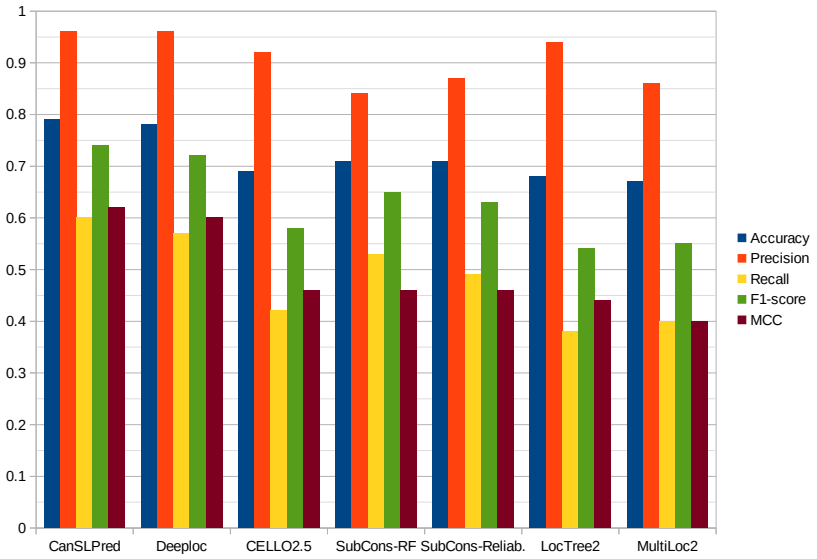

Figure 5.12: Performance results of the methods for the proteins in Trust-Test dataset of MIT.

dataset whereas MultiLoc2 reaches 0.64 MCC score. Figure 5.22 depicts the sorted performances of the predictors according to their MCC scores in Golden dataset of LYS.

5.2.8.3 Performance evaluation in Golden-Trust dataset of LYS

Table 5.23 illustrates the performance evaluation of the predictors for the proteins in Golden-Trust dataset of LYS. All the predictors except LocTree2 reach a perfect precision score of 1. CanSLPred can accurately classify 85% of the proteins and reaches the highest MCC score of 0.73, whereas MultiLoc2 has the second-highest MCC score of 0.65. Figure 5.23 depicts the sorted performances of the predictors according to their MCC scores in Golden-Trust dataset of LYS.

Table 5.13: Performance results of the methods for the proteins in Golden dataset of MIT.

Methods/Metrics	TP	FN	FP	TN	Accuracy	Precision	Recall	F1-score	MCC
MultiLoc2	145	57	39	985	0.92	0.79	0.72	0.75	0.71
LocTree2	109	93	14	1010	0.91	**0.89**	0.54	0.67	0.65
CELLO2.5	149	53	52	971	0.91	0.74	0.74	0.74	0.69
SubCons Realibity	166	36	32	992	**0.94**	0.84	0.82	0.83	**0.80**
SubCons RF	175	27	41	982	**0.94**	0.81	**0.87**	**0.84**	**0.80**
CanSLPred	175	27	47	977	**0.94**	0.79	**0.87**	0.83	0.79

Table 5.14: Performance results of the methods for the proteins in Golden-Trust dataset of MIT.

Methods/Metrics	TP	FN	FP	TN	Accuracy	Precision	Recall	F1-score	MCC
MultiLoc2	92	34	5	121	0.85	0.95	0.73	0.83	0.71
LocTree2	71	55	3	123	0.77	0.96	0.56	0.71	0.59
CELLO2.5	85	41	14	112	0.78	0.86	0.67	0.75	0.58
SubCons Realibity	106	20	6	120	0.90	0.95	0.84	0.89	0.80
SubCons RF	111	15	10	116	0.90	0.92	**0.88**	0.90	0.80
CanSLPred	107	19	2	124	**0.92**	**0.98**	0.85	**0.91**	**0.84**

5.2.9 Performance evaluation and comparison of CanSLPred for PEX proteins

The performances of the predictors are evaluated by using three datasets of PEX: Trust-Test dataset, Golden dataset, and Golden-Trust dataset, and the results are compared with the other five predictors. We present PEX performances of the predictors by using three test datasets as follows:

5.2.9.1 Performance evaluation in Trust-Test dataset of PEX

Table 5.24 illustrates the performance evaluation of the predictors for the proteins in Trust-Test dataset of PEX. SubCons achieves the highest scores at all performance metrics. SubCons can accurately classify 84% of the proteins and reaches 0.72 MCC score whereas CanSLPred achieves 0.81 accuracy score and 0.67 MCC score. Figure 5.24 depicts the sorted performances of the predictors

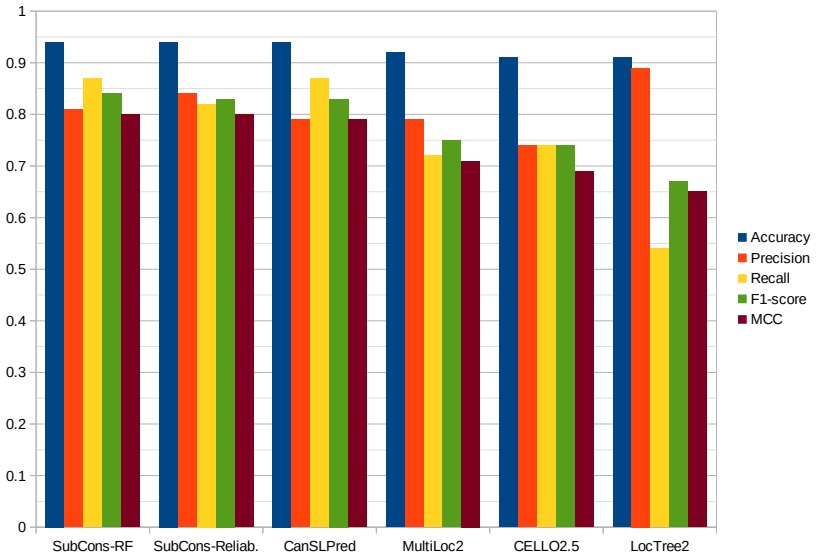

Figure 5.13: Performance results of the methods for the proteins in Golden dataset of MIT.

according to their MCC scores in Trust-Test dataset of PEX.

5.2.9.2 Performance evaluation in Golden dataset of PEX

Table 5.25 illustrates the performance evaluation of the predictors for the proteins in Golden dataset of PEX. CanSLPred reaches the highest MCC score of 0.43, whereas the second-best predictor CeLLO2.5 has 0.38 MCC score. Although CanSLPred performs low precision, it accurately classifies 1218 proteins out of 1225 in the dataset. Additionally, MutiLoc and CanSLPred accomplish a recall score of 0.43 better than the other methods. Figure 5.25 depicts the sorted performances of the predictors according to their MCC scores in Golden dataset of PEX.

5.2.9.3 Performance evaluation in Golden-Trust dataset of PEX

Table 5.26 illustrates the performance evaluation of the predictors for the proteins in Golden-Trust dataset of PEX. CanSLPred and MultiLoc2 perform the

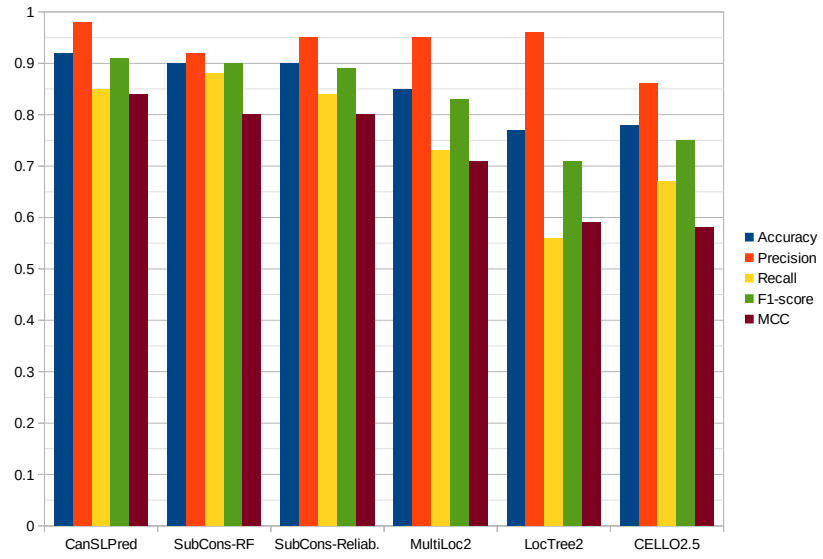

Figure 5.14: Performance results of the methods for the proteins in Golden-Trust dataset of MIT.

same in terms of all performance metrics. They reach the highest performance scores, where the MCC score is 0.52, and the precision score is 1. SubCons accomplishes 0.41 MCC, whereas CeLLO2.5 achieves 0.38 MCC score. Figure 5.26 depicts the sorted performances of the predictors according to their MCC scores in Golden-Trust dataset of PEX.

5.2.10 Comparison of the predictors in terms of MCC scores for all subcellular locations

Since MCC is one of the most reliable performance metrics to evaluate the reliability and robustness of machine learning-based methods in bioinformatics, we provide the MCC scores of the predictors for all subcellular locations together to render the comparison of the predictors. Besides, we present the overall MCC scores of each localization prediction separately for three datasets (Trust-Test, Golden dataset, Golden-Trust dataset). These overall MCC scores indicate that CanSLPred is a reliable and robust predictor since it achieves the highest overall MCC scores for all datasets. CanSLPred accomplishes the over-

Table 5.15: Performance results of the methods for the proteins in Trust-Test dataset of ERE.

Methods/Metrics	TP	FN	FP	TN	Accuracy	Precision	Recall	F1-score	MCC
MultiLoc2	16	39	8	79	0.67	0.67	0.29	0.40	0.26
LocTree2	13	42	1	86	0.70	0.93	0.24	0.38	0.37
CELLO2.5	2	53	0	87	0.63	**1**	0.04	0.08	0.15
SubCons Realibity	8	47	1	86	0.66	0.89	0.15	0.26	0.27
SubCons RF	7	48	1	86	0.65	0.88	0.13	0.23	0.24
DeepLoc	18	37	3	84	0.72	0.86	0.33	0.48	0.40
CanSLPred	31	24	5	82	**0.80**	0.86	**0.56**	**0.68**	**0.57**

Table 5.16: Performance results of the methods for the proteins in Golden dataset of ERE.

Methods/Metrics	TP	FN	FP	TN	Accuracy	Precision	Recall	F1-score	MCC
MultiLoc2	30	16	32	1148	0.96	0.48	**0.65**	0.55	0.54
LocTree2	17	29	9	1171	0.97	0.65	0.37	0.47	0.48
CELLO2.5	4	42	0	1178	0.97	**1**	0.09	0.17	0.29
SubCons Realibity	22	24	7	1173	0.97	0.76	0.48	0.59	0.59
SubCons RF	25	21	7	1173	**0.98**	0.78	0.54	**0.64**	**0.64**
CanSLPred	26	20	22	1158	0.97	0.54	0.57	0.55	0.54

all MCC scores of 0.59, 0.61 and 0.68 on Trust-Test dataset, Golden dataset, and Golden-Trust dataset respectively whereas DeepLoc achieves the second-highest overall MCC score of 0.44 on Trust-Test dataset and SubCons does the second-highest overall MCC scores of 0.56 and 0.53 on Golden dataset and Golden-Trust dataset respectively. The MCC scores and the overall MCC scores are depicted in the Tables form Table 5.27 to Table 5.29 and in the figures from Figure 5.27 to Figure 5.29.

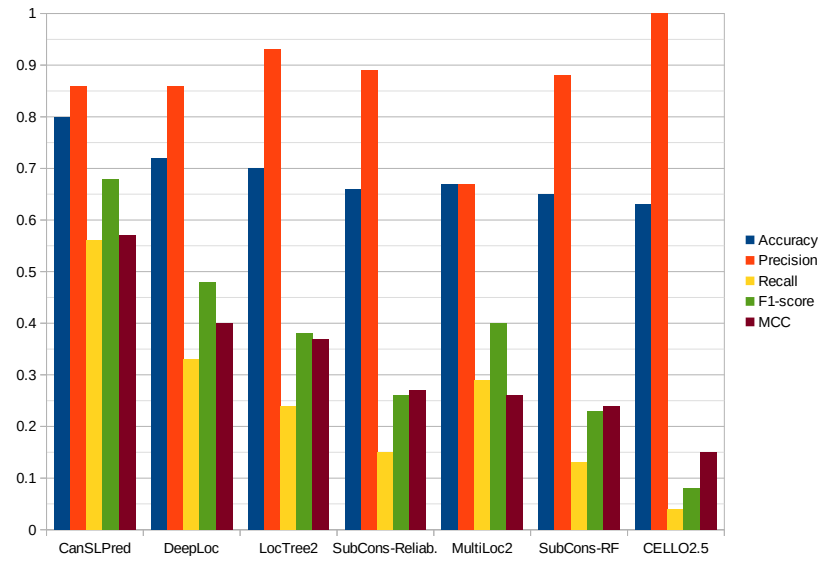

Figure 5.15: Performance results of the methods for the proteins in Trust-Test dataset of ERE.

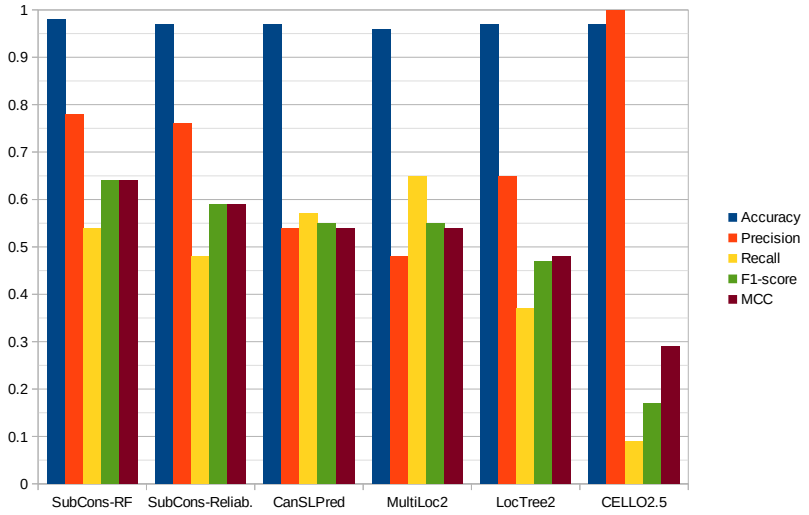

Figure 5.16: Performance results of the methods for the proteins in Golden dataset of ERE.

Table 5.17: Performance results of the methods for the proteins in Golden-Trust dataset of ERE.

Methods/Metrics	TP	FN	FP	TN	Accuracy	Precision	Recall	F1-score	MCC
MultiLoc2	15	10	3	22	0.74	0.83	0.60	0.70	0.50
LocTree2	9	16	1	24	0.66	0.90	0.36	0.51	0.40
CELLO2.5	3	22	0	25	0.54	**1**	0.12	0.21	0.25
SubCons Realibity	11	14	1	24	0.70	0.92	0.44	0.60	0.47
SubCons RF	12	13	1	24	0.72	0.92	0.48	0.63	0.50
CanSLPred	15	10	1	24	**0.78**	0.94	**0.60**	**0.73**	**0.60**

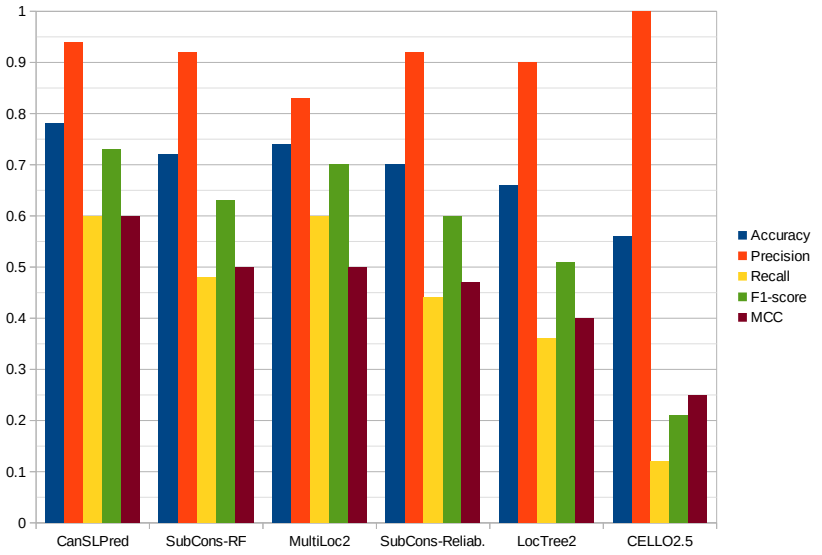

Figure 5.17: Performance results of the methods for the proteins in Golden-Trust dataset of ERE.

Table 5.18: Performance results of the methods for the proteins in Trust-Test dataset of GLG.

Methods/Metrics	TP	FN	FP	TN	Accuracy	Precision	Recall	F1-score	MCC
MultiLoc2	2	72	0	74	0.51	**1**	0.03	0.06	0.12
LocTree2	4	71	1	73	0.52	0.80	0.05	0.09	0.11
CELLO2.5	6	69	0	74	0.54	**1**	0.08	0.15	0.20
SubCons Realibity	7	68	0	74	0.54	**1**	0.09	0.17	0.22
SubCons RF	6	69	0	74	0.54	**1**	0.08	0.15	0.20
DeepLoc	10	65	1	73	0.56	0.91	0.13	0.23	0.23
CanSLPred	29	46	6	68	**0.65**	0.83	**0.39**	**0.53**	**0.36**

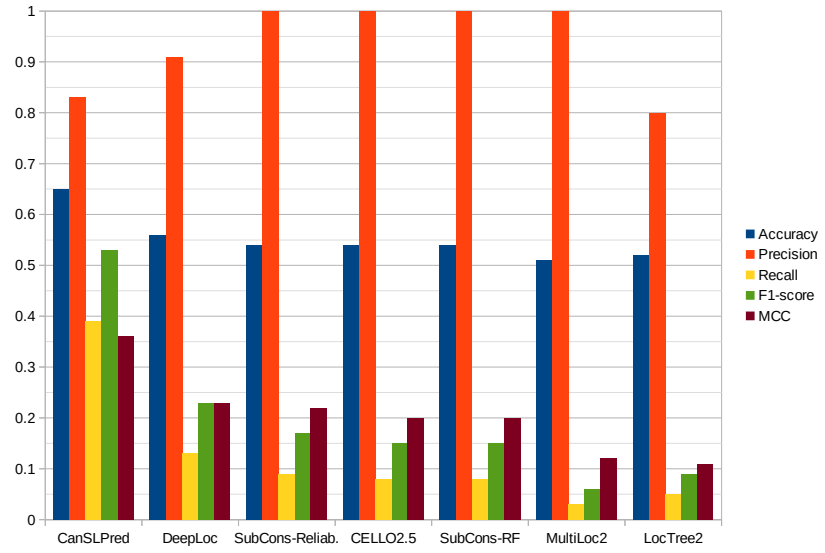

Figure 5.18: Performance results of the methods for the proteins in Trust-Test dataset of GLG.

Table 5.19: Performance results of the methods for the proteins in Golden dataset of GLG.

Methods/Metrics	TP	FN	FP	TN	Accuracy	Precision	Recall	F1-score	MCC
MultiLoc2	1	20	4	1201	0.98	0.20	0.05	0.08	0.09
LocTree2	1	20	4	1201	0.98	0.25	0.05	0.08	0.10
CELLO2.5	0	21	0	1205	0.98	0	0	0	0
SubCons Realibity	6	15	0	1205	**0.99**	**1**	0.29	0.45	0.53
SubCons RF	7	14	0	1205	**0.99**	**1**	0.33	0.50	0.57
CanSLPred	12	9	8	1197	**0.99**	0.60	**0.57**	**0.59**	**0.58**

Table 5.20: Performance results of the methods for the proteins in Golden-Trust dataset of GLG.

Methods/Metrics	TP	FN	FP	TN	Accuracy	Precision	Recall	F1-score	MCC
MultiLoc2	1	12	0	13	0.54	**1**	0.08	0.15	0.20
LocTree2	0	13	0	13	0.50	0	0	0	0
CELLO2.5	0	13	0	13	0.50	0	0	0	0
SubCons Realibity	0	13	0	13	0.50	0	0	0	0
SubCons RF	1	12	0	13	0.54	**1**	0.08	0.15	0.20
CanSLPred	8	5	0	13	**0.81**	**1**	**0.62**	**0.76**	**0.67**

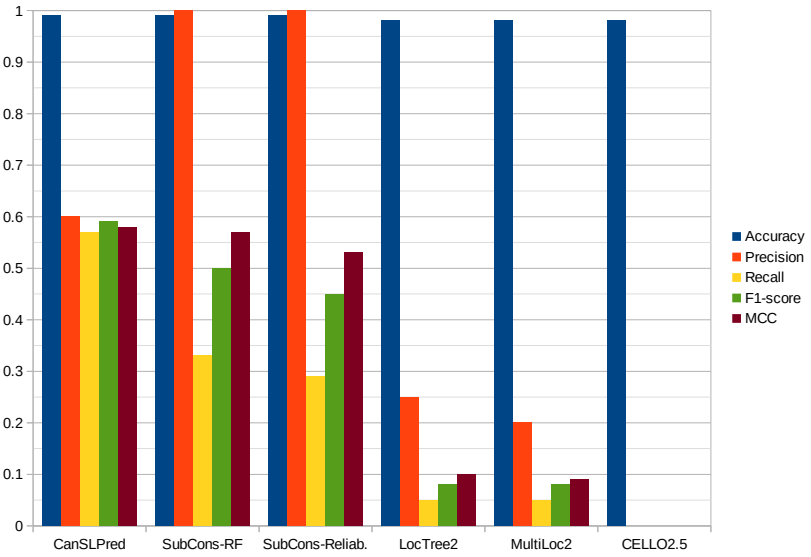

Figure 5.19: Performance results of the methods for the proteins in Golden dataset of GLG.

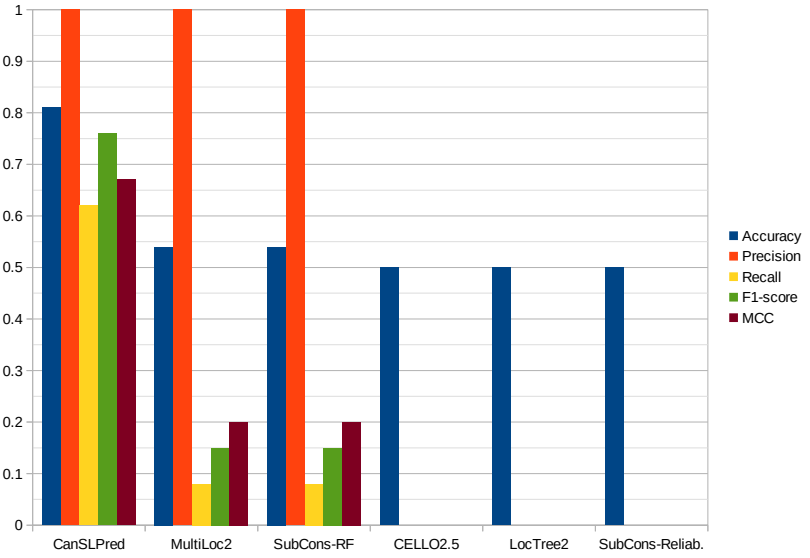

Figure 5.20: Performance results of the methods for the proteins in Golden-Trust dataset of GLG.

Table 5.21: Performance results of the methods for the proteins in Trust-Test dataset of LYS.

Methods/Metrics	TP	FN	FP	TN	Accuracy	Precision	Recall	F1-score	MCC
MultiLoc2	11	42	2	51	0.58	0.85	0.21	0.34	0.26
LocTree2	0	53	0	53	0.50	0	0	0	0
CELLO2.5	12	41	1	52	0.60	0.92	0.23	0.37	0.32
SubCons Realibity	6	47	0	53	0.56	**1**	0.11	0.20	0.24
SubCons RF	7	46	0	53	0.57	**1**	0.13	0.23	0.27
DeepLoc	3	50	0	53	0.53	**1**	0.06	0.11	0.17
CanSLPred	29	24	1	52	**0.76**	0.97	**0.55**	**0.70**	**0.59**

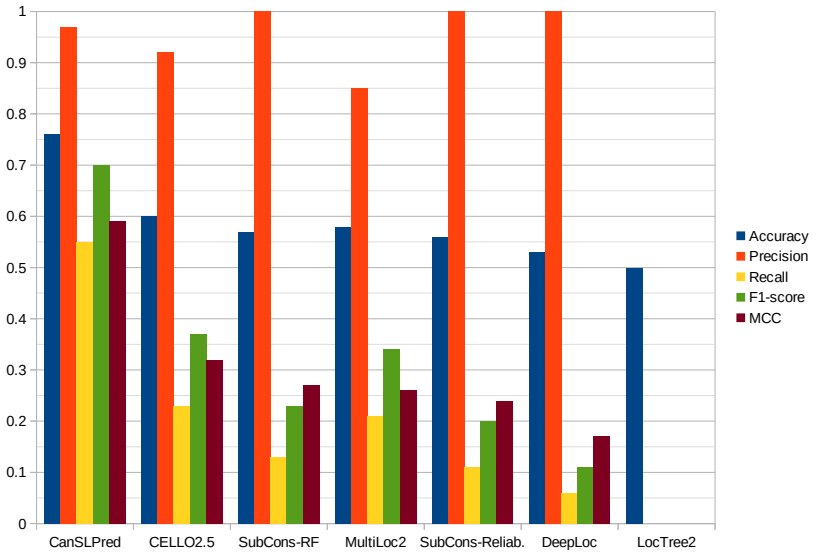

Figure 5.21: Performance results of the methods for the proteins in Trust-Test dataset of LYS.

Table 5.22: Performance results of the methods for the proteins in Golden dataset of LYS.

Methods/Metrics	TP	FN	FP	TN	Accuracy	Precision	Recall	F1-score	MCC
MultiLoc2	6	5	2	1213	**0.99**	0.75	0.55	0.63	0.64
LocTree2	0	11	0	1215	**0.99**	0	0	0	0
CELLO2.5	5	6	1	1214	**0.99**	0.83	0.45	0.58	0.61
SubCons Realibity	4	7	0	1215	**0.99**	**1**	0.36	0.53	0.60
SubCons RF	4	7	0	1215	**0.99**	**1**	0.36	0.53	0.60
CanSLPred	8	13	5	1210	**0.99**	0.62	**0.73**	**0.67**	**0.67**

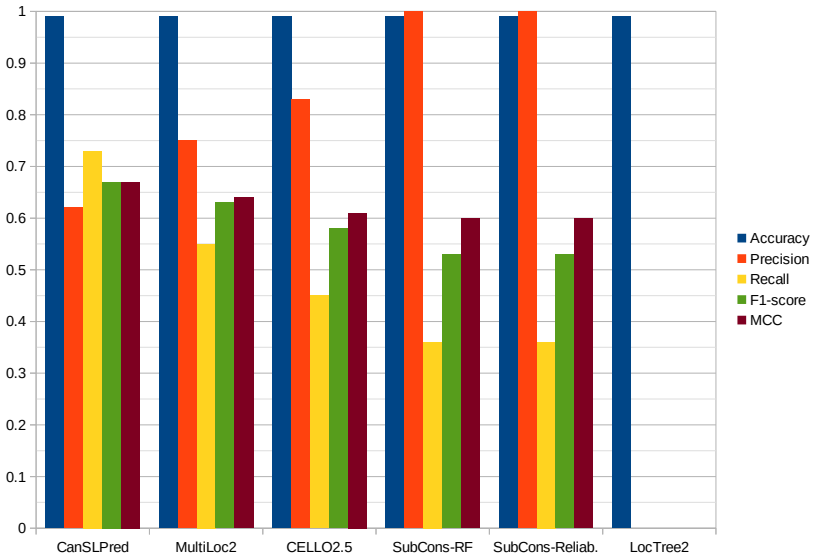

Figure 5.22: Performance results of the methods for the proteins in Golden dataset of LYS.

Table 5.23: Performance results of the methods for the proteins in Golden-Trust dataset of LYS.

Methods/Metrics	TP	FN	FP	TN	Accuracy	Precision	Recall	F1-score	MCC
MultiLoc2	6	4	0	10	0.80	**1**	0.60	0.75	0.65
LocTree2	0	10	0	10	0.50	0	0	0	0
CELLO2.5	5	5	0	10	0.75	**1**	0.50	0.67	0.58
SubCons Realibity	4	6	0	10	0.70	**1**	0.40	0.57	0.50
SubCons RF	4	6	0	10	0.70	**1**	0.40	0.57	0.50
CanSLPred	7	3	0	10	**0.85**	**1**	**0.70**	**0.82**	**0.73**

Table 5.24: Performance results of the methods for the proteins in Trust-Test dataset of PEX.

Methods/Metrics	TP	FN	FP	TN	Accuracy	Precision	Recall	F1-score	MCC
MultiLoc2	8	8	0	16	0.75	**1**	0.50	0.67	0.58
LocTree2	3	13	0	16	0.59	**1**	0.19	0.32	0.32
CELLO2.5	2	14	0	16	0.56	**1**	0.12	0.21	0.26
SubCons Realibity	9	7	0	16	0.78	**1**	0.56	0.72	0.63
SubCons RF	11	5	0	16	**0.84**	**1**	**0.69**	**0.82**	**0.72**
DeepLoc	2	14	0	16	0.56	**1**	0.12	0.21	0.26
CanSLPred	10	6	0	16	0.81	**1**	0.62	0.77	0.67

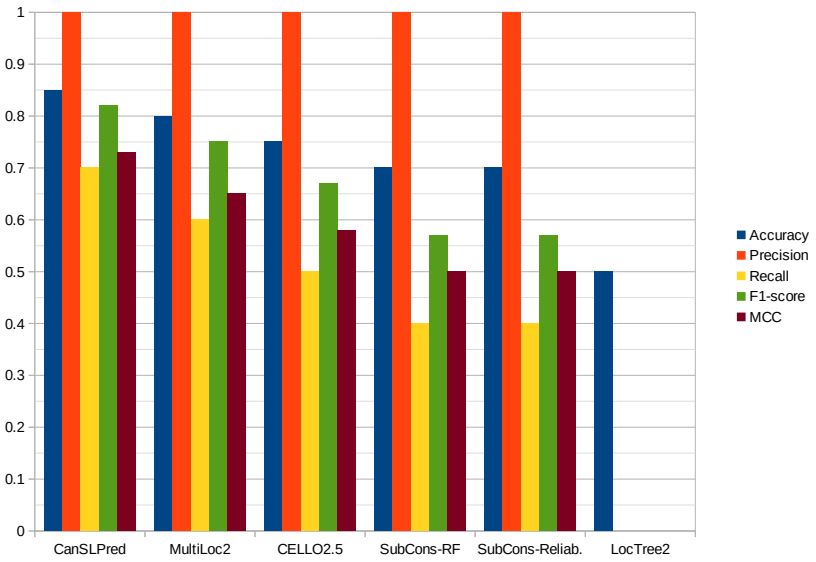

Figure 5.23: Performance results of the methods for the proteins in Golden-Trust dataset of LYS.

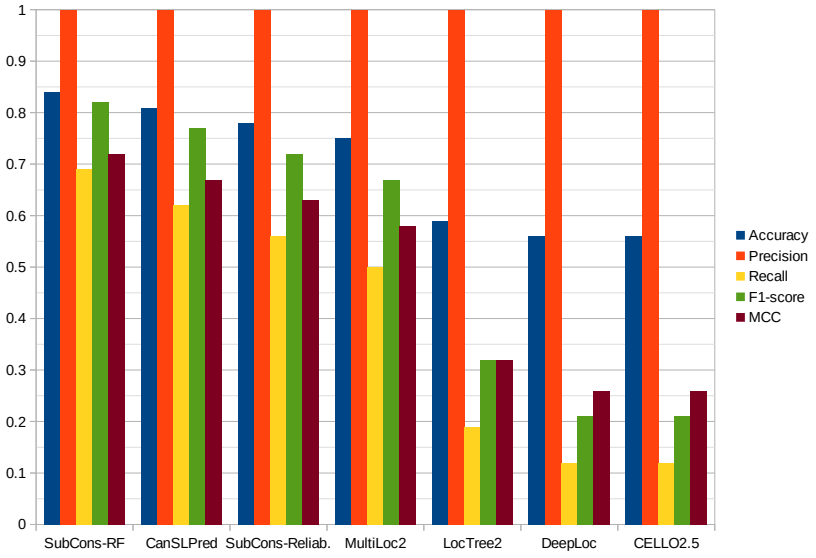

Figure 5.24: Performance results of the methods for the proteins in Trust-Test dataset of PEX.

Table 5.25: Performance results of the methods for the proteins in Golden dataset of PEX.

Methods/Metrics	TP	FN	FP	TN	Accuracy	Precision	Recall	F1-score	MCC
MultiLoc2	3	4	32	1187	0.97	0.09	**0.43**	0.15	0.18
LocTree2	0	7	1	1217	0.99	0	0	0	0
CELLO2.5	1	6	0	1218	**1**	**1**	0.14	0.25	0.38
SubCons Realibity	2	5	12	1204	0.99	0.14	0.29	0.19	0.20
SubCons RF	2	5	19	1199	0.98	0.10	0.29	0.15	0.16
CanSLPred	3	4	4	1215	0.99	0.43	**0.43**	**0.43**	**0.43**

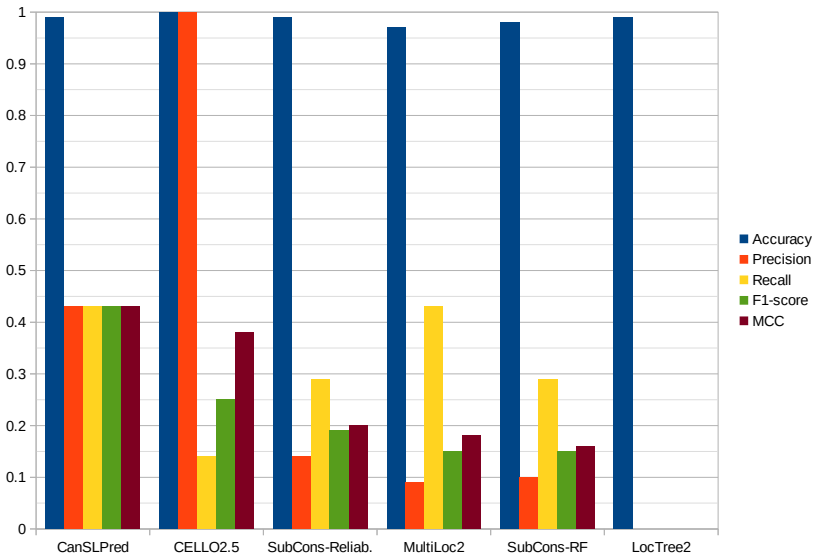

Figure 5.25: Performance results of the methods for the proteins in Golden dataset of PEX.

Table 5.26: Performance results of the methods for the proteins in Golden-Trust dataset of PEX.

Methods/Metrics	TP	FN	FP	TN	Accuracy	Precision	Recall	F1-score	MCC
MultiLoc2	3	4	0	7	**0.71**	**1**	**0.43**	**0.60**	**0.52**
LocTree2	0	7	0	7	0.50	0	0	0	0
CELLO2.5	1	6	0	7	0.57	**1**	0.14	0.25	0.38
SubCons Realibity	2	5	0	7	0.64	**1**	0.29	0.45	0.41
SubCons RF	2	5	0	7	0.64	**1**	0.29	0.45	0.41
CanSLPred	3	4	0	7	**0.71**	**1**	**0.43**	**0.60**	**0.52**

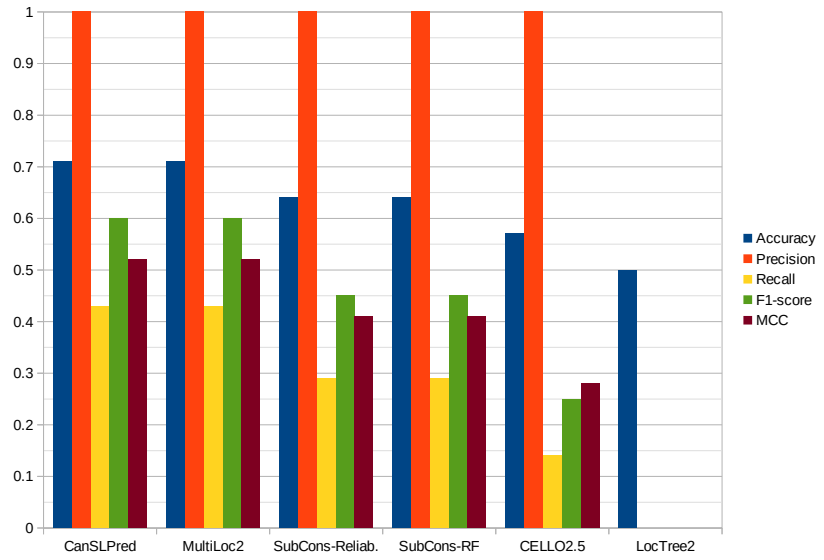

Figure 5.26: Performance results of the methods for the proteins in Golden-Trust dataset of PEX.

Table 5.27: Comparison of the predictors in terms of MCC scores for all subcellular locations by using Trust-Test dataset.

SLs/Predictors	CELLO2.5	MultiLoc2	LocTree2	SubCons-Realiab.	SubCons-RF	DeepLoc	CanSLPred
NUC	0.38	0.24	0.42	0.49	0.35	0.46	**0.55**
CYT	0.27	0.41	0.34	0.29	0.38	0.44	**0.45**
MEM	0.41	0.23	0.33	0.55	0.55	0.64	**0.67**
EXC	0.59	0.37	0.62	0.65	0.66	0.74	**0.86**
MIT	0.46	0.40	0.44	0.46	0.46	0.60	**0.62**
ERE	0.15	0.26	0.37	0.27	0.24	0.40	**0.57**
GLG	0.20	0.12	0.11	0.22	0.2	0.23	**0.36**
LYS	0.32	0.26	0	0.24	0.27	0.17	**0.59**
PEX	0.26	0.58	0.32	0.63	**0.72**	0.26	0.67
Overall	0.34	0.32	0.33	0.42	0.43	0.44	**0.59**

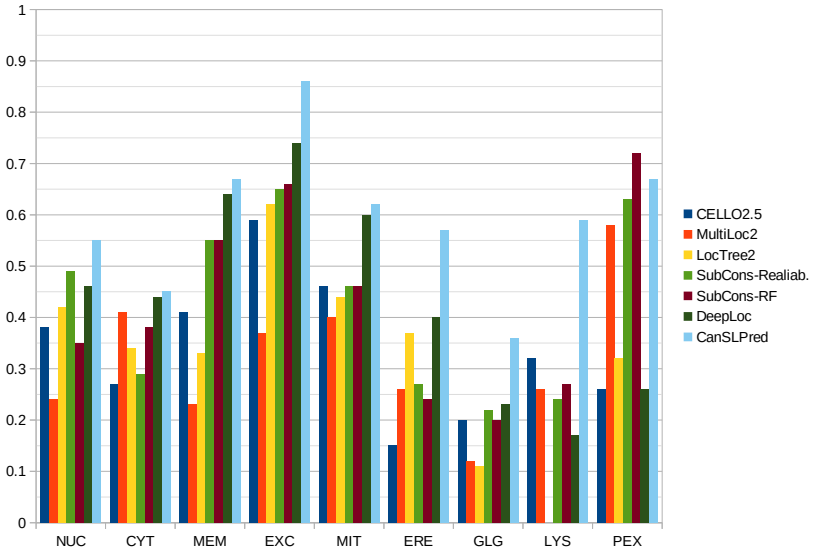

Figure 5.27: Comparison of the predictors in terms of MCC scores for all subcellular locations by using Trust-Test dataset.

Table 5.28: Comparison of the predictors in terms of MCC scores for all subcellular locations by using Golden dataset.

SLs/Predictors	CELLO2.5	MultiLoc2	LocTree2	SubCons-Realiab.	SubCons-RF	CanSLPred
NUC	0.59	0.39	0.67	0.63	0.68	**0.69**
CYT	0.3	0.25	0.32	0.17	0.36	**0.39**
MEM	0.43	0.36	0.4	0.62	0.66	**0.76**
EXC						
MIT	0.69	0.71	0.65	**0.80**	**0.80**	0.79
ERE	0.29	0.54	0.48	0.59	**0.64**	0.54
GLG	0	0.09	0.10	0.53	0.57	**0.58**
LYS	0.61	0.64	0	0.6	0.60	**0.67**
PEX	0.38	0.18	0	0.20	0.16	**0.43**
Overall	0.41	0.40	0.33	0.52	0.56	**0.61**

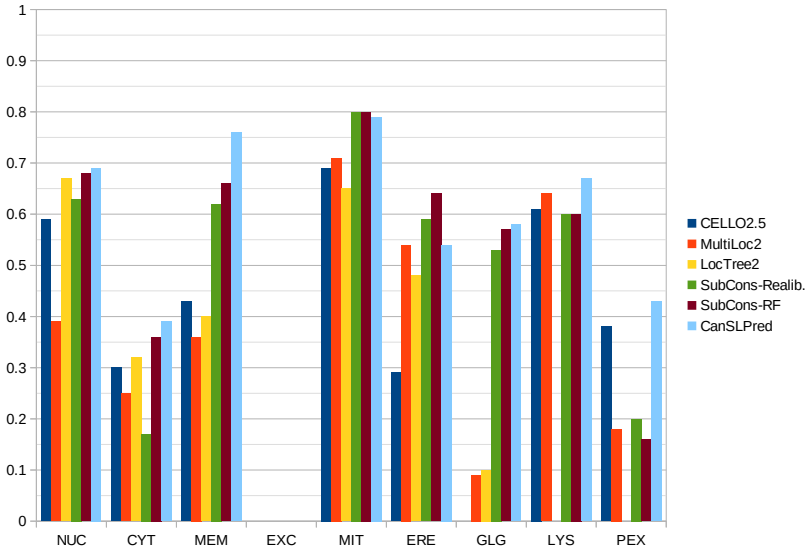

Figure 5.28: Comparison of the predictors in terms of MCC scores for all subcellular locations by using Golden dataset.

Table 5.29: Comparison of the predictors in terms of MCC scores for all subcellular locations by using Golden-Trust dataset.

SLs/Predictors	CELLO2.5	MultiLoc2	LocTree2	SubCons-Realiab.	SubCons-RF	CanSLPred
NUC	0.58	0.42	0.66	0.63	**0.72**	0.69
CYT	0.33	0.34	0.56	0.16	0.48	**0.70**
MEM	0.42	0.27	0.46	0.60	0.60	**0.65**
EXC						
MIT	0.58	0.71	0.59	0.80	0.80	**0.84**
ERE	0.25	0.5	0.40	0.47	0.50	**0.60**
GLG	0	0.20	0	0	0.20	**0.67**
LYS	0.58	0.65	0	0.50	0.50	**0.73**
PEX	0.28	**0.52**	0	0.41	0.41	**0.52**
Overall	0.38	0.45	0.33	0.45	0.53	**0.68**

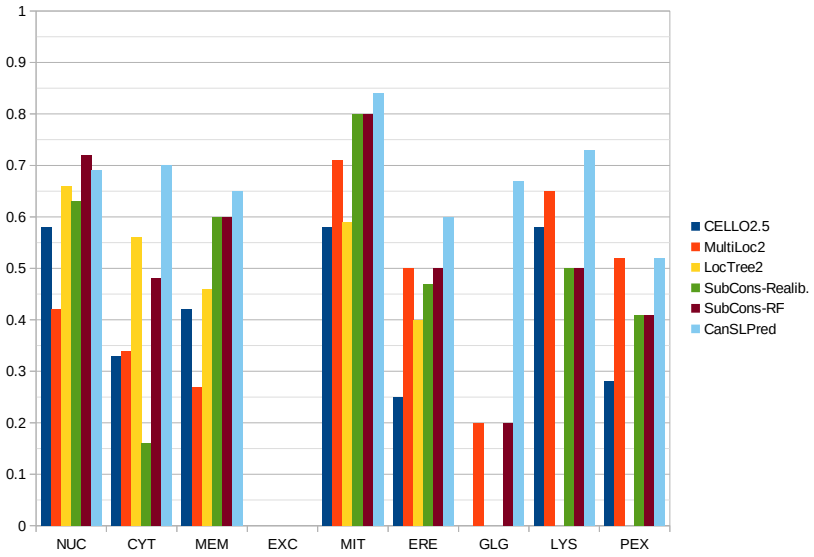

Figure 5.29: Comparison of the predictors in terms of MCC scores for all subcellular locations by using Golden-Trust dataset.

CHAPTER 6

CONCLUSION, DISCUSSION AND FUTURE WORK

6.1 Conclusion and Discussion

Determining the subcellular localization of proteins is crucial for understanding the functions of proteins, drug targeting, systems biology, and proteomics research. The subcellular localization of proteins can experimentally be identified by purification or imaging methods which are expensive and time-consuming. Therefore, several computational methods for automated prediction of protein subcellular localization are proposed in the last two decades; however, there is still room for better performance. Here, we introduce a multi-view classification method (CanSLPred) that provides subcellular localization predictions for human proteins. In the proposed multi-view approach, we employ seven feature-based probabilistic prediction models that provide seven distinct representations (protein descriptors) and seven probabilistic predictions for each protein sequence. There are three major parts that we describe in this study:

1. A newly generated subcellular location hierarchy is introduced by integrating Universal Protein Knowledge Base (UniProtKB) Subcellular Location (SL) terms to Gene Ontology (GO) Cellular Component (CC) hierarchy.
2. A dataset of protein sequences is generated by taking the proteins whose subcellular localization is experimentally annotated in UniProtKB/SwissProt and applying the new SL hierarchy to propagate the proteins according to their subcellular localization. This dataset is called Trust dataset.
3. A new classification method is described to predict the subcellular local-

ization of human proteins by employing a weighted mean voting multi-view Support Vector Machine (SVM) approach.

The new subcellular location hierarchy is formed to unite the characteristics of both hierarchies: UniProtKB SL hierarchy and GO CC hierarchy. To generate the new SL hierarchy, UniProtKB SL identifiers are first mapped to GO CC terms. GO CC hierarchy is then extracted by considering 'is_a' relations among GO CC terms in GO hierarchy. The mapping of UniProtKB SL identifiers to GO CC terms is applied at the end. CanSLPred consists of nine independently constructed classification models where each model provides predictions for one of nine subcellular locations: cytoplasm (CYT), nucleus (NUC), cell membrane (MEM), mitochondrion (MIT), endoplasmic reticulum (ERE), secreted (EXC), Golgi apparatus (GLG), lysosome (LYS), and peroxisome (PEX). The classification models are developed by considering the subcellular localization problem as a binary classification problem where each of the models decides if a protein localizes to the corresponding subcellular location or not. Each classification model predicts the subcellular localization of proteins by following four steps:

1. Feature extraction and normalization
2. Prediction by probabilistic models
3. Weighted-mean voting
4. Thresholding

In the feature extraction process, seven protein descriptors are selected out of 160 cases (40 descriptors from three tools: iFeature [4], POSSUM [5], and SPMAP [6] and 4 normalization methods: Standard normalization, MinMax normalization, Robust scaler, Power transformation), where these seven protein descriptors contribute the best in the combination of probabilistic prediction models. SVM is used to construct probabilistic prediction models, which produces probabilistic scores indicating the localization probability for a query protein sequence. A weighted score is calculated based on the obtained probabilistic scores from seven feature-based probabilistic prediction models

(SVMs) by employing weighted mean voting. Binary prediction is given by applying thresholding on the weighted score.

We evaluate CanSLPred by using three datasets: Trust-Test dataset (our in-house dataset), Golden dataset (SubCons' benchmark dataset), and Golden-Trust dataset(a refined version of Golden dataset). Trust dataset is created by applying the new SL hierarchy on the proteins whose subcellular localization is experimentally annotated in UniProtKB/SwissProt. Golden dataset consists of protein sequences whose subcellular localization is experimentally annotated in at least two out of three protein resources: mass spectrometry (Mass-Spec), SLHPA, and UniProtKB. Golden-Trust dataset is a refined version of Golden dataset where the steps we follow to generate Trust dataset are applied for the protein sequences in Golden dataset. We compare the results of CanSLPred with five state-of-the-art methods: MultiLoc2 [29], LocTree2 [31], CELLO2.5 [30], SubCons [7], and DeepLoc [32]. Although CanSLPred draws back in the classification of the proteins for some locations (PEX in Trust-Test dataset, MIT and ERE in Golden dataset, NUC in Golden-Trust dataset), it achieves the highest overall MCC scores on three test datasets, which indicates that CanSLPred shows remarkable achievement in subcellular localization prediction of human proteins. CanSLPred's overall Matthews correlation coefficient (MCC) scores are 59%, 68%, 61% overall Matthews correlation coefficient (MCC) scores on Trust-Test dataset, Golden-Trust dataset, Golden dataset, respectively whereas SubCons' overall MCC scores are 43%, 53%, and 56% (as illustrated in Chapter 5).

The achievement of CanSLPred is based on the following ideas that we apply in the construction process :

1. A carefully prepared dataset, Trust dataset, is employed in the training process.
2. It is crucial to use the most representative protein descriptors in protein feature extraction. We search for the most representative combination of seven descriptors out of 160 numerical representations, which renders a multi-view representation for the protein sequences.

3. It is essential to adopt the most appropriate machine learning algorithm for the classification of the proteins. Therefore before deciding on using SVM, we also try the other machine learning algorithms such as Logistic Regression, Naive Bayes, Neural Networks, and Tree-based algorithms.

6.2 Future work

As future work, we would like to serve CanSLPred as an online and standalone prediction tool for human proteins. We also plan to extend our approach and strengthen the reliability of the predictor by applying the double-threshold in the thresholding step, which we will leave a gray area (neither a positive prediction nor a negative prediction) and by integrating the information of protein-protein interactions (PPI) as a post-processing step. Moreover, we want to evaluate CanSLPred on all human proteins according to UniProtKB annotations.

REFERENCES

[1] B. Nerlich, "Talking organelles: A riot of metaphors," 2019.

[2] "Uniprot: the universal protein knowledgebase," *Nucleic acids research*, vol. 45, no. D1, pp. D158–D169, 2016.

[3] G. O. Consortium, "The gene ontology (go) database and informatics resource," *Nucleic acids research*, vol. 32, no. suppl_1, pp. D258–D261, 2004.

[4] Z. Chen, P. Zhao, F. Li, A. Leier, T. T. Marquez-Lago, Y. Wang, G. I. Webb, A. I. Smith, R. J. Daly, K.-C. Chou, *et al.*, "ifeature: a python package and web server for features extraction and selection from protein and peptide sequences," *Bioinformatics*, vol. 34, no. 14, pp. 2499–2502, 2018.

[5] J. Wang, B. Yang, J. Revote, A. Leier, T. T. Marquez-Lago, G. Webb, J. Song, K.-C. Chou, and T. Lithgow, "Possum: a bioinformatics toolkit for generating numerical sequence feature descriptors based on pssm profiles," *Bioinformatics*, vol. 33, no. 17, pp. 2756–2758, 2017.

[6] O. S. Sarac, Ö. Gürsoy-Yüzügüllü, R. Cetin-Atalay, and V. Atalay, "Subsequence-based feature map for protein function classification," *Computational biology and chemistry*, vol. 32, no. 2, pp. 122–130, 2008.

[7] M. Salvatore, P. Warholm, N. Shu, W. Basile, and A. Elofsson, "Subcons: a new ensemble method for improved human subcellular localization predictions," *Bioinformatics*, vol. 33, no. 16, pp. 2464–2470, 2017.

[8] EBI, "EMBL-EBI resources protein classification," 2019.

[9] K. Nakai and M. Kanehisa, "Expert system for predicting protein localization sites in gram-negative bacteria," *Proteins: Structure, Function, and Bioinformatics*, vol. 11, no. 2, pp. 95–110, 1991.

[10] P. Horton, K.-J. Park, T. Obayashi, N. Fujita, H. Harada, C. Adams-

Collier, and K. Nakai, "Wolf psort: protein localization predictor," *Nucleic acids research*, vol. 35, no. suppl_2, pp. W585–W587, 2007.

[11] O. Emanuelsson, H. Nielsen, S. Brunak, and G. Von Heijne, "Predicting subcellular localization of proteins based on their n-terminal amino acid sequence," *Journal of molecular biology*, vol. 300, no. 4, pp. 1005–1016, 2000.

[12] H. Nielsen, J. Engelbrecht, S. Brunak, and G. V. Heijne, "A neural network method for identification of prokaryotic and eukaryotic signal peptides and prediction of their cleavage sites," *International journal of neural systems*, vol. 8, no. 05n06, pp. 581–599, 1997.

[13] H. Nakashima and K. Nishikawa, "Discrimination of intracellular and extracellular proteins using amino acid composition and residue-pair frequencies," *Journal of molecular biology*, vol. 238, no. 1, pp. 54–61, 1994.

[14] 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.

[15] K.-C. Chou and Y.-D. Cai, "Prediction and classification of protein subcellular location—sequence-order effect and pseudo amino acid composition," *Journal of cellular biochemistry*, vol. 90, no. 6, pp. 1250–1260, 2003.

[16] Z. Lu, D. Szafron, R. Greiner, P. Lu, D. S. Wishart, B. Poulin, J. Anvik, C. Macdonell, and R. Eisner, "Predicting subcellular localization of proteins using machine-learned classifiers," *Bioinformatics*, vol. 20, no. 4, pp. 547–556, 2004.

[17] M.-W. Mak, J. Guo, and S.-Y. Kung, "Pairprosvm: protein subcellular localization based on local pairwise profile alignment and svm," *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol. 5, no. 3, pp. 416–422, 2008.

[18] M. S. Scott, D. Y. Thomas, and M. T. Hallett, "Predicting subcellular localization via protein motif co-occurrence," *Genome research*, vol. 14, no. 10a, pp. 1957–1966, 2004.

[19] E. Quevillon, V. Silventoinen, S. Pillai, N. Harte, N. Mulder, R. Apweiler, and R. Lopez, "Interproscan: protein domains identifier," *Nucleic acids research*, vol. 33, no. suppl_2, pp. W116–W120, 2005.

[20] J. He, H. Gu, and W. Liu, "Imbalanced multi-modal multi-label learning for subcellular localization prediction of human proteins with both single and multiple sites," *PloS one*, vol. 7, no. 6, p. e37155, 2012.

[21] K.-C. Chou and H.-B. Shen, "Hum-ploc: a novel ensemble classifier for predicting human protein subcellular localization," *Biochemical and biophysical research communications*, vol. 347, no. 1, pp. 150–157, 2006.

[22] K.-C. Chou and H.-B. Shen, "Euk-mploc: a fusion classifier for large-scale eukaryotic protein subcellular location prediction by incorporating multiple sites," *Journal of Proteome Research*, vol. 6, no. 5, pp. 1728–1734, 2007.

[23] K.-C. Chou and H.-B. Shen, "Large-scale predictions of gram-negative bacterial protein subcellular locations," *Journal of proteome research*, vol. 5, no. 12, pp. 3420–3428, 2006.

[24] S. F. Altschul, T. L. Madden, A. A. Schäffer, J. Zhang, Z. Zhang, W. Miller, and D. J. Lipman, "Gapped blast and psi-blast: a new generation of protein database search programs," *Nucleic acids research*, vol. 25, no. 17, pp. 3389–3402, 1997.

[25] W.-L. Huang, C.-W. Tung, S.-W. Ho, S.-F. Hwang, and S.-Y. Ho, "Prolocgo: utilizing informative gene ontology terms for sequence-based prediction of protein subcellular localization," *BMC bioinformatics*, vol. 9, no. 1, p. 80, 2008.

[26] S. Wan, M.-W. Mak, and S.-Y. Kung, "mgoasvm: Multi-label protein subcellular localization based on gene ontology and support vector machines," *BMC bioinformatics*, vol. 13, no. 1, p. 290, 2012.

[27] C.-S. Yu, C.-W. Cheng, W.-C. Su, K.-C. Chang, S.-W. Huang, J.-K. Hwang, and C.-H. Lu, "Cello2go: a web server for protein subcellular localization prediction with functional gene ontology annotation," *PloS one*, vol. 9, no. 6, p. e99368, 2014.

[28] S. Briesemeister, T. Blum, S. Brady, Y. Lam, O. Kohlbacher, and H. Shatkay, "Sherloc2: a high-accuracy hybrid method for predicting subcellular localization of proteins," *Journal of proteome research*, vol. 8, no. 11, pp. 5363–5366, 2009.

[29] T. Blum, S. Briesemeister, and O. Kohlbacher, "Multiloc2: integrating phylogeny and gene ontology terms improves subcellular protein localization prediction," *BMC bioinformatics*, vol. 10, no. 1, p. 274, 2009.

[30] C.-S. Yu, Y.-C. Chen, C.-H. Lu, and J.-K. Hwang, "Prediction of protein subcellular localization," *Proteins: Structure, Function, and Bioinformatics*, vol. 64, no. 3, pp. 643–651, 2006.

[31] T. Goldberg, T. Hamp, and B. Rost, "Loctree2 predicts localization for all domains of life," *Bioinformatics*, vol. 28, no. 18, pp. i458–i465, 2012.

[32] J. J. Almagro Armenteros, C. K. Sønderby, S. K. Sønderby, H. Nielsen, and O. Winther, "Deeploc: prediction of protein subcellular localization using deep learning," *Bioinformatics*, vol. 33, no. 21, pp. 3387–3395, 2017.

[33] G. O. Consortium, "The gene ontology (go) project in 2006," *Nucleic acids research*, vol. 34, no. suppl_1, pp. D322–D326, 2006.

[34] B. E. Suzek, Y. Wang, H. Huang, P. B. McGarvey, C. H. Wu, and U. Consortium, "Uniref clusters: a comprehensive and scalable alternative for improving sequence similarity searches," *Bioinformatics*, vol. 31, no. 6, pp. 926–932, 2014.

[35] L. Gatto, L. M. Breckels, S. Wieczorek, T. Burger, and K. S. Lilley, "Mass-spectrometry-based spatial proteomics data analysis using proloc and prolocdata," *Bioinformatics*, vol. 30, no. 9, pp. 1322–1324, 2014.

[36] L. Fagerberg, C. Stadler, M. Skogs, M. Hjelmare, K. Jonasson, M. Wiking, A. Åbergh, M. Uhlén, and E. Lundberg, "Mapping the subcellular protein distribution in three human cell lines," *Journal of proteome research*, vol. 10, no. 8, pp. 3766–3777, 2011.

[37] M. Uhlen, P. Oksvold, L. Fagerberg, E. Lundberg, K. Jonasson, M. Forsberg, M. Zwahlen, C. Kampf, K. Wester, S. Hober, *et al.*, "Towards a

knowledge-based human protein atlas," *Nature biotechnology*, vol. 28, no. 12, p. 1248, 2010.

[38] A. S. Rifaioglu, H. Atas, M. J. Martin, R. Cetin-Atalay, V. Atalay, and T. Dogan, "Recent applications of deep learning and machine intelligence on in silico drug discovery: methods, tools and databases," *Brief. Bioinform*, vol. 10, 2018.

[39] G. J. van Westen, R. F. Swier, J. K. Wegner, A. P. IJzerman, H. W. van Vlijmen, and A. Bender, "Benchmarking of protein descriptor sets in proteochemometric modeling (part 1): comparative study of 13 amino acid descriptor sets," *Journal of cheminformatics*, vol. 5, no. 1, p. 41, 2013.

[40] J. Dong, Z.-J. Yao, L. Zhang, F. Luo, Q. Lin, A.-P. Lu, A. F. Chen, and D.-S. Cao, "Pybiomed: a python library for various molecular representations of chemicals, proteins and dnas and their interactions," *Journal of cheminformatics*, vol. 10, no. 1, p. 16, 2018.

[41] D. T. Jones, "Protein secondary structure prediction based on position-specific scoring matrices," *Journal of molecular biology*, vol. 292, no. 2, pp. 195–202, 1999.

[42] S. Henikoff and J. G. Henikoff, "Amino acid substitution matrices from protein blocks," *Proceedings of the National Academy of Sciences*, vol. 89, no. 22, pp. 10915–10919, 1992.

[43] Z. Xie, J. Hall, I. P. McCarthy, M. Skitmore, and L. Shen, "Standardization efforts: The relationship between knowledge dimensions, search processes and innovation outcomes," *Technovation*, vol. 48, pp. 69–78, 2016.

[44] G. E. Box and D. R. Cox, "An analysis of transformations," *Journal of the Royal Statistical Society: Series B (Methodological)*, vol. 26, no. 2, pp. 211–243, 1964.

[45] I.-K. Yeo and R. A. Johnson, "A new family of power transformations to improve normality or symmetry," *Biometrika*, vol. 87, no. 4, pp. 954–959, 2000.

[46] R. Lennart and B. Westergren, "Beta mathematics handbook: concepts, theorems, methods, algorithms, formulas, graphs, tables," 1997.

[47] P. J. Rousseeuw and C. Croux, “Alternatives to the median absolute deviation,” *Journal of the American Statistical association*, vol. 88, no. 424, pp. 1273–1283, 1993.

[48] Y. Jiao and P. Du, “Performance measures in evaluating machine learning based bioinformatics predictors for classifications,” *Quantitative Biology*, vol. 4, no. 4, pp. 320–330, 2016.