Prediction of Protein methylation sites of lysine residues using machine learning algorithms

Sadia Islam

Department of Computer Science and Engineering

United International University

A thesis submitted for the degree of

MSc in Computer Science & Engineering

December 2022

Declaration

I, Sadia Islam, declare that this thesis titled, Prediction of Protein methylation sites of lysine residues using machine learning algorithms and the work presented in it are my own. I confirm that:

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

Signed: _____

Date: _____

(Sadia Islam)

Certificate

I do hereby declare that the research works embodied in this thesis entitled Prediction of Protein methylation sites of lysine residues using machine learning algorithms is the outcome of an original work carried out by Sadia Islam under my supervision.

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

Signed: _____

Date: _____

(Swakkhar Shatabda)

Department of Computer Science and Engineering,

United International University,

Dhaka-1209, Bangladesh.

Abstract

Post Translational Modification (PTM) plays an essential role in the biological and molecular mechanisms. They are also considered as a vital element in cell signaling and networking pathways. Among different PTMs, Methylation is regarded as one of the essential types. Methylation plays a crucial role in maintaining the dynamic balance, stability, and remodeling of chromatins. Methylation also leads to different abnormalities in cells and is responsible for many serious diseases. Methylation can be detected by experimental approaches such as methylation-specific antibodies, mass spectrometry, or characterizing methylation sites using the radioactive labeling method. However, these practical approaches are time-consuming and costly. Therefore, there is a demand for fast and accurate computational techniques to focus on these issues. This study proposes a machine learning based approach called LyMethySE to predict methylation sites in proteins. To build this model, we use an evolutionary-based bi-gram profile combined with predicted structural approach to extract features. To our best knowledge, no method has been used to predict the methylation site of lysine residues using combination information as feature extraction technique. Incorporating profile bigram also leads LyMethySE to keep the features size limited for different evolutionary information window size. We apply mostly used eight different classifiers from literature as predictor to evaluate our feature extraction technique. Among them, Support Vector Machine (SVM) outperforms the result. Therefore, we use SVM as our base classification technique to build LyMethySE. This study also shows the impact and comparative analysis of different base classifiers for our extracted features.

Published Papers

Work relating to the research presented in this thesis has been published by the author in the following peer-reviewed journal named Neural Computing and Applications (NCAA) [Impact Factor: 5+]

1. Sadia Islam, Shafayat Bin Shabbir Mugdha, Shubhashis Roy Dipta, MD. Easin Arafat, Swakkhar Shatabda, Hamid Alinejad Rokny and Iman Dehzangi , MethEvo: An Accurate Evolutionary Information-based Methylation site predictor, Neural Computing and Applications (NCAA) , Vol. 34, Issue. 17, August 2022, link: `https://trebuchet.public.springernature.app/get_content/b1556961-3a82-48c0-9600-00684ebbaaeb`

Acknowledgements

First of all I would like to thank to my Almighty for bringing me this far and making my paths smooth enough to complete the journey.

I would also like to express my gratitude to the people who constantly provided support in throughout the journey. Without their support it was not possible for me to complete the work.

I would like to thank my respected supervisor Dr. Swakkhar Shatabda, Associate Professor, United International University for providing his continuous guideline in lead to complete the work. I would also like to express my greatest gratitude to Dr. Iman Dehzangi, Department of Computer Science, Rutgers University, USA for enlightening my paths with his knowledge to find out the way of solutions in this research and providing me his valuable time for continuous mentoring.

I would like to provide my sincere gratitude to my family especially my parents and my husband for standing beside me continuously throughout the research journey and providing me full emotional support to complete the task.

Last but the not the least, I would like to show gratitude to United International University for providing me such an environment to conduct the study.

Contents

List of Figures **ix**

List of Tables **x**

1 Introduction **1**

1.1 Motivation . . . 1

1.2 Research Aims . . . 2

1.3 Objectives of the Thesis . . . 2

1.4 Thesis Contributions . . . 3

1.5 Organization of the Thesis . . . 3

2 Background **4**

2.1 Preliminaries . . . 4

2.1.1 Biological Preliminaries . . . 4

2.1.1.1 Proteins . . . 4

2.1.1.2 Post Translational Modification (PTM) . . . 5

2.1.1.3 Methylation . . . 6

2.1.2 Machine Learning Preliminaries . . . 6

2.1.2.1 Defining Machine Learning . . . 6

2.1.2.2 Machine Learning in PTM . . . 7

2.1.2.3 Machine Learning Algorithms . . . 7

2.2 Literature Review . . . 10

2.3 Gap Analysis . . . 11

2.3.1 Dataset . . . 11

2.3.2 Features . . . 11

2.3.3 Results . . . 13

3 Proposed Method **14**

3.1 Materials Methods . . . 14

3.1.1 Benchmark Dataset . . . 14

3.1.2 Dataset Balancing . . . 15

3.1.3 Train and Test Dataset . . . 17

3.1.4 Features Extraction . . . 17

3.1.4.1 Evolutionary Information Based Features . . . 17

3.1.4.2 Profile Bigram . . . 17

3.1.4.3 Structural Information Based Features . . . 18

3.2 Classifiers . . . 20

3.2.1 Support Vector Machine . . . 21

3.2.2 Other applied classifiers . . . 21

3.3 Performance Evaluation . . . 21

3.3.1 Validation . . . 22

3.4 Summary . . . 22

4 Experimental Analysis **24**

4.1 Independent train and test set . . . 24

4.2 Experimental Results . . . 24

4.3 Summary . . . 33

5 Conclusions, and Future Work **34**

5.1 Conclusions . . . 34

5.2 Future Work . . . 34

Bibliography **35**

List of Figures

3.1 The imbalanced data distribution of positive(1) and negative(0) class 15

3.2 The balanced data distribution of positive(1) and negative(0) class using a) KNN based under sampling method and b) SMOTE over sampling method 16

3.3 Extracting features from the neighboring amino acids using mirror method for amino acids at the two ends of a protein 19

3.4 The working methodology of LyMethySE 23

4.1 (a) MCC score and (b) f1 score of classifiers results for 10 fold cross validation using evolutionary based bigram as feature and combination of evolutionary based bigram with predicted structural information as feature 29

4.2 (a) MCC score and (b) f1 score of classifiers results for independent test set using evolutionary based bigram as feature and combination of evolutionary based bigram with predicted structural information as feature 30

4.3 ROC Curve for TPR vs FPR using (a) 10-fold cross validation and (b) independent test set 30

4.4 Precision-Recall Curve for LyMethySE model and other classifiers 31

List of Tables

2.1 Different amino acids with their reference name 5
2.2 Literature review summary table 12

4.1 Results of different classifiers using evolutionary based profile bigram as features for 10 fold cross validation 25
4.2 Results of different classifiers using evolutionary based profile bigram as features for independent test set 26
4.3 Results of LyMethSE along with other classifiers using 10 fold cross validation . 27
4.4 Results of LyMethSE along with other classifiers using independent test set . . . 27
4.5 Results of different classifiers using combinational features and SMOTE minority class over sampling method 32

Chapter 1

Introduction

This chapter includes a brief introduction on Methylation site identification necessity, problems and recent trends to apply. It gives basic problem knowledge including aims and objectives of this study. It also includes the artifact of our work. Lastly, the organization of this thesis work is also provided.

1.1 Motivation

Post Translational Modification (PTM) is an important process that increases the functional diversity of proteins [1]. PTM also plays an essential role in cell signaling and networking pathways analysis [2]. Understanding the PTM process also helps to uncover the molecular mechanism of the cells [3]. Among different PTMs, protein methylation is one of the most important ones. Methylation of proteins appears in different protein residues such as lysine, arginine, alanine, histidine, asparagine, and proline. The most common type of methylation is called N-methylation, as it occurs on nitrogen atoms. Other kinds of methylation are mainly seen in oxygen atoms of glutamate and aspartate residues or sulfur atoms in cysteine residues. Such methylations are known as O-methylation and S-methylation, respectively. Though the methylation of protein was brought to light more than forty years ago, the research in this field is still a recent trend. Methylation of histone proteins plays a vital role in maintaining the dynamic balance, stability, and remodeling of chromatins, affecting gene expression levels. Several studies in this field have revealed that the methylation of histone proteins can lead to abnormalities that causes many serious diseases such as cancer [4][5]. Non-histone proteins also undergo lysine methylation with a broad impact on different biological processes such as protein function regulation and RNA sequence processing [4][6]. It has also been shown as the

source of various cellular disorders [7].

Analyzing different research done in this field has revealed that the methylation of histone protein points out the abnormality in protein and leads to many serious diseases such as cancer [4]. Therefore, among all of the other protein residues most of the work was done focused on lysine, arginine [3]. The residue lysine can be methylated in mono, di or tri way by histone. The studies in this regard points out that as H1 and H2 histone proteins play a vital role for processing different biological analysis most of the work in this field was done focusing on these two histone protein methylations. In different biological aspects such as gene expression and protein function regulation and in processing of RNA sequence, lysine methylation plays an important part [8]. The broad analysis of protein methylations and their impacts is still evolving. As a result, identifying methylated protein and the methylation sites is an essential step towards this goal.

1.2 Research Aims

Experimental methods such as analyzing antibodies for specific methylation and mass-spectrometry are used to identify methylation sites. However, these experimental methods are very time consuming and costly to implement and execute. Therefore, proposing fast and accurate computational methods to address these problems has attracted tremendous attention [9].

Among the other residues, the lysine methylation is playing a critical role in order to find out the cellular processes of various disorders and a proper systematic and working process to identify this methylation site is still in need to explore more [7].

In this paper, we have focused on different computational methods in order to predict the site of lysine methylation of protein data. We have analyzed different classification algorithms and found out which give the best result in terms of accuracy, sensitivity, specificity and precision. We have also done a comparative analysis of different feature types along with some combinational feature to boost up the existing work that has been done in this field.

1.3 Objectives of the Thesis

The objective of the paper is to establish a better computational method in order to predict mythylation site from protein data. We want to explore the existing system, the features that have been used so far and the results that we are having. We also want to introduce

combinational features in order to achieve better results. We perform competitive analysis of features to identify their impact on mythylation site prediction process.

1.4 Thesis Contributions

The main contribution of this thesis work are,

- We have used the dataset which has shown great impact in post translational modification identification however never used in methylation site prediction. We prepared that dataset for methylation site prediction.
- We have extracted evolutionary information as PSSM and converted it with profile bigram to keep the feature size fixed even for different window size. evolutionary based information as PSSM used as features shown a great result in literature for other PTMs. Therefore, incorporating it with methylation was one of our study target.
- We have used predicted structural information using SPIDER instead of the actual structural information which is sometimes hard to extract.
- We have combined the extracted predicted structural information with evolutionary information which is extracted from PSSM. No other work was done combining this features.
- We have shown competitive analysis of results for combined and single features.
- We have investigated the impact of using different features along with the combinational one in case of methylation site prediction.
- We have also analyzed how different kind of machine learning algorithm perform with our extracted features.

1.5 Organization of the Thesis

The thesis is organised as follows:

Chapter 2 provides the related existing works.

Chapter 3 presents the proposed method.

Chapter 4 discusses the results and experimental analysis.

Chapter 5 presents the conclusions and discusses the future works.

Chapter 2

Background

In this chapter we have briefly showed the preliminaries and analyzed existing methods and their results. We have also performed gap analysis in order to identify our artifact.

2.1 Preliminaries

Here we explain two different preliminaries of our work. They are the biological preliminaries and machine learning preliminaries.

2.1.1 Biological Preliminaries

2.1.1.1 Proteins

Proteins are the complex large molecules called as building blocks of cells. They are responsible to do most of the work of cells including functionality and biomass. Protein refers to a long chain containing thousand of amino acids. This amino acids are responsible for different functionalities including different molecular motors and different signaling processes [10]. Thus it is an important part to focus on. These protein chains are basically contain 20 different amino acids. These amino acids are named as a single letter for different references. The 20 different amoni acids and their refered names are given on table 2.1.

Among these 20 different amino acids, nine are considered as most important residues as they have more impact in cell ordering. Lysine is one of these important residues which can be referred as K.

Amino acid name	Referred name
Alanine	A
Arginine	R
Asparagine	N
Aspartic acid	D
Cysteine	C
Glutamic acid	E
Glutamine	Q
Glycine	G
Histidine	H
Isoleucine	I
Leucine	L
Lysine	K
Methionine	F
Phenylalanine	M
Proline	P
Serine	S
Threonine	T
Tryptophan	W
Tyrosine	Y
Valine	V

Table 2.1: Different amino acids with their reference name

2.1.1.2 Post Translational Modification (PTM)

Post Translational Modification (PTM) refers to the change that takes place in protein residues. This changes occurs after the translation of RNA messenger to amino acids. The change arises from ribosomal biosynthetic assembly lines and both the prokaryotic and eukaryotic cells are responsible for this [11]. This also plays an important role to identify the biological molecular mechanism [3]. This modification of protein can add additional variation to proteoms. PTMs can also play a vital role in changing the way of protein folding, stability, localization to cell compartments, activation/ inactivation etc. It helps to understand the molecular mechanism of cells. There are more than 400 type of PTMs has been identified so far that can affect the cells in different ways [12].

2.1.1.3 Methylation

Among different PTMs, protein methylation is an important one. The methylation of protein data appears in different protein residues such as lysine, arginine, alanine, histidine, asparagine, proline etc. Mainly the change is happening on these residues backbone or in the side chain. This is called N-methylation as this methylation occurs on nitrogen atoms. There are some other kinds of methylation present which is seen in glutamate and aspartate of oxygen atoms, also in cysteine of sulfur atoms. They are known as O-methylation and S-methylation respectively. A vital role is played by the methylation of histone proteins to identify the stability and remodeling of chromatin of proteins and also in gene expression analyzation. Though both histone and non-histone protein has the impact of lysine methylation [6], to maintain the dynamic balance of histone proteins, methylation of lysine residue is mostly responsible. Analyzing different researches done in this field has revealed that the methylation of histone protein points out the abnormality in protein and leads to many serious diseases including cancer [4].

The protein residue, lysine, can be methylated in mono, di or tri way by histone. The studies in this regard points out that as H1 and H2 histone proteins play a vital role for processing different biological analysis most of the work in this field was done focusing on these two histone protein methylations. Along with the lysine methylation, arginine methylation also plays an important role for protein data analysis. It also has the tendency to affect the histone codes and modify the histones. The arginine methylation to impact the histone codes generally happens together with the lysine methylation. Many studies have shown that lysine-specific demethylase 1 and JmjC domain-containing histone demethylase 1 has occurred due to demethylation of histone H3 which is responsible for lysine 4 and lysine 36 respectively [9]. In different biological aspects such as gene expression and protein function regulation and in processing of RNA sequence, lysine methylation plays an important part [8].

2.1.2 Machine Learning Preliminaries

2.1.2.1 Defining Machine Learning

Machine learning refers to a branch of Computer Science and to be a little specific, it is known as a branch of Artificial Intelligence(AI). This is a technique that is implemented to imitate the human nature of learning and making decisions. This method uses some previous data to learn about the nature of work and based on those previous data, it uses some algorithm to make a decision about some unknown instances [13]. Machine learning is been used in

different diverse fields in order to have prediction about any instances. The applied fields include pattern recognition, computational biology, computer vision, entertainment, medical, spacecraft engineering, finance etc. Thus machine learning is known as a "general purpose technology" [14]. The implementation of machine learning is increasing with time due to its huge success in different fields. The nature of Machine learning tasks which learns gradually make itself eligible to make decision about unseen things made it popular enough to take decisions or predict any future instance [15].

2.1.2.2 Machine Learning in PTM

As among the other residues, the lysine methylation is playing a vital role in order to find out the cellular processes of various disorders, therefore, it is a need to find out a process to identify the methylated sites in order to take proper measures to handle different disorders. There are some traditional approaches available to identify methylation site. They are candidate approaches such as methylation-specific antibodies check test, mapping of post-translational modifications by mass, spectrometry, radioactive labeling to characterize methylation on target proteins [16]. However these methods require different machinaries and its time consuming too. A proper systematic and effective process to identify this methylation site is still in need to explore [17].

In this regard, in our work, we focus on predicting the site of lysine methylation of protein data using different machine learning classifiers. This method will be able to take protein data as input and will produce the possibilities of lysine residues getting methylated or not. Therefore, we are using some classic binary classification methods in order to find out methylation sites. This work also tried to find out a novel and effective feature extraction method that will help the classifiers to predict with the maximum output in case of different evaluation criteria. Therefore, we use a novel feature extraction method which combines the evolutionary information and predicted structural information. The structural information is extracted using SPIDER and the evolutionary information is extracted from PSSM and combined with profile bigram. We analyze nine different classification algorithms and found out which give the best result in terms of accuracy, sensitivity, specificity and precision.

2.1.2.3 Machine Learning Algorithms

The algorithms that are mostly used in literature and also used in this study are, Support Vector Machine, Decision Tree, Logistic Regression, Gradient Boosting, Gaussian Naive Bayes,

AdaBoost, Bernoulli Naive Bayes, Multi-layer Perceptron classifier and Rotation Forest. A short description of each classifier is given below,

Support Vector Machine: This is one of the most widely used algorithm in case of PTM prediction. A large use of Support Vector mchine(SVM) as predictor is also noticed in methylation site prediction [5, 7, 18–20]. SVM works by creating a hyperplane between classes and by finding out the data points on which the partition should be made which is known as support vectors [21]. Support vector is getting more popularity as time passes due to its simplicity and the flexibility that it has to adapt different kind of classification problem [22]. This method is capable of handling both classification and regression problems.

Logistic Regression: Logistic Regression is another well known machine learning algorithm used for prediction purpose. This is also capable to work with both classification and regression problem, however, it is mostly popular for classification problems [23]. This model basically works to establish a relationship between a dependent and an independent variable. This model mainly used to find out the effect that is having a categorical outcome for predictor variable [24]. This model is mostly seen to identify and disease or risk factors [24].

Gaussian Naive Bayes: Naive bayes algorithms are mostly popular for its simple probabilistic analysis. When we use the Gaussian distribution to calculate the probability for each case, it is known as Gaussian Naive bayes [25]. This algorithm always calculates the probability in an independent manner which indicates that it doesn't check for any dependency of features. However, as this model is based on probabilistic model which calculates the probability values to take decision, it is more faster in case of result generation timing [26].

Bernoulli Naive Bayes: This is also a bayesian network model based on the basic theory of making decision with probabilistic calculation. However in bernoulli naive bayes, it doesn't consider the dependencies between words and uses the feature of binary words. It also works well with small number of words or data [27]. It works by just keeping track of the presence of any feature rather than focusing on the number of count or weight [28]. Which shows some magnificent change in results for different application field.

Decision Tree: Decision tree is known as a tree based machine learning algorithm. Here in root and each internal node, it makes a decision to choose which sub tree should be chosen next. This decision is made by analyzing the attribute values and calculating the outputs based on it. From a set of instances, a decision tree is made basically using divide and conquer strategy and by generating the output decision from each internal node, we move down to the

leaf which gives us the predicted label [29]. This algorithm is non-parametric and also can deal with complex and large dataset effectively [30].

AdaBoost: The algorithms that we is explained so far are single classifiers where a single startegy or rule is followed to generate the result. However, boosting is known as ensemble classifier where more than one comparatively identified as weak rule or classifiers results are combined to take decision. They are capable to generate highly accurate reults by combining the comparatively weak rules. The first boosting algorithm which was practically implemented and mostly used and studied is Adaboost [31, 32].

Gradient Boosting: Gradient Boosting algorithm is known as one of the most powerful machine laerning algorithm with a uge variety of application. It is also possible to customize this algorithm based on our need by changing the loss function accordingly. It is also an enseble classifier which incorporates more than one decision to generate the final output [33]. However, it may not produce expected outcome all the time if the number of iterations and loss function is not handles properly[34].

Rotation Forest: Rotation Forest (RoF) is a tree-based ensemble method that randomly splits the data into k subsets, and then applies Principal Component Analysis (PCA) to each segment for feature transformation [39]. In the end, the results of classification using decision trees on those transformed features are aggregated using an ensemble to produce the final result. RoF is a supervised learning model in the area of machine learning. The idea behind RoF is to introduce a mechanism to encourage the individual accuracy and diversity of each base classifier at the same time [39]. The decision tree is mainly used as the base classifier for RoF. One of the reasons for choosing the decision tree is that it is sensitive to the rotation of the feature axes and can produce better prediction performance [39]. It has demonstrated promising results for similar studies found in the literature [49-54].

Rotation forest classifier is another example of ensemble machine learning algorithm which combines the decision of different rules. It is basically a tree based ensemble method which works by randomly splitting the dataset into some sub datasets and applies some analysis named PCA (pricipal component analysis) for each sub dataset to extract the features [35]. Decision tree is basically used as base classifier for prediction as decision tree is sensitive on how we are keeping the data and rotation of each axis, it shows different output which may help us to find the best one [35].

Multi-layer Perceptron classifier: Multi-layer Perceptron (MLP) is known as neural network based machine learning algorithm which works in different layer where the input layer

takes the input and output layer produces the final output and the hidden layers in between them are responsible to handle the processing. More than one hidden layers are possible here and all together creates a connected netwrok which indicates that all the node of upper layer has a connection with every node in the lower layer [36]. MLP is also widely used neural network based algorithm which works better with classification problems.

2.2 Literature Review

A good number of computational method have been introduced so far in order to predict methylation site. We have analyzed the most relevant and recent one to explore the current status.

In 2009, Ming Shien et al. proposed a new method called MASA to tackle this problem [5]. To build this model, they extracted information related to solvent-accessible surface area from the surroundings amino acids of a methylation site. They also used Support Vector Machine (SVM) as their classifier to build MASA.

In the same year, Jianlin Shao et al. used Bi-profile Bayes in order to extract sequential features and used SVM classifier to solve this problem [18]. Although they could not achieve better performance than MASA, they still showed the potential for using bi-profile analysis for methylation prediction.

Later on, Ping Shi et al. proposed a method named PMeS which used enhanced sequential and physicochemical-based feature extraction techniques [19]. They extracted the coding for sparse property, van der Waals volume with a normalized format, composition of amino acid with positive weights, and lastly the surface area to build their feature vector. They also used SVM as their classification technique.

In 2017, Leyi Wei et al. proposed a new method called MePred-RF [37]. This method was built using Random Forest (RF) as the classification technique and sequential based feature. They demonstrated that their model is able to outperform previous studies. Despite achieving promising results, all these studies used small benchmarks to investigate the performance of their model. They also mainly relied on sequential and physicochemical properties for feature extraction.

Later on, Qiu et al. proposed a model based on the structural information of single-residue [3]. However, they solely focused on just the methylation site for feature extraction rather than including the neighboring residues. The combined features extracted from the secondary

structure, accessible surface area, electrostatic potential, protrusion index, depth index, and the interaction network of the methylation site.

In 2020, Biggar et al. proposed a new method called MethylSight which is built by combining the prediction of in silico method with the mass spectrometry targeted value in order to identify the lysine methylation sites [7]. They also used SVM as their classification technique. They incorporated confidence score which is in between 0 to 1 as a measure of probability in order to help the users to decide the level of recall.

Most recently, Wei Zheng et al., proposed Met-predictor to predict both methylation site and methylation types (mono, di and tri-methylated) [20] While most of the protein methylation prediction technique is designed based on sequential or physicochemical properties for feature extraction, they utilized the tertiary structure information as well in order to enhance the prediction result. They combined sequential model features with predicted structural information to build their feature vector. They also used SVM as their classifier and obtained better results compared to their previous studies.

The summary of literature review is given on table 2.2.

2.3 Gap Analysis

After analyzing the existing study for methylation prediction of lysine residue, we have found out some places to study more. The points are given bellow with respect to different criteria.

2.3.1 Dataset

In our study we have selected to extrcat data from Compendium of Protein Lysine Modifications (CPLM) dataset which have not been used directly by any of the previous studies for methylation site prediction. This dataset contains many PTMs that have been detemined accurately using different experimental techniques [38]. To our best knowledge, though this dataset has not been used directly for methylation site prediction, it showed to have a good impact on other PTMs prediction in a good number of recent studies [39–42].

2.3.2 Features

Due to the limitation of sequential based features in terms of performance, studies moved forward to analyze the structural information. Since structural information was extracted from 3D structure of protein for better result, it creates a limitation for those protein where the 3D

Author's name with year	Dataset	Features	Predictor classifier	Result
Ming Shien et al.2009 [9]	4 combination. Total: 91,182	solvent-accessible area from the surroundings amino acids is used.	Support Vector Machine	Accuracy: 80.8%
Jianlin Shao et al.2009 [9]	188 positives and 564 negatives for lysine	Sequential features extracted using Bi-profile Bayes.	Support Vector Machine	Sensitivity: 70.05% Specificity: 77.08% Accuracy: 75.51%
Ping Shi et al.2012 [11]	322 positive sites and 4126 negative sites for methyllysine	sparse property, van der Waals volume with a normalized format, lastly the surface area to build feature vector.	Support Vector Machine	Specificity: 99.23% Accuracy: 91.16%
Leyi Wei et al. 2017 [12]	1,744 samples, of which true Kmethylated and non-K-methylated samples were 226 and 1518 respectively.	Sequential based feature	Random Forest	Specificity: 84.6% Accuracy: 80.7%
Wankun Deng et al. 2017 [13]	1521 positive sites and 47 972 negative sites for lysine	solvent-accessible surface area, physicochemical properties, secondary structures.	Modified GPS 3.0 algorithm	Specificity: 95.04%
Qiu et al. 2018 [14]	254 true methylated sites and 674 non methylated sites for lysine.	Structural information.	Random Forest	Sensitivity: 95.1% Specificity: 89%
Sarah Ilyas et al. 2019 [15]	670 positive samples, 984 negative samples	Position and composition relative features and statistical moments.	Neural Network based model	Accuracy: 96.7% (self consistency), 91.6% (cross-validation), 93.42% (jackknife testing)
Wei Zheng et al. 2020 [16]	5894 experimentally verified lysine methylation sites from 2849 protein sequences	Structural Features	Support Vector Machine	Accuracy: 59.7%
Biggar et al. 2020 [17]	Total 19988 in training and 4996 in testing.	Silico method with the mass spectrometry targeted value.	Support Vector Machine	Competitive analysis of recalls

Table 2.2: Literature review summary table

structure is unknown. Therefore, we have choose to use predicted structural information as our features which will predict the structural information for proteins in order to find out the methylation site. We have also combined the predicted structural information with evolutionary information to have a strong feature set for methylation site prediction. To our best knowledge, no study is found to use combinational feature of predicted structural and evolutionary information as both of them plays important role in case of methylation site prediction.

2.3.3 Results

The results that we have achieved from literature so far is still not stable in different evaluation criteria. Therefore, There is a high need to improve the results as the predictors performance is still remained limited. In this aspect we focused in this study to have a better and stable result fro different evaluation criteria.

Chapter 3

Proposed Method

In this chapter we are going to have a clear and brief explanation of the materials that we have used for the proposed system and also a clear explanation of proposed methods.

3.1 Materials Methods

In our model, LyMethySE, we have used the evolutionary and predicted structural information that have been extracted from position-specific scoring matrix (PSSM) and SPIDER 2.0 [43, 44] respectively. To extract our features, the PSSM was extracted and translated to profile bigram [45, 46]. These incorporated features using PSSM-bigram and predicted structural information are then used to predict lysine methylation site. In this study, we have used 8 different classifiers as base classifier to compare their results and to identify the one with the best performance. These eight classifiers that have been used with different parameters are Support Vector Machine (SVM) with linear, polynomial and rbf kernel, Logistic Regression(LR), Gaussian Naïve Bayes(GNB), Bernoulli Naïve Bayes, Gradient Boosting(GB), Rotation Forest(RoF), MLP, AdaBoost, Decision Tree(DT). Among all these 8 classifiers, SVM with rbf kernel gives us better performance in different evolution criteria. In this section, we present the dataset, dataset balancing technique, feature extraction process and the classifiers that are used to for the prediction purpose.

3.1.1 Benchmark Dataset

This study uses Lysine Methylation PTM available at the Compendium of Protein Lysine Modifications (CPLM). CPLM contains different Lysine PTMs data that have been generated using

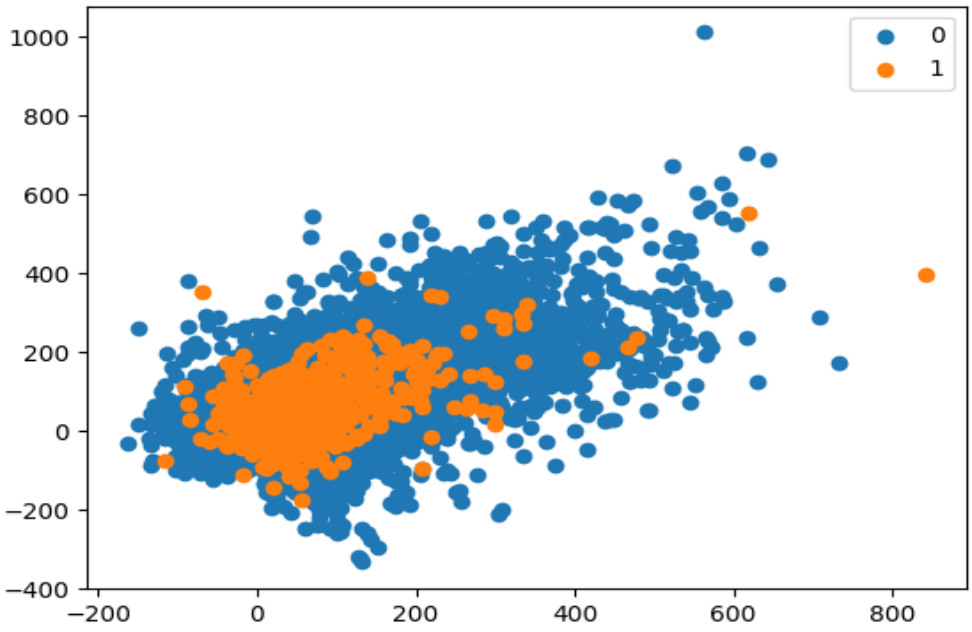

Figure 3.1: The imbalanced data distribution of positive(1) and negative(0) class

experimental methods [38]. The methylation dataset that we use includes 745 protein sequences with 1,144 lysine methylation sites. We then use CD-HIT [47–49] on this dataset to remove proteins with more than 40% sequential similarity. After removing the similar sequences, our dataset contains 633 protein sequences with 1116 methylated sites and 40,857 non-methylated sites.

3.1.2 Dataset Balancing

Our employed dataset has a ratio of 1:36 for methylated and non-methylated sites, which is highly imbalanced. Therefore, the model can positively bias the majority class, predicting non-methylated sites instead of methylated sites. Hence, it is essential to remove imbalances in data. The data distribution of our imbalanced dataset is given in figure 3.1.

A wide range of data balancing approaches is found in the literature. Broadly these approaches are divided into two groups, namely, oversampling [50] and under-sampling [51]. Oversampling refers to creating more data of minority class, whereas under-sampling points to reducing the number of instances of the majority class. However, in oversampling, the number of

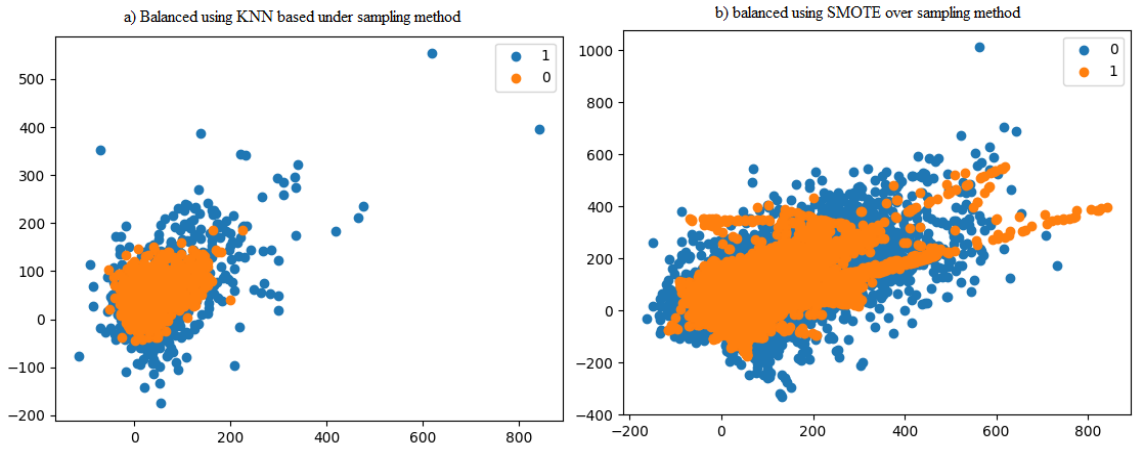

Figure 3.2: The balanced data distribution of positive(1) and negative(0) class using a) KNN based under sampling method and b) SMOTE over sampling method

samples in the minority class artificially increased, leading to overfitting or biased results. In our study, we try to investigate both the over sampling and under sampling method as well. Most widely used two undersampling and over sampling method are K-Nearest Neighbor (KNN) and Synthetic Minority Over-sampling Technique (SMOTE) respectively are investigated in this study. The data distribution visualization is given in figure 3.2. In this figure 3.2, a) represents the data distribution after balancing using KNN based under sampling of majority class and b) represents the data distribution after balancing using SMOTE over sampling of minority class. However, the KNN based under sampling method of majority class showed better performance in order to extract important information for our study. Performance details is given on next chapter.

Hence, in this study, to build our model LyMethSE, we use the under-sampling method instead of oversampling. We investigate different under-sampling approaches, among them K-Nearest Neighbor (KNN), as one of the most widely used under-sampling methods for data balancing [41, 52, 53], obtained the balancing outcome results. The KNN, used in this study is built based on the Euclidean distance between all the lysine residues. We remove those negative residues in the data, with at least one positive data within the k-th nearest neighbors. After investigating different values for k, we identify k=375, which leads to the balanced dataset with a 1:1 ratio. After balancing the dataset the final number of positive instances or methylated samples in the dataset, are 1116, and the number of negative instances or non-methylated samples are 1101.

3.1.3 Train and Test Dataset

To investigate our model's generality and avoid overfitting due to balancing our data, we divide our dataset into 80:20 instances where 80% of the dataset is used to train the model, and the other 20% is used as the test data. The testing data is not used for training or parameter tuning. We also apply 10-fold cross-validation on the training data to assess the generality of our model.

3.1.4 Features Extraction

In this section, we explain our extracted features and feature selection process. Our extracted features include evolutionary information incorporated in the form of profile Bi-gram combined with predicted structural information extracted directly from the protein sequence.

3.1.4.1 Evolutionary Information Based Features

We use evolutionary information extracted from the position-specific scoring matrix (PSSM) [43]. Evolutionary features, especially those extracted from PSSM, have been shown effective to solve similar problems [54–56]. PSSM is an N $\times$ 20 matrix where N represents the protein sequence length, that presents the substitution probability of given amino acid with other amino acids for its position along the protein sequence. To generate PSSM, we run PSI-BLAST against a non-redundant (NR) protein databank with three iterations and a threshold (E) of 0.001. We then transform PSSM in terms of profile bigram [45, 46].To the best of our knowledge, PSSM has never been used for feature extraction for the Methylation site prediction problem.

3.1.4.2 Profile Bigram

Profile bigram is a well-established method to extract significant discriminatory information from PSSM matrix [16][25]. Evolutionary-based profile bigram is extracted in the following manner.

$$Bi = Bi_{p,q} \sum_{k=1}^{30} r_{k,p} r_{k+1,q} \tag{3.1}$$

Where, Bi denotes the profile bigram of the PSSM matrix R and r_{ij} is indicating a single element of R. The value of p and q here is from 1 to 20 equal to the number of amino acids building the protein sequence which is representing a 20 x 20 matrix. This 20 x 20 matrix is then converted into a flat 400-dimension feature vector. This resulted matrix is considered as

our feature vector.

$$ProBi = [Bi_{1,1}, Bi_{1,2}, \ldots, Bi_{1,20}, Bi_{2,1}, Bi_{2,2} \ldots, Bi_{20,1}, Bi_{20,2}, \ldots, Bi_{20,19}, Bi_{20,20}] \quad (3.2)$$

In this study, the window size that is used to produce the PSSM is 15. This denotes that 15 amino acid from upside and 15 from downside is considered as window for PSSM generation. We investigated different window sizes to extract profile bigram from PSSM which among them, using 15 demonstrated the best results. If the part of protein sequence that we are considering is named as PE, which is having a lysine residue at the middle of the portion and 15 amino acids at upside and 15 more at downside, then the visibility of that segment will be,

$$P_E = A_{-15}, A_{-14}, \ldots., , A_{-1}, K, A_1 \ldots., A_{14}, A_{15} \quad (3.3)$$

Where K is the targeted lysine residue for evolutionary features. The amino acids named as A_{-15} to A_{-1} and A_1 to A_{15} are the upside amino acids and downside amino acids respectively for the targeted lysine in the middle.

In a case of protein head and tail parts, where there are not enough amino acids to build the complete window, we use mirror method to fill the required number. In this method, we filling out the missing amino acids from the data that comes when we mirror reflect the other part. This method is shown in figure 3.3

3.1.4.3 Structural Information Based Features

In this study, we use the secondary structure, the backbone torsion angles and the Accessible Surface Area (ASA) as structural features. The secondary structure have coil, strand and helix named three states and the torsion angles have θ, τ, ψ, and ϕ named four angles. From the previous studies [41, 52–54, 57, 58], we have seen that the structural information such as secondary structure, ASA and torsion angles plays an important role in prediction or classification work. In this study, we use SPIDER 2.0 [43, 44] which can give us a predicted value for those selected parameters. SPIDER 2.0 is considered as a tool that helps us to find out the predicted structural information of a protein sequence. This is known as a machine learning based tool that is built with an architecture of deep learning methods. As an output, SPIDER 2.0 provides us the predicted values of amino acid incorporated with the protein sequence in a matrix form (named the matrix as SPD2). This tool is also considered as one of the best tool in order to predict the three structural features (secondary structure, ASA, torsion angles) [59–61]. For these study, the structural features that we have used are explained in the following parts,

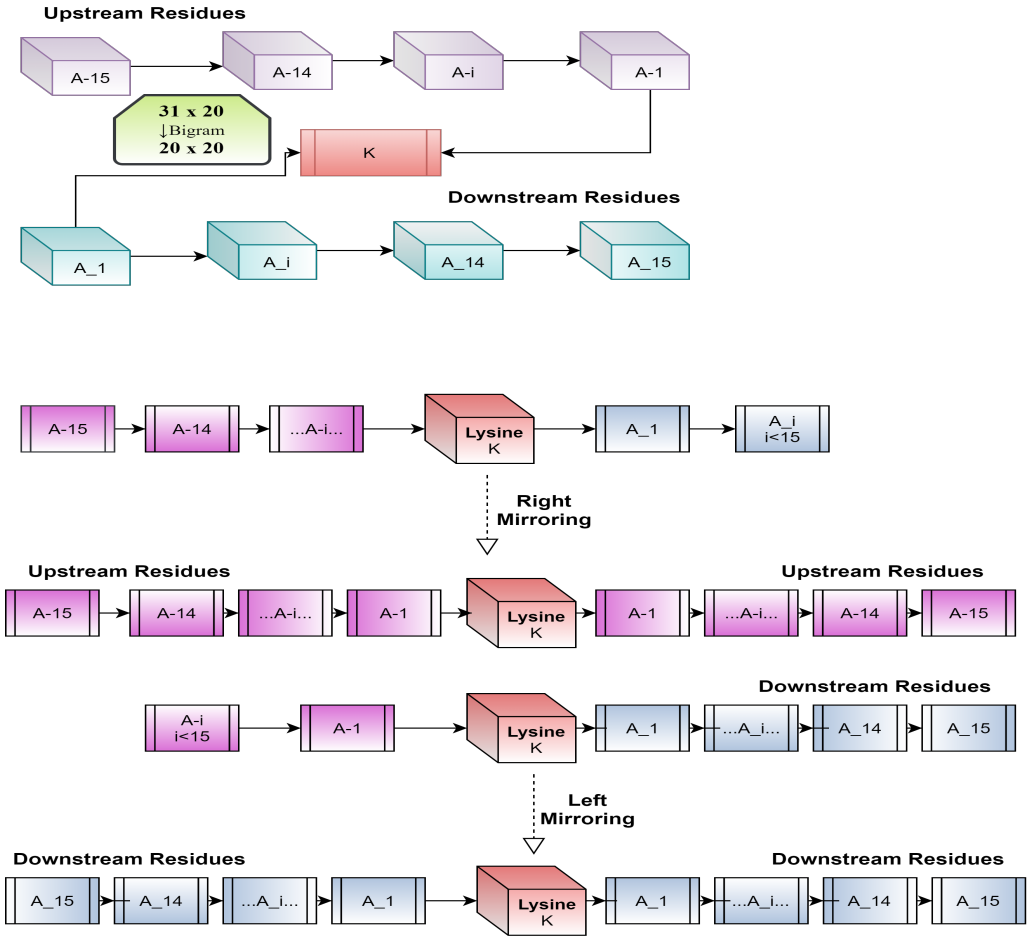

Figure 3.3: Extracting features from the neighboring amino acids using mirror method for amino acids at the two ends of a protein

- **Secondary Structure:** The secondary structure is used to identify the basic three dimensional structure of a protein data. Coil, strand and helix are the three objects of a secondary structure. When we are predicting the secondary structure of a protein, it is basically giving us the prediction of building one of these three components for every amino acid. This information helps us to differentiate the amino acids that are more stable and the amino acids that are more capable to interrelate with other molecules. As an output of the secondary structure, the SPD2 produces a matrix of N x 3 size where the N denotes the length of the protein sequence and the 3 columns are denoting the three different components of secondary structure [62].

- **Accessible Surface Area:** The surroundings of the methylation site also plays an important role as structural information which is named as a solvent-accessible surface area (ASA) of methylation site. SPD2 is known as the best way to find out the structural information and the ASA of protein sequence [5]. The ASA provides us information about the amino acids that relies on the surface area and which has more chances to endure Post Transitional Modifications. We have generated the final ASA result by analyzing the SPD2 result for each sequences of proteins.

- **Torsion Angles:** The torsion angles are used as a local backbone angle of a protein sequence. It represents the local structure of protein. While the secondary structure give us basic knowledge about the local configurations, the torsion angle provides us a successive information of the protein sequence local structure. It basically provides us the successive information about interactivity of amino acid. The torsion angles generally refer to four angles that are θ, τ, ψ, and ϕ [43, 44]. The probabilistic value of these angles are produced as an output of SPD2.

3.2 Classifiers

To investigate the effectiveness of our model, we employed 8 classifiers where most of them have been successfully used for PTM site prediction task [5–7, 9, 18, 20, 56]. Note that these classifiers were studied just using sequential and structural features. Hence, investigating the effectiveness of these classifiers, we provide a comparison between different classifiers using combinational features of evolutionary-based profile bigram and predicted structural information as input feature. The classifiers that are explored in this study are, Support Vector Machine (SVM) with linear, polynomial and rbf kernel, Logistic Regression(LR), Gaussian Naïve Bayes(GNB),

Bernoulli Naïve Bayes, Gradient Boosting(GB), Rotation Forest(RoF), MLP, AdaBoost, Decision Tree(DT). Among all these 9 classifiers, we achieve the best result using SVM with respect to different evolution criteria. Therefore, we build our own model LyMethSE using the SVM classifier.

3.2.1 Support Vector Machine

Support Vector Machine (SVM) is known as one of the most widely used Machine Learning algorithms. This classical machine learning algorithm is proved to perform better than many other classifiers for different tasks. It can solve both linear and nonlinear problems and has been used in a wide range of real-world scenarios. SVM aims to find the largest marginal hyper plane between classes based on support vector theory to reduce the prediction error and enhance the generality of the classification task [21]. Different kernel functions are used along with the SVM algorithm, such as linear, nonlinear, polynomial, radial basis function (RBF), and sigmoid [63]. SVM has also been widely used for different Biological Applications [21, 63]. In case of protein sequence analysis, SVM shows a very promising performance compared to other classifiers [64–66]. SVM is also widely used to predict methylation sites in the literature as well [5, 7, 18–20]. To build LyMethSE, we use SVM with rbf as its kernel function, demonstrating the best results compared to other kernels.

3.2.2 Other applied classifiers

Though we used SVM as a classifier to build LyMethSE, we applied some other classifiers to our extracted features in order to identify the methylated and non-methylated sites. We used different single model classifiers, tree based classifiers, ensemble classifiers, tree based ensemble classifiers and neural network models to find out how these classifiers perform with our extracted features and also to prove the generosity of our feature engineering method. The algorithms that are used for performance evaluation are, Support vector machine(SVM), Logistic regression, Gaussian Naive Bayes, Bernoulli Naive Bayes, Decision Tree, Rotation Forest, Adaboost, Gradient Boosting and Multi layer Perceptron Classifier.

3.3 Performance Evaluation

In order to have an accurate analysis result, it is really important to design a proper performance evaluation matrix. To evaluate our model, we use six different evaluation criteria for

performance analysis namely, Accuracy (Acc), Sensitivity (Sn), Specificity (Sp), Precision (Pr), F1-score, and Matthews correlation coefficient (MCC). These evaluation metrics are calculated as follows,

$$Acc = \frac{TP + TN}{TP + TN + FP + FN} * 100 \tag{3.4}$$

$$Sn = \frac{TP}{TP + FN} * 100 \tag{3.5}$$

$$Sp = \frac{TN}{TN + FP} * 100 \tag{3.6}$$

$$Pr = \frac{TP}{TN + FP} \tag{3.7}$$

$$F1 = \frac{2 * SN * PR}{PR + SN} \tag{3.8}$$

$$MCC = \frac{(TP * TN) - (FP * FN)}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \tag{3.9}$$

Where, TP, TN, FP and FN are True Positive, True Negative, False Positive and False Negative, respectively. True Positive (TP) refers to those instances that are methylated and correctly classified as methylated sites, True Negative (TN) refers to the instances that are non-methylated and correctly classified as non-methylated, False Positive (FP) reflects the number of those instances that are classified as Methylated sites while they are actually non-methylated sites, and False Negative (FN) value refers the number of instances that are classified as non-methylated while they are methylated sites.

3.3.1 Validation

In order to evaluate the performance of our model, we use both independent test set, and 10-fold cross-validation. In K-fold cross-validation we divide the employed dataset into K segments and then use K-1 segments to train the model and the last segment is used to test. We then repeat this model K times to use all the samples exactly once for testing. Note that we use 10-fold cross-validation on the training data. The testing data is not used for any parameter tuning. The overall working mechanism and design of our model is shown in a diagram given on figure 3.4.

3.4 Summary

In this chapter we briefly describe the dataset that we have used and the pre-processing steps of dataset preparation. It also describes the feature preparation phases with detailed description of

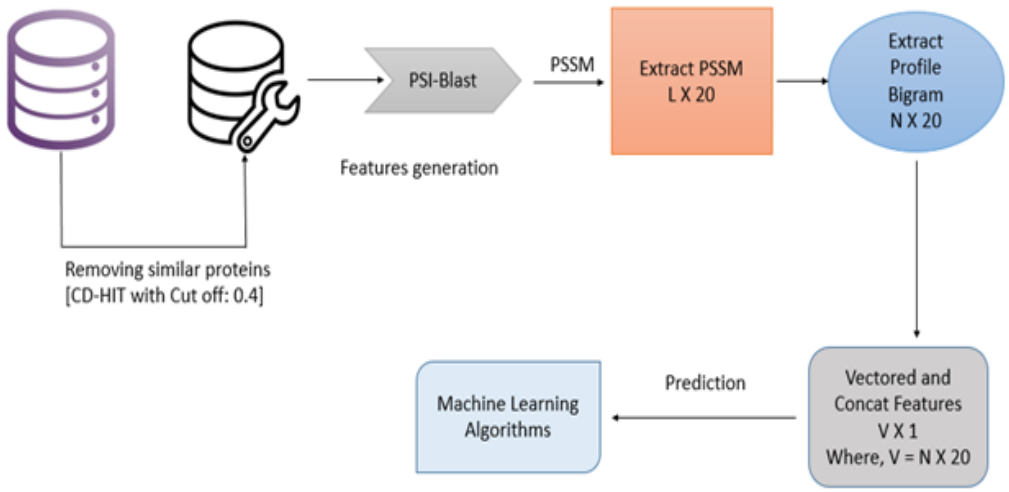

Figure 3.4: The working methodology of LyMethySE

the features that we prepare to use. Moreover it describes the implemented classifiers along with the classifier that we have chosen to use as base classifier to predict methylation for LyMethSE. The performane eveluation method and validation method are also described in this chapter.

Chapter 4

Experimental Analysis

In this chapter we describe the results that we have achieved using combinational features. We also describe the competitive analysis of using different features extraction methods.

4.1 Independent train and test set

In this study we have used a balanced CPLM dataset which is having 1116 methylated samples and 1101 non-methylated samples. To validate our model, we divided the dataset and 20% of it is used to test the model and 80% of the dataset is used to train our model.

4.2 Experimental Results

From literature, we see that non of the study used evolutionary based features. However, it showed an amazing output in case of other PTM prediction. Again, evolutionary information extracted from PSSM and transforming it to profile bigram also showed an amazing feature extraction technique for PTM. Therefore, we extracted the evolutionary based profile bigram as features first and applied different machine learning algorithms to identify the influence of this feature in case of methylation site prediction. The results that we have achieved using evolutionary based profile bigram as extracted features for 10 fold cross validation is given on table 4.1.

From table 4.1 we can see that most of the classifiers (except Gaussian NB and Bernoulli NB) is having a good performance in case of different evaluation criteria. We can also say that Naive Bayes based algorithm, were we follow the principle of conditional probability is not going with this prediction method where we have used evolutionary based information converted to

Classifier	ACC(%)	MCC	Precision (%)	Roc AUC score (%)	F1 score	Sensitivity (%)	Specificity (%)
Logistic Regression (LR)	97.4	0.94	97.8	97.4	0.96	96.9	97.6
Decision Tree (DT)	92.1	0.83	92.7	91.1	0.90	91.5	93.1
Gradient Boosting (GB)	95.1	0.90	95.7	94.3	0.93	93.2	95.8
Gaussian NB	73.5	0.49	84.5	73.3	0.67	56.6	90.0
AdaBoost	98.1	0.96	97.6	98.3	0.97	97.4	98.6
BernoulliNB	87.3	0.75	82.9	87.4	0.87	94.1	81.8
MLP	97.2	0.94	97.8	97.0	0.97	96.6	97.9
SVM	98.7	0.96	98.2	97.9	0.98	98.4	98.8

Table 4.1: Results of different classifiers using evolutionary based profile bigram as features for 10 fold cross validation

profile bigram as feature set. We can also observe that neural network based algorithm MLP and ensemble learning based algorithm AdaBoost and Gradient Boosting is having very nearer result to single and basic classifier Logistic regression and SVM. It shows the generosity of our extracted features and also shows the impact of evolutionary based features in case of methylation site prediction.

In order to investigate more precisely, we also generated the independent test set results for evolutionary based profile bigram extracted features for different machine learning algorithms. Our observation of its performance are given on table 4.2. From the table we can see that here also all the different classifiers except the naive bayes based classifiers are providing a good result in case of different performance evaluation criteria. Here the decision tree classifier is giving a little less result compared to other ensemble and single classifiers, however, we cannot ignore the results as it is having accuracy, precision ROC AUC score, sensitivity and specificity above 90%. In case of other present ensembles and single classifiers, they are providing accuracy, precision ROC AUC score, sensitivity and specificity over 95%. From both table 4.1 and table 4.2 where the 10 fold cross validation and independent test set results are given for evolutionary based profile bigram features we can say state evolutionary based information is having a great influence in methylation site prediction. They are having a very promising result in different performance measure criteria when we are using different kind of machine learning algorithms.

To investigate more and in the hope to improve the results more, we incorporate the pre-

Classifier	ACC(%)	MCC	Precision (%)	Roc AUC score (%)	F1 score	Sensitivity (%)	Specificity (%)
Logistic Regression (LR)	97.1	0.93	96.8	97.2	0.96	97.3	96.9
Decision Tree (DT)	91.2	0.82	90.7	91.3	0.91	91.5	90.9
Gradient Boosting (GB)	96.2	0.92	95.5	96.3	0.96	96.8	95.8
Gaussian NB	74.2	0.52	88.4	73.9	0.67	54.9	93.0
AdaBoost	97.1	0.94	96.8	97.1	0.97	97.3	96.9
BernoulliNB	85.9	0.72	82.5	86.1	0.86	90.6	81.3
MLP	97.8	0.94	96.5	97.8	0.97	98.8	96.5
SVM	97.3	0.94	98.3	97.6	0.97	97.9	98.2

Table 4.2: Results of different classifiers using evolutionary based profile bigram as features for independent test set

dicted structural based features with evolutionary based bigram profile and try to apply the same algorithms to identify how they perform with combinational feature. From literature we already identify that structural information has great influence in methylation site prediction. Therefore, we want to incorporate two different important features together in this prediction technique to achieve better results. In this regard, we build our model LyMethSE using incorporated combinational feature of evolutionary based bigram profile information and predicted structural information. As from table 4.1 and table 4.2 we identify that evolutionary based information is working very well with SVM. We use SVM as base classifier to predict the methylation site in our model LyMethSE. The results that we achieve using combinational feature extraction method for our model for 10 fold cross validation are given on table 4.3. From the table we can see that LyMethSE is having a satisfying result in comparison with other classifiers. In table 4.3 we can find that in LyMethSE is having 98.5% accuracy, 0.97 MCC, 98.6% precision value, 98.5% Roc AUC score, 0.99 F1 score, 98.5% sensitivity and 98.5% specificity. In comparison with other classifiers, LyMethSE is outperforming in all the evaluation criteria. However, we can not ignore the results that we are having for other different classifiers. Except the naive bayes based classifiers, almost all other classifiers are performing very well with combinational features and also providing a very promising results in different performance measurement criteria.

To investigate the combinational feature more, we also generate results for independent test

Classifier	ACC(%)	MCC	Precision (%)	Roc AUC score (%)	F1 score	Sensitivity (%)	Specificity (%)
Logistic Regression (LR)	97.3	0.95	97.6	97.3	0.97	97.1	97.6
Decision Tree (DT)	91.4	0.83	92.1	91.4	0.91	90.9	91.9
Gradient Boosting (GB)	96.1	0.92	97.2	96.1	0.96	95.0	97.3
Gaussian NB	74.6	0.51	84.3	74.7	0.71	60.9	88.4
AdaBoost	97.7	0.96	97.8	97.7	0.98	97.8	97.7
BernoulliNB	86.6	0.74	82.1	86.6	0.88	94.4	78.9
MLP	97.0	0.94	97.5	97.0	0.98	96.5	97.5
Rotation Forest	93.0	0.86	93.6	93.0	0.93	92.5	93.5
LyMethSE	**98.5**	**0.97**	**98.6**	**98.5**	**0.99**	**98.5**	**98.5**

Table 4.3: Results of LyMethSE along with other classifiers using 10 fold cross validation

score for the same classifiers that we have used for 10 fold to compare the results. The results that we achieve for independent test set are given in table 4.4.

Classifier	ACC(%)	MCC	Precision (%)	Roc AUC score (%)	F1 score	Sensitivity (%)	Specificity (%)
Logistic Regression (LR)	98.0	0.96	99.1	98.0	0.98	96.9	99.1
Decision Tree (DT)	93.5	0.87	95.3	93.5	0.93	91.5	95.5
Gradient Boosting (GB)	96.2	0.92	96.8	96.2	0.96	95.5	96.8
Gaussian NB	73.5	0.50	84.9	73.6	0.69	57.6	89.5
AdaBoost	98.4	0.97	100	98.4	0.98	96.9	100
BernoulliNB	89.2	0.79	85.4	89.2	0.89	94.6	83.7
MLP	97.5	0.95	99.5	97.5	0.97	95.5	99.5
Rotation Forest	91.9	0.84	93.1	91.9	0.92	90.6	93.2
LyMethSE	**98.9**	**0.98**	**100**	**98.9**	**0.99**	**97.8**	**100**

Table 4.4: Results of LyMethSE along with other classifiers using independent test set

As shown in Table 4.4, LyMethSE also outperforms and produced a significant result for independent test set as well in comparison with other eight classifiers. LyMethSE achieves

98.9%, 100%, 97.8%, 100%, 0.98, and 0.99 in terms of Accuracy, Precision, Sensitivity, Specificity, MCC, and F1-score, respectively. This result scheme is having best result in case of all the evaluation criteria. Achieving consistently higher results for both 10-fold cross-validation and independent test set demonstrates the generality of LyMethSE and its preference over other classifiers with combinational feature. Other different machine learning algorithms are performing well here too and the results they produced is having a very near one with LyMethSE. It proves the success of our feature extraction method and generation of combinational features with all the important information to predict methylation site.

As Results of combinational features demonstrated in Table 4.3 and Table 4.4 reflects the performance of different classifiers for combinational feature of evolutionary bi-profile and structural information. If we analyze the results more, we can get that the classification results are not varying much and almost all the classifiers are proving prediction accuracy results which is more than 90% in average (except GNB). As shown in these tables, all classifiers attain high prediction performance (excluding GNB) which demonstrates the effectiveness of using the feature of bi-gram profile to extract evolutionary information from PSSM and combining them with predicted structural information to tackle this problem. Using this combination of feature extraction process helps to keep the feature size constant even for different window sizes. This helps our model to obtain discriminatory information and also to keep the computational cost low for large amount of data. Therefore, not only LyMethSE is providing higher performance with respect to different evaluation scores, but also process large amount of protein sequences with a fixed size features which is beneficiary in terms of time consumption.

As from the discussion above we can say that using evolutionary based bi profile feature we have a very good result for different classifiers. However, combining it with predicted structural information help us to improve the result a little more. To compare and show the difference of using evolutionary based bi profile as features and also using combination of it with predicted structural information as feature, we generate a bar chart of MCC score and f1 score for both of the results that we achieve for this two feature extraction methods. The bar chart of MCC and f1 score for evolutionary based bigram profile as features and combination of it with predicted structural information as feature for 10 fold cross validation is shown in figure 4.1. We generate the same thing for independent test score as well which is given on figure 4.2. As our dataset was imbalanced and we balanced it using under sampling method, we focus to show the comparison result based on MCC and f1 score. From the bar chart we can say that for most of the classifiers, it worked a little better for combinational feature in terms of MCC

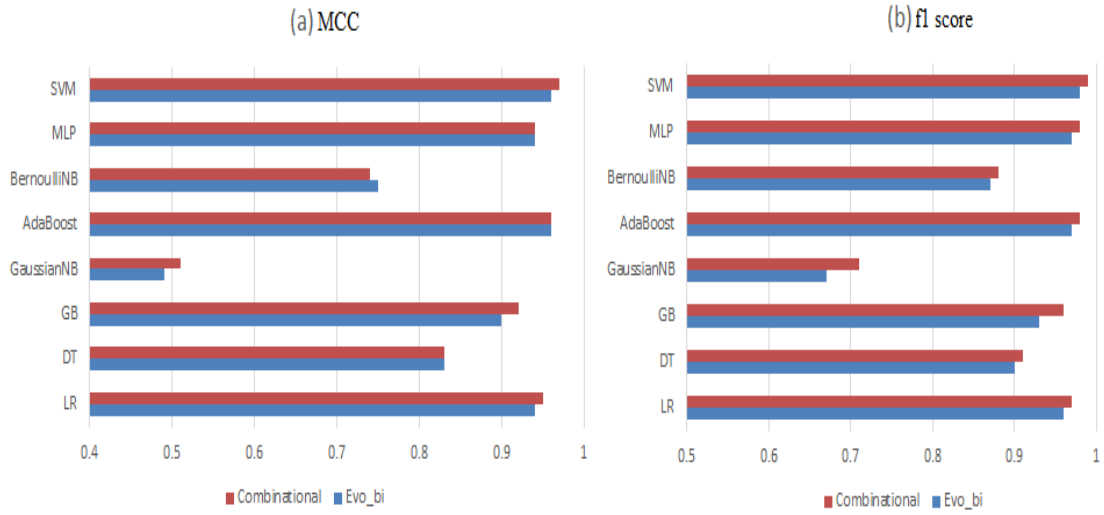

Figure 4.1: (a) MCC score and (b) f1 score of classifiers results for 10 fold cross validation using evolutionary based bigram as feature and combination of evolutionary based bigram with predicted structural information as feature

and f1 score for both cross validation and independent test set. It shows that combining the predicted structural information with the evolutionary based bigram profile help the model to improve its results in terms of different evaluation criteria.

To provide more insight to our achieve results, we generate the Receiver Operating Characteristic (ROC) curves for different used classifiers and our model LyMethSE. the ROC curves for 10-fold cross-validation and independent test set for different classifiers along with LyMethSE are given in Figure 4.3.

In figure 4.3, the x-axis shows the False Positive Rate or FPR and the y-axis shows the True Positive Rate or TRP. TPR and FPR basically refers to the sensitivity and specificity, respectively. We want to reach the optimum point for the TRP and FRP graph where the TRP touches the maximum value and FRP remains in minimum value. The left ROC curve reflects the results for 10-fold cross-validation whereas the right one is for independent test results. As shown in this figure, LyMethSE achieves better results compared to other classifiers which demonstrates the generality and effectiveness of this model. However, we can see from the ROC curve that some other classifiers such as Adaboost(AB), Logistic Regression(LR) and Gradient Boosting(GB) is having a promising result in the curve too. Note that the other classifiers that are shown in the curve are also using the same extracted features which are extracted using our novel approach. Promising and competitive results that are achieved by other classifiers using

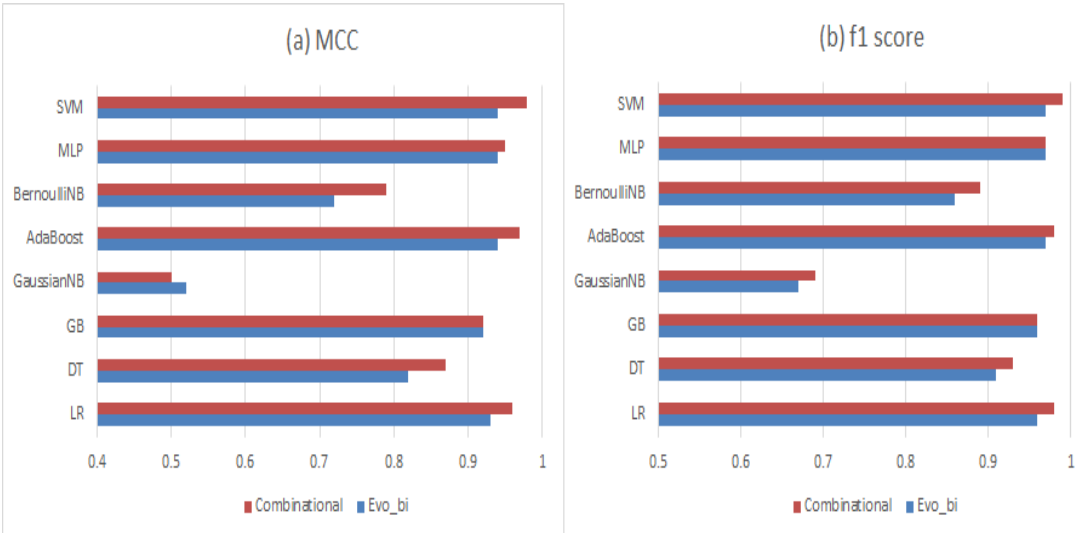

Figure 4.2: (a) MCC score and (b) f1 score of classifiers results for independent test set using evolutionary based bigram as feature and combination of evolutionary based bigram with predicted structural information as feature

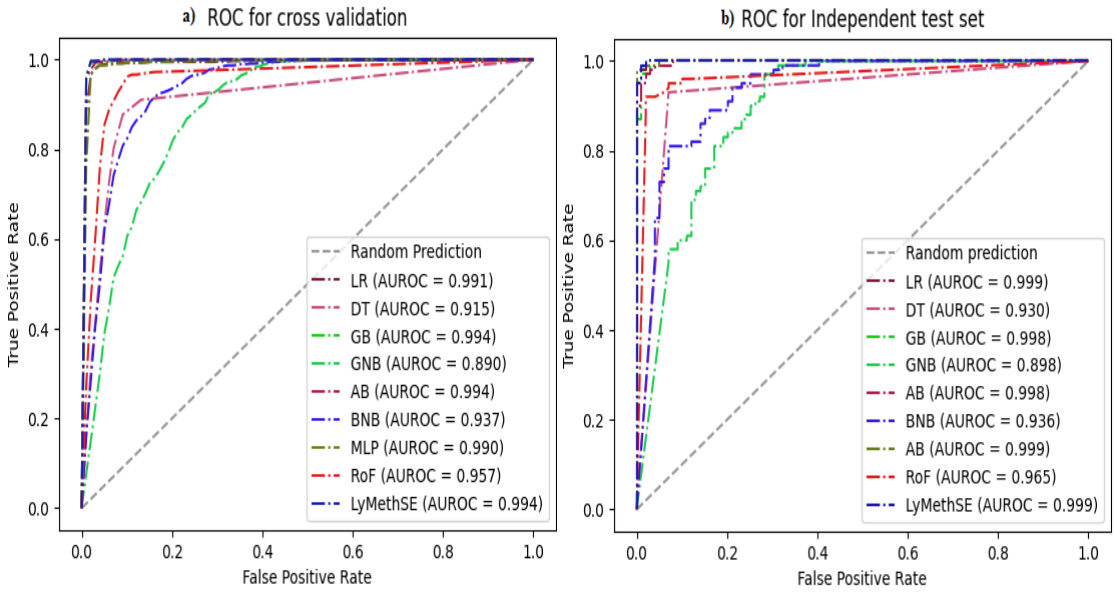

Figure 4.3: ROC Curve for TPR vs FPR using (a) 10-fold cross validation and (b) independent test set

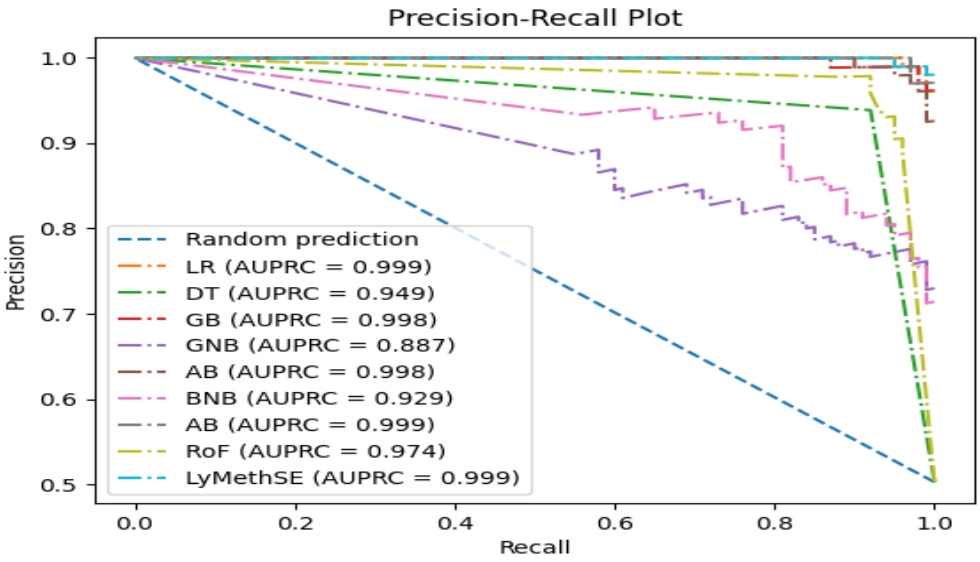

Figure 4.4: Precision-Recall Curve for LyMethySE model and other classifiers

our extracted features, demonstrates their effectiveness of this feature extraction technique in methylation site prediction task. From figure 4.3 (a) we can see that the AUC score of LyMethSE for cross validation is 0.994 and for the independent test set in figure 4.3 (b) is 0.999. Though in case of cross validation, Adaboost(AB) and Gradient Boosting(GB) and in case of independent test set Adaboost(AB) and Logistic Regression(LR) is having the same value for area under the ROC curve, by analyzing the curves, we can say that LyMethSE is holding the best area for both of the cases.

As it is discussed in the literature, achieving even the best AUC for ROC curve does not guarantee to have the best area under the curve for Precision-Recall (PR) curve [67]. Therefore, in figure 4.4, the precision-recall curves for LyMethSE compared to our employed classifiers are shown. In this curve, x-axis shows the level of recall and y-axis shows the level of precision. As shown in Figure 4.4, LyMethSE demonstrate AUC of 0.999 for PR curve which is in line with those reported for ROC curve and better than those reported for other classifiers. From the curve we also can see that Logistic regression(LR) and AdaBoost(AB) is also having the same area under PR curve. However, from the curve we can analyze and find out that LyMethSE is holding the best result for precision and recall. In other way, the outstanding results which is near to LyMethSE is actually reflecting the benefit of using combinational features of evolutionary and predicted structural information in the field of methylation site prediction.

As we discussed about different investigation methods to prove the generosity of our model and feature extraction technique, we did all this experiments using a dataset which was being balanced using KNN based under sampling method. In order to ensure if we are not using any important data while under sampling, we decide to implement a over sampling method as well and check the outcome to compare if our balancing method is able to extract the important information. In order to do this, we applied a most common and widely used minority class over sampling method Synthetic Minority Over-sampling Technique (SMOTE). In our dataset, there were 1116 positive instances which indicates the methylated site data and 40857 negative instances which indicated non-methylated sites. We applied the SMOTE to over sample the positive class using euclidean distance. After applying this balancing method, we have 40857 positive and negative instances. In this balanced dataset with combinational feature of evolutionary bigram and predicted structural information, we applied the previously best performing algorithms for independent test set. The results that we get is given on table 4.5. From the results we can identify that all the best performed machine learning algorithms with knn based under sampling method are giving much lower results with SMOTE over sampling method. Therefore, we can come to a point that this over sampling method is adding unnecessary information and thus our important information are not considered properly. However, it also proves that using knn based under sampling method of majority class here performed well for our dataset to extract and keep all the important information.

Classifier	**ACC(%)**	**MCC**	**Precision (%)**	**Roc AUC score (%)**	**F1 score**	**Sensitivity (%)**	**Specificity (%)**
Logistic Regression (LR)	73.2	0.46	72.4	73.1	0.74	75.6	70.6
Decision Tree (DT)	93.3	0.86	91.5	93.3	0.93	95.7	90.9
Gradient Boosting (GB)	86.9	0.73	85.1	86.8	0.87	89.7	84.0
AdaBoost	82.4	0.64	80.9	82.4	0.83	85.2	79.5
MLP	97.8	0.95	97.6	97.8	0.97	98.0	97.6
SVM	92.0	0.84	89.2	92.0	0.92	95.8	88.2

Table 4.5: Results of different classifiers using combinational features and SMOTE minority class over sampling method

The comparison of results achieved using LyMethSE compared to other classifiers indicates the preference of our proposed model compared to other classifiers used for this task. In

addition, achieving promising results for all the classifiers investigated in this study demonstrate the effectiveness of our proposed combinational feature of evolutionary bi-profile and predicted staructural information to predict protein Methylation sites.

4.3 Summary

In this chapter, we provide the brief result analysis of different classifiers along with LyMethSE. We have also shown the generality of our feature extraction method in this chapter. We analyze the ROC curve for the results of different classifiers with our extracted features. We generate the Precision-Recall curve as well for better comparison of curve areas.

Chapter 5

Conclusions, and Future Work

5.1 Conclusions

In this work, we introduce a novel predictor LyMethSE for methylation site prediction of lysine residues from a protein sequence. This method shows the effectiveness of using combination of evolutionary bi-profile information and predicted structural information as features for methylation site prediction. The evolutionary information extracted from protein sequence and then profile bigram was used to keep the features size fixed for different window size consideration. This features are then combined with structural information and generated the final features list. After extracting the features, data balancing was done using K-nearest Neighbor algorithm to avoid any biasness in result generation process. Different classifiers are used for predicting the sites for the extracted features and our model achieves a significantly outstanding result for such extracted features. The accuracy, sensitivity, specificity and MCC results that LyMethSE achieves are 98.9%, 97.8%, 100% and 0.98 for individual test score. In this study, we also present the influence of different classifiers in our extracted features. Thus, we see that LyMethSE along with SVM with rbf kernel achieves promising results in this aspect. In spite of achieving such promising results, the main lacking of such computational method is that we cannot identify the reason behind getting a lysine methylated or non-methylated.

5.2 Future Work

In future, we want to analyze the reason behind getting a lysine methylated or non-methylated. We also want to apply our extracted novel combinational features in different benchmarks and different methylated site prediction to achieve better results.

Bibliography

[1] E. M. Cornett, L. Ferry, P.-A. Defossez, and S. B. Rothbart, "Lysine methylation regulators moonlighting outside the epigenome," *Molecular cell*, vol. 75, no. 6, pp. 1092–1101, 2019. 1

[2] W.-R. Qiu, X. Xiao, W.-Z. Lin, and K.-C. Chou, "imethyl-pseaac: identification of protein methylation sites via a pseudo amino acid composition approach," *BioMed research international*, vol. 2014, 2014. 1

[3] H. Qiu, Y. Guo, L. Yu, X. Pu, and M. Li, "Predicting protein lysine methylation sites by incorporating single-residue structural features into chou's pseudo components," *Chemometrics and Intelligent Laboratory Systems*, vol. 179, pp. 31–38, 2018. 1, 2, 5, 10

[4] X.-J. Cao, A. M. Arnaudo, and B. A. Garcia, "Large-scale global identification of protein lysine methylation in vivo," *Epigenetics*, vol. 8, no. 5, pp. 477–485, 2013. 1, 2, 6

[5] D.-M. Shien, T.-Y. Lee, W.-C. Chang, J. B.-K. Hsu, J.-T. Horng, P.-C. Hsu, T.-Y. Wang, and H.-D. Huang, "Incorporating structural characteristics for identification of protein methylation sites," *Journal of computational chemistry*, vol. 30, no. 9, pp. 1532–1543, 2009. 1, 8, 10, 20, 21

[6] H. Liu, M. Galka, E. Mori, X. Liu, Y.-f. Lin, R. Wei, P. Pittock, C. Voss, G. Dhami, X. Li *et al.*, "A method for systematic mapping of protein lysine methylation identifies functions for hp1β in dna damage response," *Molecular cell*, vol. 50, no. 5, pp. 723–735, 2013. 1, 6

[7] K. K. Biggar, F. Charih, H. Liu, Y. B. Ruiz-Blanco, L. Stalker, A. Chopra, J. Connolly, H. Adhikary, K. Frensemier, M. Galka *et al.*, "Proteome-wide prediction of lysine methylation reveals novel histone marks and outlines the methyllysine proteome," *bioRxiv*, p. 274688, 2020. 2, 8, 11, 20, 21

[8] S. Ilyas, W. Hussain, A. Ashraf, Y. D. Khan, S. A. Khan, and K.-C. Chou, "imethylk-pseaac: Improving accuracy of lysine methylation sites identification by incorporating statistical moments and position relative features into general pseaac via chou's 5-steps rule," *Current Genomics*, vol. 20, no. 4, pp. 275–292, 2019. 2, 6

[9] H. Chen, Y. Xue, N. Huang, X. Yao, and Z. Sun, "Memo: a web tool for prediction of protein methylation modifications," *Nucleic acids research*, vol. 34, no. suppl_2, pp. W249–W253, 2006. 2, 6, 20

[10] O. Keskin, A. Gursoy, B. Ma, and R. Nussinov, "Principles of protein- protein interactions: what are the preferred ways for proteins to interact?" *Chemical reviews*, vol. 108, no. 4, pp. 1225–1244, 2008. 4

[11] C. Walsh, *Posttranslational modification of proteins: expanding nature's inventory*. Roberts and Company Publishers, 2006. 5

[12] S. Ramazi and J. Zahiri, "Post-translational modifications in proteins: resources, tools and prediction methods," *Database*, vol. 2021, 2021. 5

[13] Q. Bi, K. E. Goodman, J. Kaminsky, and J. Lessler, "What is machine learning? a primer for the epidemiologist," *American journal of epidemiology*, vol. 188, no. 12, pp. 2222–2239, 2019. 6

[14] E. Brynjolfsson and T. Mitchell, "What can machine learning do? workforce implications," *Science*, vol. 358, no. 6370, pp. 1530–1534, 2017. 7

[15] I. El Naqa and M. J. Murphy, "What is machine learning?" in *machine learning in radiation oncology*. Springer, 2015, pp. 3–11. 7

[16] S. M. Carlson and O. Gozani, "Emerging technologies to map the protein methylome," *Journal of molecular biology*, vol. 426, no. 20, pp. 3350–3362, 2014. 7

[17] K. K. Biggar, Y. Ruiz-Blanco, F. Charih, Q. Fang, J. Connolly, K. Frensemier, H. Adhikary, S. Li, and J. Green, "Methylsight: Taking a wider view of lysine methylation through computer-aided discovery to provide insight into the human methyl-lysine proteome," *bioRxiv*, p. 274688, 2018. 7

[18] J. Shao, D. Xu, S.-N. Tsai, Y. Wang, and S.-M. Ngai, "Computational identification of protein methylation sites through bi-profile bayes feature extraction," *PloS one*, vol. 4, no. 3, p. e4920, 2009. 8, 10, 20, 21

[19] S.-P. Shi, J.-D. Qiu, X.-Y. Sun, S.-B. Suo, S.-Y. Huang, and R.-P. Liang, "Pmes: prediction of methylation sites based on enhanced feature encoding scheme," *PloS one*, vol. 7, no. 6, p. e38772, 2012. 10

[20] W. Zheng, Q. Wuyun, M. Cheng, G. Hu, and Y. Zhang, "Two-level protein methylation prediction using structure model-based features," *Scientific reports*, vol. 10, no. 1, pp. 1–15, 2020. 8, 11, 20, 21

[21] W. S. Noble, "What is a support vector machine?" *Nature biotechnology*, vol. 24, no. 12, pp. 1565–1567, 2006. 8, 21

[22] D. A. Pisner and D. M. Schnyer, "Support vector machine," in *Machine learning*. Elsevier, 2020, pp. 101–121. 8

[23] P. C. Sen, M. Hajra, and M. Ghosh, "Supervised classification algorithms in machine learning: A survey and review," in *Emerging Technology in Modelling and Graphics: Proceedings of IEM Graph 2018*. Springer, 2020, pp. 99–111. 8

[24] T. G. Nick and K. M. Campbell, "Logistic regression," *Topics in biostatistics*, pp. 273–301, 2007. 8

[25] C. Bustamante, L. Garrido, and R. Soto, "Comparing fuzzy naive bayes and gaussian naive bayes for decision making in robocup 3d," in *MICAI 2006: Advances in Artificial Intelligence: 5th Mexican International Conference on Artificial Intelligence, Apizaco, Mexico, November 13-17, 2006. Proceedings 5*. Springer, 2006, pp. 237–247. 8

[26] R. D. Raizada and Y.-S. Lee, "Smoothness without smoothing: why gaussian naive bayes is not naive for multi-subject searchlight studies," *PloS one*, vol. 8, no. 7, p. e69566, 2013. 8

[27] A. McCallum, K. Nigam *et al.*, "A comparison of event models for naive bayes text classification," in *AAAI-98 workshop on learning for text categorization*, vol. 752, no. 1. Madison, WI, 1998, pp. 41–48. 8

[28] V. Narayanan, I. Arora, and A. Bhatia, "Fast and accurate sentiment classification using an enhanced naive bayes model," in *Intelligent Data Engineering and Automated Learning–IDEAL 2013: 14th International Conference, IDEAL 2013, Hefei, China, October 20-23, 2013. Proceedings 14*. Springer, 2013, pp. 194–201. 8

[29] J. R. Quinlan, "Learning decision tree classifiers," *ACM Computing Surveys (CSUR)*, vol. 28, no. 1, pp. 71–72, 1996. 9

[30] Y.-Y. Song and L. Ying, "Decision tree methods: applications for classification and prediction," *Shanghai archives of psychiatry*, vol. 27, no. 2, p. 130, 2015. 9

[31] R. E. Schapire, "Explaining adaboost," *Empirical Inference: Festschrift in Honor of Vladimir N. Vapnik*, pp. 37–52, 2013. 9

[32] A. Vezhnevets and V. Vezhnevets, "Modest adaboost-teaching adaboost to generalize better," in *Graphicon*, vol. 12, no. 5. Citeseer, 2005, pp. 987–997. 9

[33] A. Natekin and A. Knoll, "Gradient boosting machines, a tutorial," *Frontiers in neurorobotics*, vol. 7, p. 21, 2013. 9

[34] C. Bentéjac, A. Csörgő, and G. Martínez-Muñoz, "A comparative analysis of gradient boosting algorithms," *Artificial Intelligence Review*, vol. 54, pp. 1937–1967, 2021. 9

[35] J. J. Rodriguez, L. I. Kuncheva, and C. J. Alonso, "Rotation forest: A new classifier ensemble method," *IEEE transactions on pattern analysis and machine intelligence*, vol. 28, no. 10, pp. 1619–1630, 2006. 9

[36] S. Wan, Y. Liang, Y. Zhang, and M. Guizani, "Deep multi-layer perceptron classifier for behavior analysis to estimate parkinson's disease severity using smartphones," *IEEE Access*, vol. 6, pp. 36 825–36 833, 2018. 10

[37] L. Wei, P. Xing, G. Shi, Z. Ji, and Q. Zou, "Fast prediction of protein methylation sites using a sequence-based feature selection technique," *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol. 16, no. 4, pp. 1264–1273, 2017. 10

[38] Z. Liu, Y. Wang, T. Gao, Z. Pan, H. Cheng, Q. Yang, Z. Cheng, A. Guo, J. Ren, and Y. Xue, "Cplm: a database of protein lysine modifications," *Nucleic acids research*, vol. 42, no. D1, pp. D531–D536, 2014. 11, 15

[39] B. Trost, F. Maleki, A. Kusalik, and S. Napper, "Dapple 2: a tool for the homology-based prediction of post-translational modification sites," *Journal of Proteome Research*, vol. 15, no. 8, pp. 2760–2767, 2016. 11

[40] W. Bao and Z. Jiang, "Prediction of lysine pupylation sites with machine learning methods," in *International Conference on Intelligent Computing*. Springer, 2017, pp. 408–417.

[41] M. M. Islam, S. Saha, M. M. Rahman, S. Shatabda, D. M. Farid, and A. Dehzangi, "iprotgly-ss: Identifying protein glycation sites using sequence and structure based features," *Proteins: Structure, Function, and Bioinformatics*, vol. 86, no. 7, pp. 777–789, 2018. 16, 18

[42] M. Hasan, M. Khatun, and H. Kurata, "Computational modeling of lysine post-translational modification: an overview," *Curr Synthetic Sys Biol*, vol. 6, no. 137, pp. 2332–0737, 2018. 11

[43] R. Heffernan, K. Paliwal, J. Lyons, A. Dehzangi, A. Sharma, J. Wang, A. Sattar, Y. Yang, and Y. Zhou, "Improving prediction of secondary structure, local backbone angles and solvent accessible surface area of proteins by iterative deep learning," *Scientific reports*, vol. 5, no. 1, pp. 1–11, 2015. 14, 17, 18, 20

[44] Y. Yang, R. Heffernan, K. Paliwal, J. Lyons, A. Dehzangi, A. Sharma, J. Wang, A. Sattar, and Y. Zhou, "Spider2: a package to predict secondary structure, accessible surface area, and main-chain torsional angles by deep neural networks," in *Prediction of protein secondary structure*. Springer, 2017, pp. 55–63. 14, 18, 20

[45] A. Sharma, J. Lyons, A. Dehzangi, and K. K. Paliwal, "A feature extraction technique using bi-gram probabilities of position specific scoring matrix for protein fold recognition," *Journal of theoretical biology*, vol. 320, pp. 41–46, 2013. 14, 17

[46] A. Dehzangi, Y. López, S. P. Lal, G. Taherzadeh, J. Michaelson, A. Sattar, T. Tsunoda, and A. Sharma, "Pssm-suc: Accurately predicting succinylation using position specific scoring matrix into bigram for feature extraction," *Journal of theoretical biology*, vol. 425, pp. 97–102, 2017. 14, 17

[47] W. Li and A. Godzik, "Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences," *Bioinformatics*, vol. 22, no. 13, pp. 1658–1659, 2006. 15

[48] Y. Huang, B. Niu, Y. Gao, L. Fu, and W. Li, "Cd-hit suite: a web server for clustering and comparing biological sequences," *Bioinformatics*, vol. 26, no. 5, pp. 680–682, 2010.

[49] L. Fu, B. Niu, Z. Zhu, S. Wu, and W. Li, "Cd-hit: accelerated for clustering the next-generation sequencing data," *Bioinformatics*, vol. 28, no. 23, pp. 3150–3152, 2012. 15

[50] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "Smote: synthetic minority over-sampling technique," *Journal of artificial intelligence research*, vol. 16, pp. 321–357, 2002. 15

[51] H. He, Y. Bai, E. A. Garcia, and S. Li, "Adasyn: Adaptive synthetic sampling approach for imbalanced learning," in *2008 IEEE international joint conference on neural networks (IEEE world congress on computational intelligence)*. IEEE, 2008, pp. 1322–1328. 15

[52] A. Dehzangi, Y. López, S. P. Lal, G. Taherzadeh, A. Sattar, T. Tsunoda, and A. Sharma, "Improving succinylation prediction accuracy by incorporating the secondary structure via helix, strand and coil, and evolutionary information from profile bigrams," *PloS one*, vol. 13, no. 2, p. e0191900, 2018. 16, 18

[53] H. M. Reddy, A. Sharma, A. Dehzangi, D. Shigemizu, A. A. Chandra, and T. Tsunoda, "Glystruct: glycation prediction using structural properties of amino acid residues," *BMC bioinformatics*, vol. 19, no. 13, pp. 55–64, 2019. 16

[54] S. Shatabda, S. Saha, A. Sharma, and A. Dehzangi, "iphloc-es: identification of bacteriophage protein locations using evolutionary and structural features," *Journal of theoretical biology*, vol. 435, pp. 229–237, 2017. 17, 18

[55] M. R. Uddin, A. Sharma, D. M. Farid, M. M. Rahman, A. Dehzangi, and S. Shatabda, "Evostruct-sub: An accurate gram-positive protein subcellular localization predictor using evolutionary and structural features," *Journal of theoretical biology*, vol. 443, pp. 138–146, 2018.

[56] M. W. Ahmad, M. E. Arafat, G. Taherzadeh, A. Sharma, S. R. Dipta, A. Dehzangi, and S. Shatabda, "Mal-light: Enhancing lysine malonylation sites prediction problem using evolutionary-based features," *IEEE Access*, vol. 8, pp. 77 888–77 902, 2020. 17, 20

[57] A. Dehzangi, K. Paliwal, J. Lyons, A. Sharma, and A. Sattar, "Enhancing protein fold prediction accuracy using evolutionary and structural features," in *IAPR International Conference on Pattern Recognition in Bioinformatics*. Springer, 2013, pp. 196–207. 18

[58] S. Y. Chowdhury, S. Shatabda, and A. Dehzangi, "idnaprot-es: identification of dna-binding proteins using evolutionary and structural features," *Scientific reports*, vol. 7, no. 1, pp. 1–14, 2017. 18

[59] X. Xu, D.-J. Lee, S. Antani, and L. R. Long, "A spine x-ray image retrieval system using partial shape matching," *IEEE Transactions on Information Technology in Biomedicine*, vol. 12, no. 1, pp. 100–108, 2008. 18

[60] E. Faraggi, T. Zhang, Y. Yang, L. Kurgan, and Y. Zhou, "Spine x: improving protein secondary structure prediction by multistep learning coupled with prediction of solvent accessible surface area and backbone torsion angles," *Journal of computational chemistry*, vol. 33, no. 3, pp. 259–267, 2012.

[61] J. Lyons, A. Dehzangi, R. Heffernan, A. Sharma, K. Paliwal, A. Sattar, Y. Zhou, and Y. Yang, "Predicting backbone cα angles and dihedrals from protein sequences by stacked sparse auto-encoder deep neural network," *Journal of computational chemistry*, vol. 35, no. 28, pp. 2040–2046, 2014. 18

[62] S. R. Dipta, G. Taherzadeh, M. W. Ahmad, M. E. Arafat, S. Shatabda, and A. Dehzangi, "Semal: Accurate protein malonylation site predictor using structural and evolutionary information," *Computers in Biology and Medicine*, vol. 125, p. 104022, 2020. 20

[63] A. Patle and D. S. Chouhan, "Svm kernel functions for classification," in *2013 International Conference on Advances in Technology and Engineering (ICATE)*. IEEE, 2013, pp. 1–9. 21

[64] C. Cai, L. Han, Z. L. Ji, X. Chen, and Y. Z. Chen, "Svm-prot: web-based support vector machine software for functional classification of a protein from its primary sequence," *Nucleic acids research*, vol. 31, no. 13, pp. 3692–3697, 2003. 21

[65] D. P. Lewis, T. Jebara, and W. S. Noble, "Support vector machine learning from heterogeneous data: an empirical analysis using protein sequence and structure," *Bioinformatics*, vol. 22, no. 22, pp. 2753–2760, 2006.

[66] Y. Guo, L. Yu, Z. Wen, and M. Li, "Using support vector machine combined with auto covariance to predict protein–protein interactions from protein sequences," *Nucleic acids research*, vol. 36, no. 9, pp. 3025–3030, 2008. 21

[67] J. Davis and M. Goadrich, "The relationship between precision-recall and roc curves," in *Proceedings of the 23rd international conference on Machine learning*, 2006, pp. 233–240. 31