

Received November 5, 2020, accepted November 7, 2020, date of publication November 11, 2020, date of current version November 24, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.3037343

Discrimination of Golgi Proteins Through Efficient Exploitation of Hybrid Feature Spaces Coupled With SMOTE and Ensemble of Support Vector Machine

MUHAMMAD TAHIR¹, (Senior Member, IEEE), FAZLULLAH KHAN², (Senior Member, IEEE), MOHAMMAD KHALID IMAM RAHMANI¹, (Member, IEEE), AND VINH TRUONG HOANG³, (Member, IEEE)

¹Department of Computer Science, College of Computing and Informatics, Saudi Electronic University, Riyadh 11673, Saudi Arabia

²Institute of Social and Economic Research, Duy Tan University, Da Nang 550000, Vietnam

³Faculty of Computer Science, Ho Chi Minh City Open University, Ho Chi Minh City 70000, Vietnam

Corresponding author: Fazlullah Khan (fazlullahkhan@duytan.edu.vn)


ABSTRACT Many organelles inside and outside a living cell depend on the perfect behavior of Golgi apparatus for smooth and normal functioning. Its poor performance may lead to many inheritable diseases like diabetes and cancer. Therefore, it is highly crucial to detect any strange behavior of Golgi apparatus in advance. Accurate discrimination of *cis*-Golgi from *trans*-Golgi proteins surely helps researchers identify the role of Golgi proteins in various diseases and assist pharmacists in drug development. In this work, various hybrid models of Bi-Profile Bayes, Bigram PSSM, Di-Peptide Composition, and Split Amphiphilic Pseudo Amino Acid Composition with SMOTE oversampling technique have been employed to discriminate Golgi protein types. Multiple linear Support Vector Machines have been used to exploit the discrimination power of these models. The proposed prediction system: *Golgi-predictor* has shown significant performance and achieved promising results compared to other existing state-of-the-art techniques. Through the 10-fold cross-validation, the proposed system achieved an accuracy value of 97.6%, sensitivity value of 98.8%, specificity value of 96.5%, G-mean value of 97.6%, MCC value of 0.95, and F-score value of 0.97. Similarly, through the jackknife cross-validation, the achieved values for accuracy, sensitivity, specificity, G-mean, MCC, and F-score are respectively, 96.5%, 97.8%, 95.2%, 96.4%, 0.93, and 0.96. Moreover, through the independent dataset testing, *Golgi-predictor* demonstrated significant enhancement in performance over other techniques. The proposed methodology aims at supporting drug designers in pharmaceutical industry and assisting researchers from the fields of bioinformatics and computational biology towards better innovation in predicting the behavior of Golgi proteins.

INDEX TERMS Amphiphilic pseudo amino acid composition, bigram features, bi-profile Bayes, Golgi proteins, support vector machine, synthetic minority oversampling technique.

I. INTRODUCTION

Cells in different organisms, as well as in different parts of the same organism, perform unique functions, and possess distinctive features. However, most of the cellular processes are identical almost in all organisms, namely reproduction, energy conversion, and molecule transportation. The eukaryotic cell holds a defined nucleus, genetic material in the form

of DNA, and several other organelles, including cytoplasm, mitochondria, lysosome, endoplasmic reticulum, and Golgi apparatus that help it to carry out different activities such as digestion, movement, and reproduction [1]. Golgi apparatus is among the essential proteins that is composed of flattened sacs [2]. It further processes proteins and lipids received from endoplasmic reticulum [3] and package them for transportation to the exterior of cell or other locations in the same cell through secretory vesicles. Further processing of proteins and lipids inside Golgi apparatus happens systematically. There

The associate editor coordinating the review of this manuscript and approving it for publication was Yanjiao Chen .

are two faces of a Golgi apparatus: the *cis* face and *trans* face. The part in-between *cis* and *trans* faces is referred to as medial. Golgi apparatus receives proteins and lipids from endoplasmic reticulum at the *cis* face, whereas at the *trans* face, it ships its products towards various destinations [4]. Many inheritable diseases, such as cancer and diabetes, are due to the poor functionality of Golgi apparatus. Moreover, Golgi apparatus is considered an early target of Parkinson's and Alzheimer's diseases. Currently, anti-inflammatory and neuroprotective therapies exist that involve the usage of chemical drugs. However, they do not guarantee a permanent solution to these diseases [4]. Therefore, detection of any dysfunction or aberrant behavior in Golgi apparatus is needed ahead of time so that researchers may identify its role in the aforementioned diseases. Solutions, inspired from machine learning, have been proposed in recent years for solving a wide range of problems related to bioinformatics and computational biology domains [5], [6]. However, only a few have focused on Golgi proteins to classify them in *cis*-Golgi and *trans*-Golgi. In this connection, Ding *et al.* [7] proposed a modification of Pseudo Amino Acid Composition (PseAAC) coupled with modified Mahalanobis discriminant achieving 74.7% accuracy through jackknife testing. In another of their work, they used Di-Peptide Composition (DPC) features with Support Vector Machine (SVM) [8] that achieved 85.4% accuracy. In 2016, Yang *et al.* [9] suggested the utilization of Common Spatial Pattern features coupled with Random Forest based Recursive Feature Elimination. The resultant selected features are further exploited by Random Forest classifier to predict the two classes from Golgi proteins. The accuracies obtained are 90.1%, 88.5%, and 93.8% using 10-fold cross-validation, jackknife cross-validation, and independent dataset testing, respectively. In 2017, Ahmad *et al.* [4] have employed Bigram PSSM, split PseAAC, and DPC features in conjunction with SMOTE oversampling and Fisher's feature selection that achieved 94.9% accuracy through jackknife and 10-fold cross-validation testing. In continuation of their previous work, Ahmad and Hayat [10] proposed to use DPC with gap 3, SAAC, and PSSM based features in conjunction with SMOTE oversampling technique. Following this, Majority-Voting based Feature Selection (MVFS) is applied that selects high ranked features from the integrated space of selected features with 11 different feature selection techniques. The high-ranked features have been exploited by kNN classifier for their discriminative power that achieved 95.85% performance accuracy with 10-fold cross-validation. Cui *et al.* [11] have come up with a new idea of using part of the Golgi protein sequence instead of using the complete sequence. They used Enhanced Amino Acid Content Encoding [12] to encode the Golgi protein sequences in parts. In this way, they obtained 529 protein sequences in their dataset, which are exploited by Random Forest classifier with 1000 trees. The achieved performance accuracy is 80.0% by their proposed method. Recently, Zhou *et al.* [13] developed their prediction model by constructing the fused feature space of PseAAC, DPC, PsePSSM, and Encoding

Based on Grouped Weight (EBGW) that are passed through feature reduction technique namely conditional co-variance minimization. The feature space is oversampled with SMOTE technique to produce a balanced set, which is classified using XGBoost algorithm with performance accuracy of 92.1% through Jackknife testing protocol.

The literature review suggests the development of a more accurate and reliable model that could classify sub-Golgi proteins with higher success rates to meet the needs of the bioinformatics community. This paper presents a majority voting based ensemble of multiple SVMs that exploits the discriminative power of various hybrid models constructed from Bi-Profile Bayes (BPB), Bigram PSSM, DPC, and split Amphiphilic PseAAC (Amph-PseAAC) features. SMOTE oversampling technique is utilized to curb the imbalance present in the dataset. Our key contributions are:

- 1) We investigated the effect of different feature extraction strategies individually and collectively under the influence of SMOTE oversampling technique.
- 2) We computed Amph-PseAAC on C-terminus, N-terminus, and middle part of the sequence separately thus obtaining Amph-PseAAC for three subsequences.
- 3) We provide sufficient evaluations to demonstrate the improved performance of Linear-SVM.
- 4) We constructed an enhanced ensemble of better performing classifications.

The organization of this article is planned as follows. Section II describes the utilized datasets, proposes the *Golgi-predictor* system, explains the SMOTE oversampling, and closes the section with brief introduction of Support Vector Machines. Section III highlights the performance measures. Section IV presents simulation results and evaluates the performance of *Golgi-predictor* on benchmark datasets. Section V analyzes the comparative performance of *Golgi-predictor* and state-of-the-art techniques from the existing literature. Section VI concludes the article with final comments.

II. MATERIAL AND METHODS

In this section, we first present a brief description of the utilized datasets. Next, we discuss the proposed system. Then, we provide details about SMOTE oversampling. Finally, SVM based classification is presented towards the end of this section.

A. DATASETS

In this study, we utilized a highly imbalanced dataset comprising of 87 *cis*-Golgi and 217 *trans*-Golgi protein sequences originally constructed by Yang *et al.* [9]. Additionally, we used another dataset for independent testing that was proposed by Ding *et al.* [7] and employed by other investigators [4], [8], [9]. The independent dataset is also imbalanced consisting of 13 *cis*-Golgi and 51 *trans*-Golgi protein sequences. Both the datasets are publicly available at¹:

¹<https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4783950/>

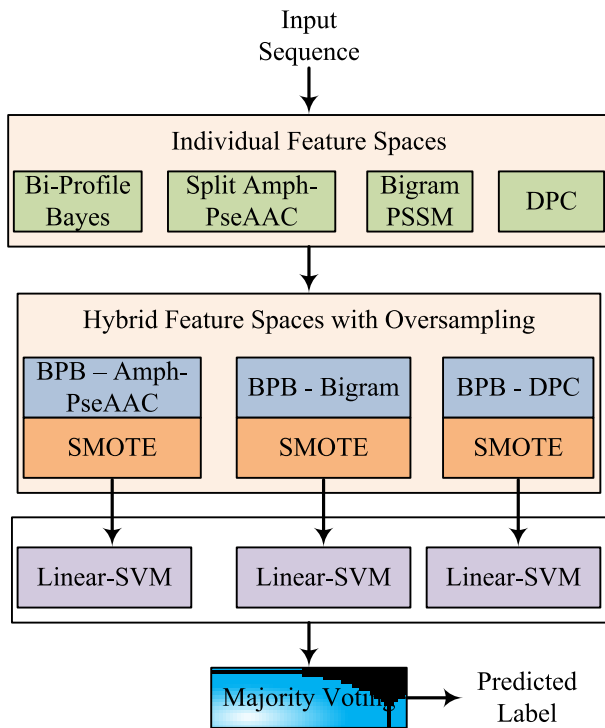


FIGURE 1. The proposed Golgi-predictor Schema.

B. THE PROPOSED SYSTEM

We propose the utilization of majority voting based ensemble strategy for enhanced prediction of Golgi sub-classes. Fig. 1 depicts the proposed prediction scheme: *Golgi-predictor*. In this work, multiple SVMs with linear kernel are first tuned using SMOTE based hybrid feature spaces of BPB, DPC, Bigram PSSM, and Amph-PseAAC. Then, majority voting based approach is adopted to combine the predictions from individual SVMs that leads to higher prediction performance. The details of different components of *Golgi-predictor* are presented as follows.²

1) FEATURE EXTRACTION STRATEGY

The protein sequences are represented by a fixed-length feature vector that is off course an essential entity of the computation model development. Therefore, the extraction of discriminative information from protein sequences greatly depends upon the underlying feature extraction technique. To build an efficient and accurate computational model, we need to extract discriminative features from protein sequences that could truly imitate the target class. In the current work, split Amph-PseAAC, DPC, BPB, and PSSM based bigram features have been utilized.

Amphiphilic Pseudo Amino Acid Composition (Amph-PseAAC), proposed in [14] and later utilized in [15], [16] not only keeps the record of frequency information but also traces the correlation information associated with physicochemical characteristics of two residues in a protein sequence.

²The implementation code for replicating the experiments presented in this paper is available upon request.

Hydrophobicity and hydrophilicity of amino acids in a protein sequence are the key physicochemical properties that efficiently reflect Amphiphilic features of a protein sequence. Therefore, they are very effective in formulating the sequence-order correlation factors of protein sequences [14]. Hydrophobicity and hydrophilicity based physicochemical patterns can generate Amph-PseAAC features that are capable of discriminating protein sequences accurately. Generally, the Amph-PseAAC method is applied to the entire sequence, and a fixed-length feature descriptor is obtained for each sequence. The length of the feature vector is $20 + 2\lambda$ where 20 is the amino acids frequency in a protein sequence, and 2λ represents the Amphiphilic correlation factors reflecting the distribution of different hydrophobicity and hydrophilicity patterns in a protein sequence. In the current study, the input protein sequence is first split into three parts, where 20 residues are extracted from each of N and C termini, thus resulting in N-part, C-part, and the middle part. Amph-PseAAC is then applied to each of the three sub-sequences separately. The three feature vectors corresponding to three sub-sequences are then concatenated to construct a single feature vector of dimensionality 90-D that represents the entire sequence. This point forward, we will use Amph-PseAAC to refer to the split Amph-PseAAC.

Amino Acid Composition (AAC) is considered among the most primitive feature extraction techniques [17]. However, its capabilities are minimal and are capable of only extracting the occurrence frequency information of amino acid residues in a given protein sequence [18]. AAC may be useful in situations where protein sequences of different categories are diverse enough that could be discriminated on the basis of frequency information only. However, it may fail where the protein sequences belong to different classes and frequency information is not sufficient for discrimination [19]. As a remedy, Di-Peptide Composition (DPC) is used that not only extracts the occurrence frequency information but also exploits the structural information of amino acid residues in a protein sequence. DPC considers amino acid residues in pairs with some specified gap [20]. Dipeptide D_i is calculated using (1).

$$D_i = \frac{\text{Total occurrences of } D_i}{\text{Total number of all possible dipeptides}} \tag{1}$$

Since DPC will look for all possible pairs, it will result into a 400-D feature vector. Inherently, DPC considers the adjacent dipeptides, however, DPC with extended capabilities considering the variable gaps between pair of amino acid residues have also been introduced [9]. Given a protein sequence P , DPC with variable gap can be computed using (2).

$$P_{(i,j)} = \frac{\sum N_{i,j} \times 100}{\sum N_i \sum N_j} \tag{2}$$

where i and j locate the amino acid residues at positions i and $i+1$, respectively. $N_{i,j}$ shows the frequency of i -type residue followed by j -type residue. The term in the denominator of (2) represents the total frequency of type i and type j residues.

In this work, we are utilizing DPC with gap 3 as discussed in [4].

Another feature extraction technique used in this study is Bi-Profile Bayes (BPB) that was first proposed by Shao *et al.* [21] for the classification of methylation sites proteins. It has also been employed for the identification of Protein Puplytation sites [22]. BPB feature extraction technique can be stated briefly as: Let $S = s_1, s_2, s_3, \dots, s_n$ represents a protein sequence where s_j is one amino acid and n is the length of protein sequence. S either belongs to C_1 or C_{-1} where C_1 and C_{-1} represent *cis*-Golgi and *trans*-Golgi proteins, respectively. The feature vector is formulated as given in (3).

$$\vec{P} = (p_1, p_2, \dots, p_n, p_{n+1}, \dots, p_{2n}) \quad (3)$$

where p_1, p_2, \dots, p_n represent the posterior probabilities of each amino acid at each position in protein sequence from category C_1 (*cis*-Golgi proteins) and p_{n+1}, \dots, p_{2n} represent the posterior probabilities of each amino acid at each position in protein sequence from category C_{-1} (*trans*-Golgi proteins) that is known as Bi-Profile of a protein sequence. In the current study, the posterior probabilities are calculated using the occurrence of each amino acid at each position in the training dataset. The dimensionality of the feature vector is 50-D.

Position Specific Scoring Matrix (PSSM) represents the evolutionary information about a protein sequence. PSSMs are calculated using the steps mentioned in [4]. The size of a PSSM matrix depends on the length L of a protein sequence in a given dataset. For example, the size of PSSM will be L -by-20 for a protein sequence of length L . Since each protein sequence is of different length, each PSSM thus formed will also be of different size making it impossible for any predictor to process it. In order to have equal length representations of variable length protein sequences, phenomenon of Bigram PSSM is adopted. Bigram features of each protein sequence are calculated from linear probabilities of PSSM, which results in a Bigram probability matrix B of size 20-by-20 and hence the feature vector is of 400-D. The Bigram probability matrix B can be defined as given in (4).

$$B_{m,n} = \sum_{i=1}^{L-1} p_{i,m} p_{i+1,n}, \text{ where } 1 \leq m \leq 20 \text{ and } 1 \leq n \leq 20 \quad (4)$$

Here L is the length of primary protein sequence, whereas m and n show the 20 amino acids. The Bigram feature vector is constructed according to (5).

$$F = [B_{1,1}, B_{1,2}, \dots, B_{1,20}; B_{2,1}, B_{2,2}, \dots, B_{2,20}; \dots; B_{20,1}, B_{20,2}, \dots, B_{20,20}]^T \quad (5)$$

2) SMOTE OVERSAMPLING

Imbalance data may affect the learning capability of a classifier and degrade its classification performance. Due to the imbalance, the classifier is always biased towards the majority class. However, we should keep in mind that when the

minority class samples are increased synthetically, the classifier learning bias may be shifted towards the minority class. In this work, the utilized dataset is highly imbalanced. The minority class comprises merely 28.6% of the total instances in the dataset as mentioned in section II-A where 87 instances belong to *cis*-Golgi class out of the total 304 instances. As a result, the decisions of the classifier might be more biased towards the majority class samples. Due to the increased affinity towards the majority class, the minority class samples may completely be ignored in the decision-making process, and ultimately the overall performance may be degraded. The higher accuracy values obtained may be the result of higher specificity values that is not an acceptable situation in machine learning and pattern recognition.

A number of different techniques can be adopted to reduce imbalance between class samples in a dataset. Some researchers proposed to either replicate the minority class samples or remove the majority class samples [23] that may either lead to information redundancy or information loss. Replicating the minority class samples simply by duplication introduces redundancy in the minority class that may lead to overfitting, because of the similar regions in the feature space [24]. SMOTE is yet another technique used to handle imbalanced datasets. Contradicting to other oversampling techniques, SMOTE performs its operations in the feature space. Minority class sample is selected, and new synthetic samples are introduced along the line segments that connect some or all the k -nearest neighbors of the minority class under consideration. Synthetic samples thus produced may increase the generalization capability of the classifier. Interested readers may find complete details of SMOTE in [25].

3) SUPPORT VECTOR MACHINE

Support Vector Machine (SVM) [26], [27] is a widely used classification algorithm. Its strengths have been widely exploited in the fields of computational biology, bioinformatics, and pattern recognition [15], [28], [29]. SVM is basically a binary classification algorithm that utilizes the structural risk minimization principle for classification. It exploits the kernel functions to transform the input feature space into a high dimensional feature space for efficient classification. In the current work, Linear-SVM from LIBSVM package [30] is used to develop the prediction model. First, multiple SVMs have been trained, and then their predictions are combined through the majority voting scheme. Grid search approach is followed to optimize the SVM parameter C where the search space specified by (6).

$$\{0.0001 \leq C \leq 20000, \text{ step size} = 1\} \quad (6)$$

III. EVALUATION MEASURES

Three testing protocols including jackknife testing, 10-fold cross-validation testing, and independent dataset testing are employed to analyze the consistency and reliability of the proposed technique [16], [31], [32]. Jackknife testing protocol always gives unique results that make it widely acceptable

testing technique in the bioinformatics community for assessing the performance of their proposed models [32]–[37]. In jackknife testing, for N given samples, $N-1$ samples are used for training and one sample is used for testing [38]. The test is repeatedly performed for each instance in the dataset; due to this reason, the execution time of jackknife testing is higher and depends upon the number of samples in the dataset. In k -fold cross-validation testing, the entire dataset is first split into k mutually exclusive subsets. Then $k-1$ subsets are used for training and one is used for testing [4], [39]. Although similar in mechanics, k -fold cross-validation testing is much faster compared to jackknife testing. Usually, the test is repeatedly performed k times, and the final output is reported as the average of k testing results.

In the Independent dataset testing protocol, the training and testing datasets are entirely different from each other. That is why uncertainty in prediction results is expected to be higher. It is, therefore, crucial for the testing dataset to be much larger so that most of the samples from the training dataset are covered [4]. In situations where training and testing datasets are not available separately, the training dataset is divided into two mutually exclusive subsets that are used as training and testing sets by the classifier. We briefly present descriptions of some standard performance measures including Accuracy, Sensitivity, Specificity, Geometric mean, Mathews Correlation Coefficient, and F-score.

A. ACCURACY

Accuracy [40] measure is used to assess the overall performance of *Golgi-predictor*. It takes into account TP, TN, FP, and FN that represent true positive, true negative, false positive, and false negative protein sequences, respectively. The majority class samples greatly influence accuracy, that is why it may produce a misleading assessment in case of imbalanced data. Equation (7) is used to calculate accuracy.

$$Accuracy = ((TP + TN)/(TP + FN + TN + FP)) \times 100 \quad (7)$$

B. SENSITIVITY/SPECIFICITY

Sensitivity, as described in [40], [41], assesses the actual percentage of correctly predicted samples from positive class whereas specificity [41], [42] examines the actual fraction of negative samples that are correctly predicted. Equations (8) and (9) are used to mathematically express sensitivity and specificity.

$$Sensitivity = (TP/(TP + FN)) \times 100 \quad (8)$$

$$Specificity = (TN/(TN + FP)) \times 100 \quad (9)$$

C. G-MEAN

Geometric mean (G-mean) [43] is the performance measure that considers both the sensitivity and specificity in order to show the balance of classifier over the majority as well as minority classes. G-mean can be calculated using (10).

$$G - mean = \sqrt{Sensitivity \times Specificity} \quad (10)$$

D. MATHEWS CORRELATION COEFFICIENT

Mathews Correlation Coefficient (MCC) is a statistical metric that can show a confusion matrix as a scalar value [42]. Comparatively, it is less influenced by imbalanced data. It is therefore recognized as a stable performance metric among others. MCC takes into account all the entries of a confusion matrix including TP, FP, TN, and FN in its calculations. MCC produces its output in the range of -1 and $+1$ where the former is returned for inverse predictions and the later is for perfect predictions whereas 0 is returned for average random predictions. MCC is calculated using (11).

$$MCC = (TP \times TN - FP \times FN) / \sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)} \quad (11)$$

E. F-SCORE

F-score is a measure of balance between precision and recall [44], [45]. It is calculated as harmonic mean of precision and recall. F-score is null in case either of the precision or recall is null. The output of F-score varies between 0 and 1 where 0 highlights poor performance of the model on positive class and 1 indicates better performance of the model on positive class. The values between 0 and 1 exclusively identify trade-off between precision and recall. F-score is given by (12).

$$F - score = 2 \times ((precision \times recall)/(precision + recall)) \quad (12)$$

where precision and recall are calculated using (13) and (14), respectively.

$$precision = TP/(TP + FP) \quad (13)$$

$$recall = TP/(TP + FN) \quad (14)$$

Here, precision [42] is the number of true positives divided by the sum of true positives and false positives predicted by the classifier whereas recall [11] is the number of true positives divided by the sum of all positives actually present in the positive class.

IV. RESULTS AND DISCUSSION

Performance of Linear-SVM has been assessed using different feature spaces with and without SMOTE oversampling technique. The performance is reported using Accuracy (Acc), Sensitivity (Sn), Specificity (Sp), Geometric mean (G-mean), MCC, and F-score. Results are obtained through 10-fold cross-validation testing, jackknife testing, and independent dataset testing. All the simulations are performed in MATLAB using LIBSVM package [30].

A. ANALYSIS OF INDIVIDUAL FEATURES WITH AND WITHOUT OVERSAMPLING

The performance of Linear-SVM using various individual feature spaces without oversampling is presented in Table 1. Linear-SVM yielded the highest accuracy value of 86.1% for

TABLE 1. Performance accuracies of Linear-SVM using various individual feature spaces without oversampling.

Feature	Testing	Acc	Sn	Sp	G-mean	MCC	F-Score
BPB	10-fold	86.1	80.4	88.4	84.3	0.69	0.80
	Jackknife	81.9	74.2	84.9	79.3	0.59	0.74
	Independent	100	100	100	100	1	1
DPC	10-fold	67.7	48.8	75.3	60.6	0.24	0.51
	Jackknife	61.8	41.8	69.8	54.0	0.11	0.43
	Independent	82.8	87.5	81.6	84.4	0.63	0.75
Amph – PseAAC	10-fold	69.7	46.9	78.8	60.7	0.27	0.50
	Jackknife	65.4	41.9	74.8	55.9	0.17	0.44
	Independent	75	45.1	82.5	60.9	0.28	0.47
Bigram	10-fold	75.6	55.0	83.9	67.9	0.42	0.60
	Jackknife	74.0	53.6	82.1	66.3	0.38	0.58
	Independent	89.0	67.6	94.5	79.9	0.70	0.76

¹ BPB for Bi-Profile Bayes² DPC for Di-Peptide Composition³ Amph-PseAAC for Amphiphilic Pseudo Amino Acid Composition⁴ Bigram for PSSM based bigram features

BPB features using 10-fold cross-validation protocol. The performance of Linear-SVM for the same features is promising using jackknife testing with 81.9% accuracy. However, the accuracy is 100% for same features using independent dataset testing. Since accuracy is considered a poor performance measure in case of imbalance dataset therefore, we consider G-mean, MCC, and F-score values. The G-mean value of 84.3% is highest for BPB features using 10-fold cross-validation. As we also know that sensitivity measures the accuracy for positive class, which is minority class in this case therefore, F-score is a good measure to be discussed here. The higher F-score value of 0.80 indicates that performance of Linear-SVM is better on the positive class. This shows the significance of BPB features that possess discriminative power even in the presence of imbalanced data.

This is evident from the spare diagram in Fig. IV-A where only the BPB feature space shows separability in features. It can be clearly observed that BPB features offer more discrimination capability compared to other feature spaces under consideration.

The performance of Linear-SVM for Amph-PseAAC using independent testing is poor as shown by G-mean value of 60.9%. Similarly, F-score with value 0.47 is an indication of poor performance of Linear-SVM over the positive class from testing dataset. Table 2 presents the performance accuracies of Linear-SVM using individual features with oversampling. The highest accuracy value of 87.3% is achieved for BPB features using 10-fold cross-validation. Through the jackknife testing, the accuracy is 85.9%. The G-mean value of 87.2% for BPB features indicates that performance of the proposed model using 10-fold cross-validation is promising. In addition, the same features performed well using jackknife testing that achieved 85.8% G-mean value. Likewise, F-score value of 0.87 for BPB features shows that performance of the model over positive class is better compared to negative class.

B. ANALYSIS OF HYBRID FEATURES WITH AND WITHOUT OVERSAMPLING

Table 3 highlights the performance of Linear-SVM for hybrid feature spaces without oversampling using 10-fold

TABLE 2. Performance accuracies of Linear-SVM using various individual feature spaces with oversampling.

Feature	Testing	Acc	Sn	Sp	G-mean	MCC	F-Score
BPB	10-fold	87.3	89.0	85.5	87.2	0.74	0.87
	Jackknife	85.9	87.6	84.2	85.8	0.71	0.86
	Independent	100	100	100	100	1	1
DPC	10-fold	79.9	84.6	75.2	79.7	0.60	0.80
	Jackknife	79.4	84.9	74.0	79.2	0.59	0.80
	Independent	82.8	91.4	80.6	85.8	0.65	0.76
Amph – PseAAC	10-fold	74.1	78.1	70.2	74.0	0.48	0.75
	Jackknife	71.4	75.3	67.5	71.2	0.43	0.72
	Independent	75	68.2	76.7	72.3	0.42	0.62
Bigram	10-fold	80.4	87.2	73.6	80.1	0.62	0.81
	Jackknife	78.1	85.5	70.6	77.6	0.57	0.79
	Independent	96.8	94.5	97.4	95.9	0.92	0.94

¹ BPB for Bi-Profile Bayes² DPC for Di-Peptide Composition³ Amph-PseAAC for Amphiphilic Pseudo Amino Acid Composition⁴ Bigram for PSSM based bigram features**TABLE 3. Performance accuracies of Linear-SVM using various hybrid feature spaces without oversampling.**

Feature	Testing	Acc	Sn	Sp	G-mean	MCC	F-Score
BPB + Bigram	10-fold	85.5	79.5	87.9	83.5	0.67	0.79
	Jackknife	82.5	75.7	85.2	80.3	0.61	0.75
	Independent	98.4	95.3	99.2	97.2	0.96	0.97
BPB + DPC	10-fold	82.5	74.6	85.7	79.9	0.60	0.75
	Jackknife	80.2	71.1	83.9	77.2	0.55	0.72
	Independent	98.4	95.3	99.2	97.2	0.96	0.97
BPB + Amph – PseAAC	10-fold	87.8	80.6	90.6	85.4	0.72	0.82
	Jackknife	84.8	77.4	87.8	82.4	0.66	0.78
	Independent	96.8	90.7	98.4	94.4	0.91	0.94
hybrid – all	10-fold	82.2	71.0	86.7	78.4	0.59	0.74
	Jackknife	79.9	69.2	84.2	76.3	0.54	0.71
	Independent	96.8	90.7	98.4	94.4	0.91	0.94

¹ BPB for Bi-Profile Bayes² DPC for Di-Peptide Composition³ Amph-PseAAC for Amphiphilic Pseudo Amino Acid Composition⁴ Bigram for PSSM based bigram features⁵ hybrid-all for BPB + Bigram + DPC + Amph-PseAAC

cross-validation, jackknife testing, and independent dataset testing. The 10-fold cross-validation accuracy value of 87.8%, achieved for the hybrid feature space of BPB and Amph-PseAAC, is highest among other hybrid models. Likewise, the accuracy obtained using jackknife testing for the same hybrid model is also in better range that is 84.8%. All the accuracy values using 10-fold cross-validation are above the percentage of 80 that show the importance of hybrid models compared to their individual counterparts where only BPB features achieved accuracy value over 80% as shown in Table 1.

Similarly, the performance of Linear-SVM for hybrid features with oversampling is shown in Table 4. The highest accuracy value achieved by Linear-SVM for the hybrid model of BPB and DPC is 91.4% using 10-fold cross-validation. The accuracy using jackknife testing is 91.2% for the same hybrid model. The G-mean value of 91.4% obtained using 10-fold cross-validation testing for the hybrid features of BPB and DPC with oversampling is highest. Overall, the performance of Linear-SVM for the hybrid models with SMOTE oversampling is higher compared to the individual feature spaces.

The spare diagram of hybrid feature spaces with and without oversampling are depicted in Fig. 3 that highlights the visual separability among different instances in the

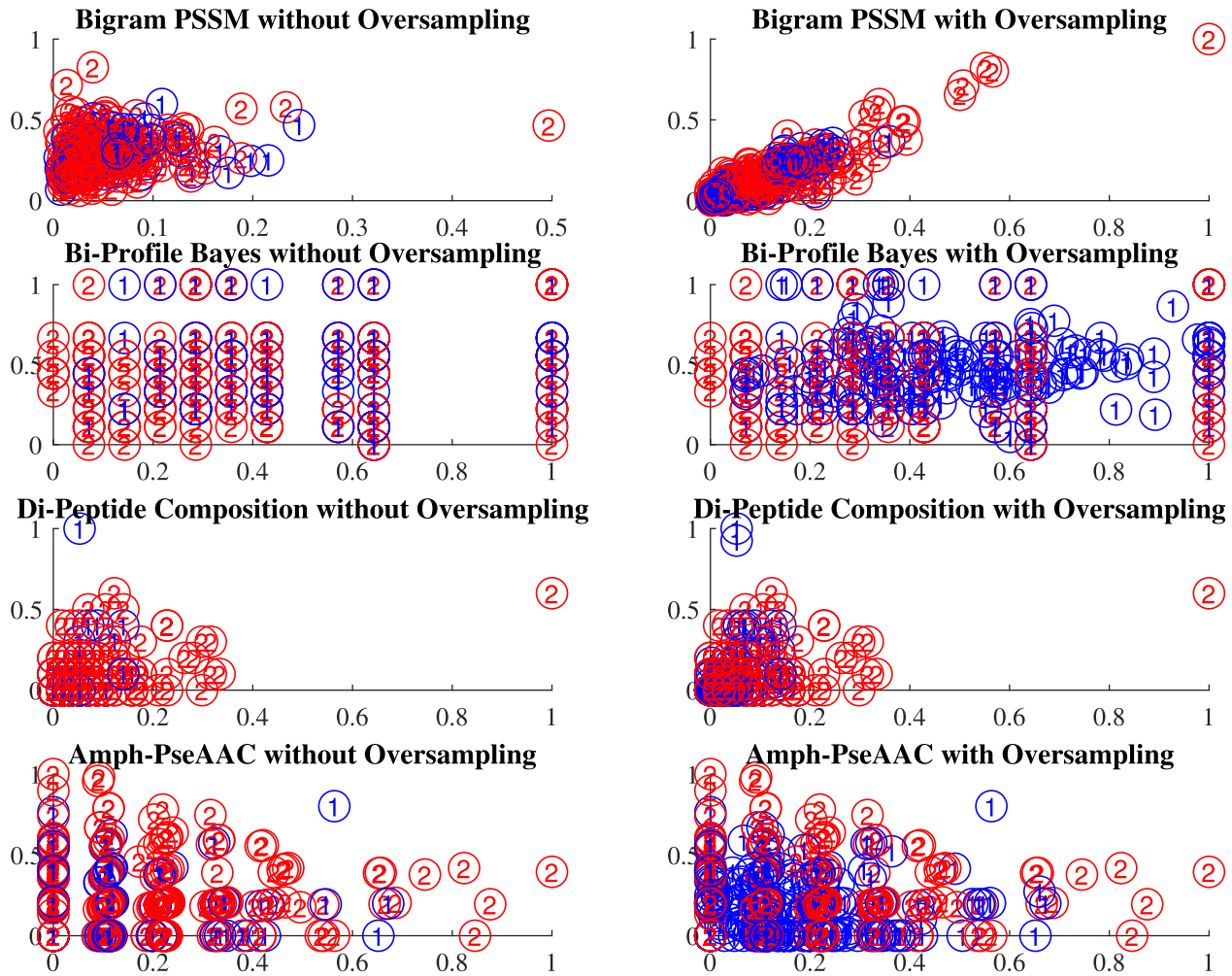


FIGURE 2. Individual feature spaces with and without oversampling.

TABLE 4. Performance accuracies of Linear-SVM using various hybrid feature spaces with oversampling.

Feature	Testing	Acc	Sn	Sp	G-mean	MCC	F-Score
BPB + Bigram	10-fold	90.7	93.0	88.4	90.6	0.81	0.90
	Jackknife	88.4	91.4	85.4	88.3	0.77	0.88
	Independent	100	100	100	100	1	1
BPB + DPC	10-fold	91.4	94.1	88.8	91.4	0.83	0.91
	Jackknife	91.2	93.7	88.7	91.1	0.82	0.91
	Independent	96.8	90.7	98.4	94.4	0.91	0.94
BPB + Amph - PseAAC	10-fold	90.3	91.2	89.4	90.2	0.80	0.90
	Jackknife	89.1	90.4	87.9	89.1	0.78	0.89
	Independent	96.8	90.7	98.4	94.4	0.91	0.94
hybrid - all	10-fold	92.1	94.0	90.3	92.1	0.84	0.92
	Jackknife	91.2	93.3	89.1	91.1	0.82	0.91
	Independent	98.4	95.3	99.2	97.2	0.96	0.97

¹ BPB for Bi-Profile Bayes
² DPC for Di-Peptide Composition
³ Amph-PseAAC for Amphiphilic Pseudo Amino Acid Composition
⁴ Bigram for PSSM based bigram features
⁵ hybrid-all for BPB + Bigram + DPC + Amph-PseAAC

constructed feature space. It is evident that BPB based hybrid models possess higher discrimination power compared to other individual feature spaces. This enhanced the learning capability of Linear-SVM that ultimately improved the ensemble performance. In the next section, we are presenting

TABLE 5. Performance of Linear-SVM ensemble for hybrid feature spaces with oversampling.

	Testing	Acc	Sn	Sp	G-mean	MCC	F-Score
Ensemble	10-fold	97.6	98.8	96.5	97.6	0.95	0.97
	Jackknife	96.5	97.8	95.2	96.4	0.93	0.96
	Independent	98.4	95.3	99.2	97.2	0.96	0.97

the performance of ensemble classification comprising of multiple Linear Support Vector Machines.

C. ANALYSIS OF ENSEMBLE CLASSIFICATION

This study considers the first three hybrid models from Table 4 for constructing *Golgi-predictor*: an ensemble of Linear Support Vector Machines. The achieved results are highlighted in Table 5.

The ensemble accuracy is highest among the outputs of individual SVM classifications. That is why ensemble-based predictions are always important because they provide reliable and more accurate results in any problem domain.

Through the 10-fold cross-validation protocol, *Golgi-predictor* has achieved 97.6% ensemble accuracy that is 5.5%

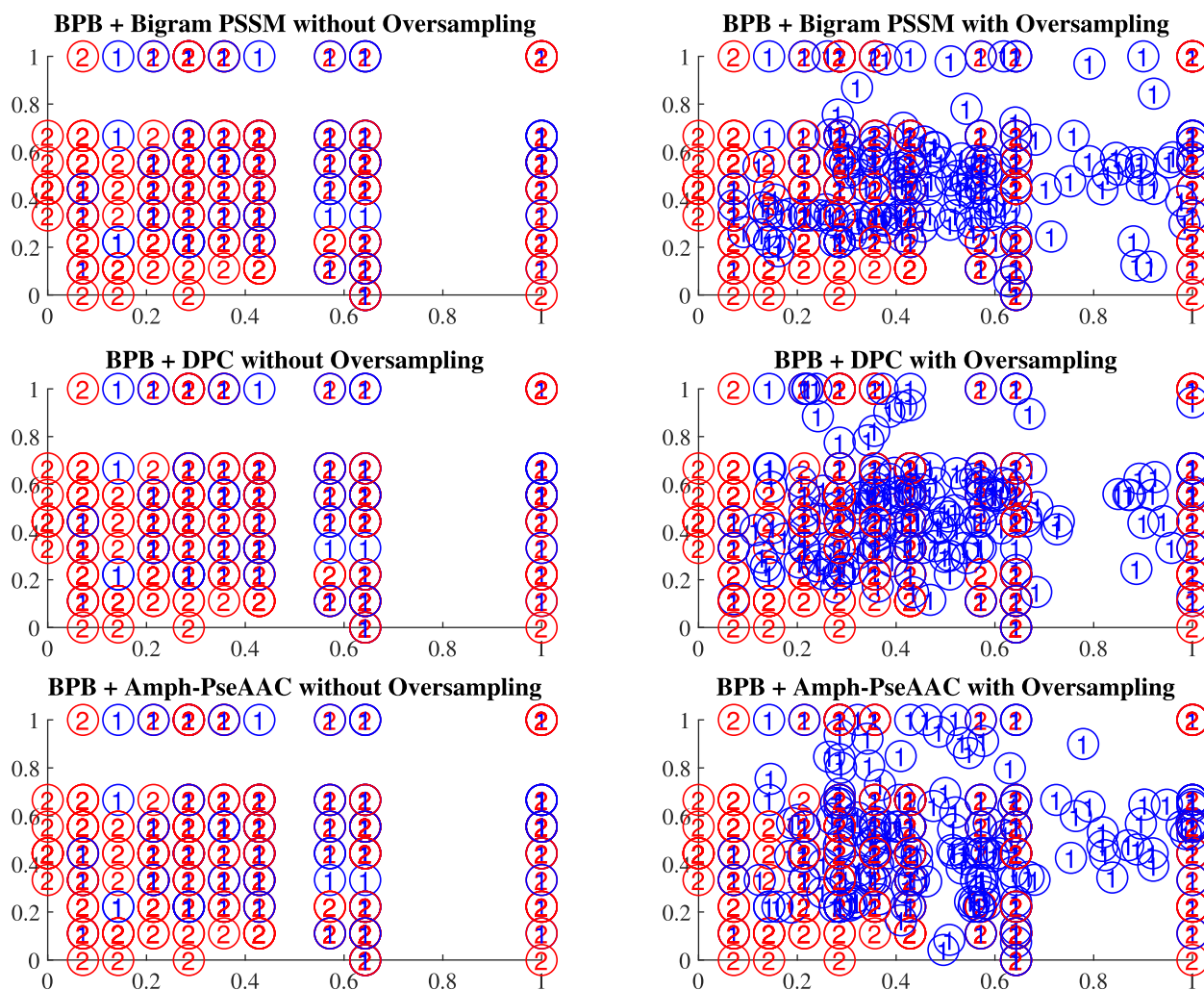


FIGURE 3. Hybrid feature spaces with and without oversampling.

higher than the accuracy value of stand-alone Linear-SVM on hybrid-all features as shown in Table 4. Similarly, jackknife testing accuracy is 5.3% higher that shows the significance of ensemble method. From Table 5, the G-mean value of 97.6% is also promising. The F-score value of 0.97 shows performance enhancement on the positive class. Similarly, through the jackknife testing, the obtained performance accuracy, G-mean, and F-score values are 96.5%, 96.4%, and 0.96, respectively. Moreover, through the independent dataset testing, *Golgi-predictor* achieved 98.4% accuracy, 97.2% G-mean, and 0.97 F-score values. Improved predictions are due to the differences in classifier’s learning capabilities, which are different with different feature extraction strategies. Each feature extraction strategy extracts different information from protein sequences that affects the learning of classifiers differently. When the predictions are combined, the output results are improved.

V. PERFORMANCE COMPARISON

Comparative analysis of the proposed technique against the existing techniques is presented in Table 6. The comparison

is provided based on accuracy, sensitivity, specificity, and MCC. In order to make the comparison more interesting, we also calculated G-mean for the other methods, as shown in Table 6. F-score values are provided for our work only.

Ding et al. [7] have computed results only through jackknife and independent dataset testing. In another work, Ding et al. [8] reported their results only through jackknife testing protocol. Yang et al. [9] reported the performance predictions of their system using 10-fold cross-validation, jackknife, and independent dataset testing protocols. They achieved accuracy values of 90.1%, 88.5%, and 93.8%, respectively, for the three testing protocols. The work published by Ahmad et al. [4] have reported the highest performance accuracies compared to the other mentioned techniques. In another work by Ahmad and Hayat [10], MVFS based method has achieved accuracy values of 95.8%, 98.1%, and 94.0% through 10-fold cross-validation, jackknife, and independent dataset testing protocols, respectively.

In the current paper, we proposed a Linear-SVM based ensemble that outperformed the existing techniques in terms of accuracy values through 10-fold cross-validation and

TABLE 6. Performance comparison with the existing techniques.

	Method	Acc	Sn	Sp	G-mean	MCC	F-Score
10-fold	Ding et al. [7]	-	-	-	-	-	-
	Ding et al. [8]	-	-	-	-	-	-
	Yang et al. [9]	90.1	90.8	89.4	90.0	0.80	-
	Ahmad et al. [4]	94.9	97.2	92.6	94.8	0.90	-
	Ahmad and Hayat [10]	95.8	97.2	94.4	-	0.92	-
	Cui et al. [11]	80.0	-	-	-	-	-
	Zhou et al. [13]	-	-	-	-	-	-
	<i>Golgi-predictor</i>	97.6	98.8	96.5	97.6	0.95	0.97
Jackknife	Ding et al. [7]	74.7	69.6	79.6	74.4	0.51	-
	Ding et al. [8]	85.4	73.8	90.5	81.7	0.65	-
	Yang et al. [9]	88.5	88.9	88.0	88.4	0.76	-
	Ahmad et al. [4]	94.9	97.2	92.6	94.8	0.90	-
	Ahmad and Hayat [10]	98.1	98.6	97.7	-	0.96	-
	Cui et al. [11]	-	-	-	-	-	-
	Zhou et al. [13]	92.1	89.7	94.4	-	0.84	-
	<i>Golgi-predictor</i>	96.5	97.8	95.2	96.4	0.93	0.96
Independent	Ding et al. [7]	-	-	-	-	-	-
	Ding et al. [8]	-	-	-	-	-	-
	Yang et al. [9]	93.8	92.3	94.1	93.1	0.82	-
	Ahmad et al. [4]	94.8	94.0	93.9	93.9	0.86	-
	Ahmad and Hayat [10]	94.0	81.4	96.8	-	0.84	-
	Cui et al. [11]	-	-	-	-	-	-
	Zhou et al. [13]	86.5	98.1	75.0	-	0.75	-
	<i>Golgi-predictor</i>	98.4	95.3	99.2	97.2	0.96	0.97

independent dataset testing protocols. However, it shows slightly lower performance in case of jackknife testing against the work of Ahmad and Hayat [10]. In case of 10-fold cross-validation, our proposed model achieved 97.6% accuracy and outperformed the existing techniques by 1.8% beating the previous highest accuracy reported by Ahmad and Hayat [10]. The G-mean, which is considered a balance measure in the presence of imbalanced data, has produced better values compared to other methods. Similarly, through jackknife testing, our proposed model obtained 96.5% accuracy that is 1.6% better than the previous highest reported value. The G-mean value, in this case, is also highest. Our proposed model also outperformed the existing state-of-the-art methods using independent dataset testing. This work achieved 98.6% accuracy and proved the efficiency of our proposed ensemble based technique. Despite the majority voting based ensemble, the hybrid-all feature space shown in Table 4 has also outperformed the reported performance of Yang et al. [9]. This shows the significance of our proposed models.

VI. CONCLUSION

The proposed *Golgi-predictor* is a novel and reliable computational model for the identification of sub-Golgi proteins. We proposed the utilization of hybrid models in conjunction with SMOTE oversampling technique that are exploited by Linear-SVM for classification. The final output is obtained by combining the predictions of multiple SVMs using the majority voting approach.

The proposed *Golgi-predictor* has been validated for its accuracy, reliability, and efficiency using three standard testing protocols, including 10-fold cross-validation, jackknife testing, and independent dataset testing. Simulation results demonstrated significant performance in classifying the sub-Golgi proteins. The hybrid models exploited by individual SVMs have shown marginal performance compared to existing state-of-the-art methods. However, the majority

voting based ensemble of SVMs has boosted the accuracy to higher levels and outperformed the existing methods. Results obtained through different testing protocols have proved that the proposed *Golgi-predictor* is reliable and efficient that could be used by researchers for drug development and diagnostic purposes. Since the predictions are based on hybrid models as well as ensemble classification, it would be worthy for future research to identify critical characteristics of different feature extraction techniques specific to individual protein sequences. This would help in identifying case-by-case characteristics of individuals. We further add that the overall accuracy may be improved by removing irrelevant features using some feature selection technique that is also helpful in reducing the computational complexity of a model. Some of the recent works [46], [47] have demonstrated the effectiveness of feature selection. In future, we intend to enhance the prediction capability of our model using feature selection techniques.

REFERENCES

- [1] J. E. Darnell, H. F. Lodish, and D. Baltimore, *Molecular Cell Biology*, 2nd ed. New York, NY, USA: Scientific American Books Inc., 1990.
- [2] T. Pollard, W. Earnshaw, J. Lippincott-Schwartz, and G. Johnson, *Cell Biology*, 3rd ed. Amsterdam, The Netherlands: Elsevier, 2017.
- [3] G. M. Cooper and R. E. Hausman, *The Cell*. Washington, DC, USA: ASM Press, 2000.
- [4] J. Ahmad, F. Javed, and M. Hayat, "Intelligent computational model for classification of sub-golgi protein using oversampling and Fisher feature selection methods," *Artif. Intell. Med.*, vol. 78, pp. 14–22, May 2017.
- [5] W.-R. Qiu, S.-Y. Jiang, Z.-C. Xu, X. Xiao, and K.-C. Chou, "iRNAm5C-PseDNC: Identifying m⁵-methylcytosine sites by incorporating physical-chemical properties into pseudo dinucleotide composition," *Oncotarget*, vol. 8, no. 25, p. 41178, 2017.
- [6] L. Nanni and S. Brahnam, "Multi-label classifier based on histogram of gradients for predicting the anatomical therapeutic chemical class/classes of a given compound," *Bioinformatics*, vol. 33, no. 18, pp. 2837–2841, Sep. 2017.
- [7] H. Ding, L. Liu, F.-B. Guo, J. Huang, and H. Lin, "Identify golgi protein types with modified Mahalanobis discriminant algorithm and pseudo amino acid composition," *Protein Peptide Lett.*, vol. 18, no. 1, pp. 58–63, Jan. 2011.
- [8] H. Ding, S.-H. Guo, E.-Z. Deng, L.-F. Yuan, F.-B. Guo, J. Huang, N. Rao, W. Chen, and H. Lin, "Prediction of golgi-resident protein types by using feature selection technique," *Chemometric Intell. Lab. Syst.*, vol. 124, pp. 9–13, May 2013.
- [9] R. Yang, C. Zhang, R. Gao, and L. Zhang, "A novel feature extraction method with feature selection to identify golgi-resident protein types from imbalanced data," *Int. J. Mol. Sci.*, vol. 17, no. 2, p. 218, Feb. 2016.
- [10] J. Ahmad and M. Hayat, "MFSC: Multi-voting based feature selection for classification of golgi proteins by adopting the general form of Chou's PseAAC components," *J. Theor. Biol.*, vol. 463, pp. 99–109, Feb. 2019.
- [11] Q. Cui, Y. Cao, W. Bao, B. Yang, and Y. Chen, "SubRF_Seq: Identification of sub-golgi protein types with random forest with partial sequence information," *Sci. Program.*, vol. 2020, pp. 1–7, Jul. 2020.
- [12] Z. Chen, N. He, Y. Huang, W. T. Qin, X. Liu, and L. Li, "Integration of a deep learning classifier with a random forest approach for predicting malonylation sites," *Genomics, Proteomics Bioinf.*, vol. 16, no. 6, pp. 451–459, Dec. 2018.
- [13] H. Zhou, C. Chen, M. Wang, Q. Ma, and B. Yu, "Predicting golgi-resident protein types using conditional covariance minimization with XGBoost based on multiple features fusion," *IEEE Access*, vol. 7, pp. 144154–144164, 2019.
- [14] K.-C. Chou, "Using amphiphilic pseudo amino acid composition to predict enzyme subfamily classes," *Bioinformatics*, vol. 21, no. 1, pp. 10–19, Jan. 2005.

- [15] X.-B. Zhou, C. Chen, Z.-C. Li, and X.-Y. Zou, "Using Chou's amphiphilic pseudo-amino acid composition and support vector machine for prediction of enzyme subfamily classes," *J. Theor. Biol.*, vol. 248, no. 3, pp. 546–551, Oct. 2007.
- [16] H. Ding, L. Luo, and H. Lin, "Prediction of cell wall lytic enzymes using chous amphiphilic pseudo amino acid composition," *Protein Peptide Lett.*, vol. 16, no. 4, pp. 351–355, Apr. 2009.
- [17] J. Rahman, N. I. Mondal, K. B. Islam, and A. M. Hasan, "Feature fusion based SVM classifier for protein subcellular localization prediction," *J. Integrative Bioinf.*, vol. 13, no. 1, pp. 23–33, Mar. 2016.
- [18] M. Bhasin and G. P. Raghava, "Classification of nuclear receptors based on amino acid composition and dipeptide composition," *J. Biol. Chem.*, vol. 279, no. 22, pp. 23262–23266, 2004.
- [19] G.-H. Liu, H.-B. Shen, and D.-J. Yu, "Prediction of protein–protein interaction sites with machine-learning-based data-cleaning and post-filtering procedures," *J. Membrane Biol.*, vol. 249, nos. 1–2, pp. 141–153, Apr. 2016.
- [20] J. X. Guo and N. N. Rao, "The influence of dipeptide composition on protein folding rates," in *Advanced Materials Research*, vol. 378. Kapellweg 8, Switzerland: Trans Tech Publications Ltd., 2012, pp. 157–160.
- [21] 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, Mar. 2009.
- [22] X. Zhao, J. Zhang, Q. Ning, P. Sun, Z. Ma, and M. Yin, "Identification of protein pupylation sites using bi-profile Bayes feature extraction and ensemble learning," *Math. Problems Eng.*, vol. 2013, pp. 1–7, Oct. 2013.
- [23] M. J. Pazzani, C. J. Merz, P. M. Murphy, K. M. Ali, T. Hume, and C. A. Brunk, "Reducing misclassification costs," in *Proc. 11th Int. Conf. Mach. Learn.* San Francisco, CA, USA: Morgan Kaufmann, 1994, pp. 217–225.
- [24] C. X. Ling and C. Li, "Data mining for direct marketing: Problems and solutions," *Knowl. Discovery Databases*, vol. 98, pp. 73–79, Aug. 1998.
- [25] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: Synthetic minority over-sampling technique," *J. Artif. Intell. Res.*, vol. 16, pp. 321–357, Jun. 2002.
- [26] V. N. Vapnik, "An overview of statistical learning theory," *IEEE Trans. Neural Netw.*, vol. 10, no. 5, pp. 988–999, Sep. 1999.
- [27] S. R. Gunn, "Support vector machines for classification and regression," *ISIS Tech. Rep.*, vol. 14, no. 1, pp. 5–16, 1998.
- [28] S. Li, J. T. Kwok, H. Zhu, and Y. Wang, "Texture classification using the support vector machines," *Pattern Recognit.*, vol. 36, no. 12, pp. 2883–2893, Dec. 2003.
- [29] J.-Y. Shi, S.-W. Zhang, Q. Pan, Y.-M. Cheng, and J. Xie, "Prediction of protein subcellular localization by support vector machines using multi-scale energy and pseudo amino acid composition," *Amino Acids*, vol. 33, no. 1, pp. 69–74, Jul. 2007.
- [30] C.-C. Chang and C.-J. Lin, "LIBSVM: A library for support vector machines," *ACM Trans. Intell. Syst. Technol.*, vol. 2, no. 3, pp. 27:1–27:27, 2011.
- [31] C.-J. Zhang, H. Tang, W.-C. Li, H. Lin, W. Chen, and K.-C. Chou, "iOriHuman: Identify human origin of replication by incorporating dinucleotide physicochemical properties into pseudo nucleotide composition," *Oncotarget*, vol. 7, no. 43, p. 69783–69793, 2016.
- [32] H. Ding and D. Li, "Identification of mitochondrial proteins of malaria parasite using analysis of variance," *Amino Acids*, vol. 47, no. 2, pp. 329–333, Feb. 2015.
- [33] H. Lin and W. Chen, "Prediction of thermophilic proteins using feature selection technique," *J. Microbiological Methods*, vol. 84, no. 1, pp. 67–70, Jan. 2011.
- [34] L. Dai, C. B. Renner, and P. S. Doyle, "The polymer physics of single DNA confined in nanochannels," *Adv. Colloid Interface Sci.*, vol. 232, pp. 80–100, Jun. 2016.
- [35] H. Ding, P.-M. Feng, W. Chen, and H. Lin, "Identification of bacteriophage virion proteins by the ANOVA feature selection and analysis," *Mol. BioSyst.*, vol. 10, no. 8, pp. 2229–2235, 2014.
- [36] K.-C. Chou and C.-T. Zhang, "Prediction of protein structural classes," *Crit. Rev. Biochemistry Mol. Biol.*, vol. 30, no. 4, pp. 275–349, 1995.
- [37] X.-X. Chen, H. Tang, W.-C. Li, H. Wu, W. Chen, H. Ding, and H. Lin, "Identification of bacterial cell wall lyases via pseudo amino acid composition," *BioMed Res. Int.*, vol. 2016, pp. 1–8, Oct. 2016.
- [38] Y.-D. Cai, X.-J. Liu, X.-B. Xu, and K.-C. Chou, "Prediction of protein structural classes by support vector machines," *Comput. Chem.*, vol. 26, no. 3, pp. 293–296, Feb. 2002.
- [39] R. Kohavi, "A study of cross-validation and bootstrap for accuracy estimation and model selection," in *Proc. IJcai*, vol. 14, Montreal, QC, Canada, 1995, pp. 1137–1145.
- [40] C.-Q. Feng, Z.-Y. Zhang, X.-J. Zhu, Y. Lin, W. Chen, H. Tang, and H. Lin, "ITerm-PseKNC: A sequence-based tool for predicting bacterial transcriptional terminators," *Bioinformatics*, vol. 35, no. 9, pp. 1469–1477, May 2019.
- [41] R. Su, H. Wu, B. Xu, X. Liu, and L. Wei, "Developing a multi-dose computational model for drug-induced hepatotoxicity prediction based on toxicogenomics data," *IEEE/ACM Trans. Comput. Biol. Bioinf.*, vol. 16, no. 4, pp. 1231–1239, Jul. 2019.
- [42] S. Zhang, T. Zhang, and C. Liu, "Prediction of apoptosis protein subcellular localization via heterogeneous features and hierarchical extreme learning machine," *SAR QSAR Environ. Res.*, vol. 30, no. 3, pp. 209–228, Mar. 2019.
- [43] M. Kubat and S. Matwin, "Addressing the curse of imbalanced training sets: One-sided selection," in *Proc. 4th Int. Conf. Mach. Learn.*, vol. 97, Nashville, TN, USA, Jul. 1997, pp. 179–186.
- [44] G. H. Nguyen, A. Bouzerdoum, and S. L. Phung, "Learning pattern classification tasks with imbalanced data sets," in *Pattern Recognition*. Rijeka, Croatia: InTech, 2009.
- [45] M. Bekkar, H. K. Djemaa, and T. A. Alitouche, "Evaluation measures for models assessment over imbalanced datasets," *J. Inf. Eng. Appl.*, vol. 3, no. 10, pp. 27–38, 2013.
- [46] L. Ali, C. Zhu, M. Zhou, and Y. Liu, "Early diagnosis of Parkinson's disease from multiple voice recordings by simultaneous sample and feature selection," *Expert Syst. Appl.*, vol. 137, pp. 22–28, Dec. 2019.
- [47] L. Ali, C. Zhu, Z. Zhang, and Y. Liu, "Automated detection of Parkinson's disease based on multiple types of sustained phonations using linear discriminant analysis and genetically optimized neural network," *IEEE J. Transl. Eng. Health Med.*, vol. 7, pp. 1–10, 2019.



MUHAMMAD TAHIR (Senior Member, IEEE) received the M.S. degree in computer science from the National University of Computer and Emerging Sciences, Islamabad, in 2009, and the Ph.D. degree in computer science from the Pakistan Institute of Engineering and Applied Sciences, Islamabad, in 2014. He was with the City University of Science and Information Technology, Peshawar, as an Assistant Professor, from 2014 to 2015. Since 2015, he has been with the Department of Computer Science, College of Computing and Informatics, Saudi Electronic University, as an Assistant Professor. His research interests include machine learning, pattern recognition, image processing, deep learning, and bioinformatics.



FAZLULLAH KHAN (Senior Member, IEEE) is currently a Researcher with the Institute of Social and Economic Research, Duy Tan University, Da Nang, Vietnam. His research has been published in the IEEE TRANSACTIONS ON INDUSTRIAL INFORMATICS, the IEEE TRANSACTION ON VEHICULAR TECHNOLOGY, IEEE ACCESS, *Future Generations Computer Systems* (Elsevier), *Journal of Network and Computer Applications* (Elsevier), *Computers and Electrical Engineering* (Elsevier), and *Mobile Networks and Applications* (Springer). His research interests include security and privacy, the Internet of Things, machine learning, software-defined networks, fog computing, and big data analytics. He served as a Guest Editor for IEEE Access Journal, *Multimedia Technology and Applications* (Springer), and *Mobile Networks and Applications* (Springer).



MOHAMMAD KHALID IMAM RAHMANI (Member, IEEE) received the B.Sc.Engg. degree in computer engineering from A. M. U., Aligarh, the M.Tech. degree in computer engineering from M. D. U., Rohtak, and the Ph.D. degree in digital image retrieval algorithms. He has more than 20 years of professional experience. He was with the Head of the Department of Computer Science and Engineering. He was also with the Accredited College. He has been a Chief Mentor,

a Convenor/Coordinator, and the Editor with the National Conferences, such as RTCSIT in 2013, RTCSIT in 2012, and RTEEE in 2012. He is currently an Assistant Professor with the Department of Computer Science, College of Computing and Informatics, Saudi Electronic University. He is having more than 30 research papers in reputed journals and conferences. His research interests include the IoT, information security, machine learning, content-based image retrieval, and distributed computing, algorithms, and data structures.



VINH TRUONG HOANG (Member, IEEE) graduated from the University of Montpellier and the University of the Littoral Opal Coast, France. He is currently an Assistant Professor and the Head of the Image Processing and Computer Graphics Department, Faculty of Computer Science, Ho Chi Minh City Open University, Vietnam. His research interests include texture classification, biometric recognition, semi-supervised learning, artificial intelligence in healthcare, climate change, object recognition/detection/tracking, optical character recognition, face analysis, kinship verification, and plant identification.

• • •