where, $K(x^i, x^t) = exp(-\gamma\,||x^i - x^t||^2)$

$y_i \in (+1, -1)$,

$\gamma = \frac{1}{2\sigma^2}$

σ is the width of the function,

α_i is the lagrange multipliers.

The subcellular location of target protein P_t will be predicted as:

$$Loc(x^t) = \bigcup_{j=1}^{d} \{j : S_j(x^t) > 0\} \tag{14}$$

If $Loc(x^t) = \emptyset$, then the number of subcellular locations is set to one and the location is given by:

$$Loc(x^t) = arg_{j}^{d}{}_{=1} min f_j \tag{15}$$

where, f_j is the functional value returned by each SVM classifier. This idea of avoiding zero prediction is adapted from Wan et al. work [9].

4 Result and Discussion

In statistical prediction, two of the most crucial issues are what metrics should be used and what kind of test strategy should be followed. For test strategy selection, we adopt the technique used in [19, 20], and that technique was also followed by other researchers, mentioned in the both works.

4.1 Metrics in Multi-label Systems

In biological context, a protein may exist in more than one location, so, our prediction problem is a multi-label problem. To evaluate the anticipated performance, here we adopted two well defined and widely used metrics from the work [9]: actual accuracy and locative accuracy.

Actual accuracy (AA) is defined as the exact match of classifier's predicted labels with the actual labels of a target protein.

$$AA = \frac{\sum_{i=1}^{N} \Delta[M(P_i),\ L\ (P_i)]}{N_{AA}} * 100 \tag{16}$$

Where, $\Delta[M(P_i), L\ (P_i)] = \begin{cases} 1, & if\ \ M(P_i) \equiv L(P_i) \\ 0, & otherwise \end{cases}$

$L(P_i)$ represents true label set for i^{th} protein

$M(P_i)$ represents predicted label set for i^{th} protein

N_{AA} is total number of unique proteins

On the other hand, if the predicted label $M(P_i)$ of target protein P_i matches with any label of the true label set $L(P_i)$ then the accuracy is considered as *Locative Accuracy*. It can be defined as-

$$LA = \frac{\sum_{i=1}^{N} \Delta|M(P_i) \cap L\ (P_i)|}{N_{LA}} * 100 \tag{17}$$

N_{LA} is total number of locative proteins

4.2 Cross-Validation and Success Rate

In most of the statistical prediction problem, for examining the predictor's strength, independent dataset test, K-fold cross validation and jackknife test are widely used. As mentioned in both [19, 20], only the jackknife test always provide a unique result for a given benchmark dataset but the computational cost is very high compared to the other techniques.

In this study, however, to reduce the computational cost, we adopted the 8-fold cross-validation method, as done by many researchers with support vector machine. In this case, in each pass, seven portions were used as training data, remaining set was used as test data and this process repeats until all the proteins goes to test set. 8-fold was selected as there are only 8 proteins at nucleoid location and more than 8 proteins in other locations in our Gram-negative bacterial dataset. We have ensured that each fold contains at least one protein sequence of each class.

As SVM with RBF kernel was our prediction engine and parameter selection may bias the predictor hence, we need to select optimal combination of γ and regularization coefficient c for obtaining highest accuracy. For achieving better accuracy, we select the parameters from $\gamma = \{2^{-8}, 2^{-7}, \ldots, 2^{0}, \ldots, 2^{8}\}$ and $c = \{2^{-2}, 2^{-1}, 2^{0}, 2^{1}, 2^{2}\}$. Since AAIDPAAC feature outperforms other features with SVM, the corresponding search space in Gram Negative dataset are shown in Figure 3.

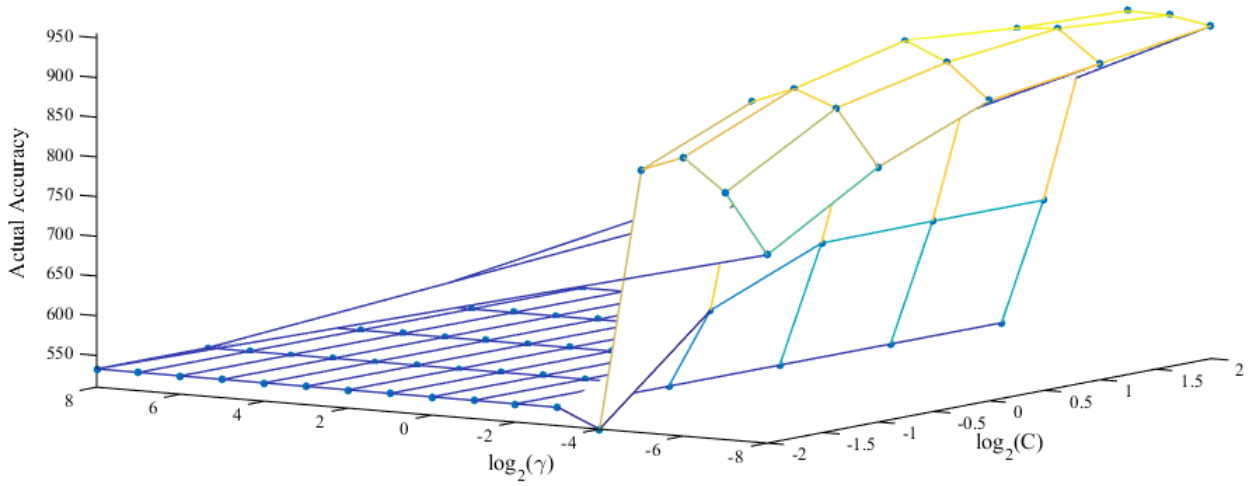

Figure 3: Tuning SVM with AAIDPAAC in Gram Negative Dataset

Here, the experimental result of single features and fused feature sets is given bellow:

Table 3: Performance for different features using SVM

Features	Actual Accuracy (AA) (%)	Locative Accuracy (LA) (%)
PseAAC	65.04	68.5
PPM	61.3	68.8
AAID	65.4	69.25
AAIDPAAC	68.7	70.7
PPMPAAC	67.99	71.05

It can be seen from the following graph Figure 4 that the best actual accuracy 68.7% is achieved by AAIDPAAC which is 3.5%, 7% and 3% greater than PseAAC, PPM and AAID respectively. AAIDPAAC as the fuse form of AAID and PseAAC feature representation has

achieved this best performance for $\gamma = 2^{-6}$, $c = 2^{0}$. The second highest actual accuracy 67.99% is also achieved by feature fusion representation PPMPAAC.

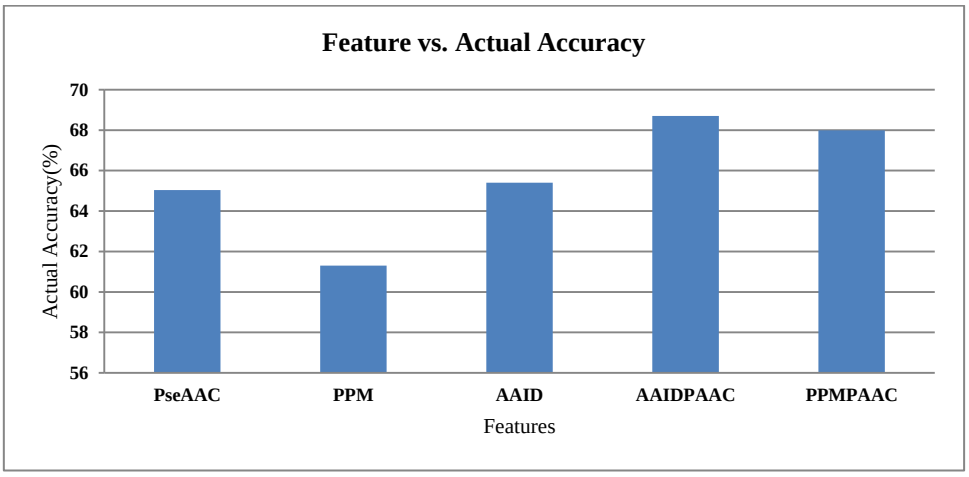

Figure 4: Actual Accuracy (%) for different Features

In the following Figure 5, we observe that best locative accuracy 71.05% is achieved by PPMPAAC which is also a feature fusion representation. The locative accuracy of PPMPAAC is 2.5%, 2.2% and 2% heigher than PseAAC, PPM and AAID respectively.

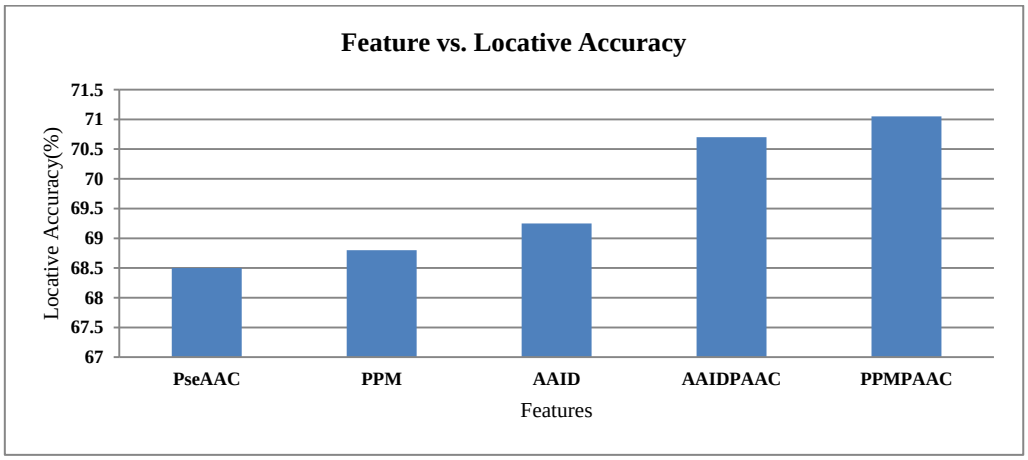

Figure 5: Locative Accuracy (%) for different Features

The feature fusion representations perform better actual and locative accuracy than single features as illustrated in Figure 4 and Figure 5.

5 Conclusion

In this paper, we present an approach to improve the accuracy for protein subcellular localization prediction. In protein subcellular localization prediction using machine learning technique, informative feature extraction methods from protein sequence mostly affect the performance. If inappropriate, noisy and less informative feature extraction methods are selected for classification then the accuracy is decreased instead of increasing. In this paper, we have used SVM, one of the widely used machine learning techniques with five distinct features for subcellular localization prediction. Among them, three is single feature representations and two is feature fusion representations. From the result, we have seen that feature fusion representation performs better than single features. In this paper, we have evaluated the performance on Gram negative bacterial dataset. The actual and locative

accuracy using fusion representation such as AAIDPAAC and PPMPAAC are at least 3% and 2% higher respectively in comparison to single feature representation. Nevertheless, still it is a challenge to achieve higher accuracy by using more efficient methods and it is the important part of our future work.

References

[1] S. Wang and S. Liu. Protein sub-Nuclear localization based on effective fusion representations and dimension reduction algorithm LDA. International Journal of Molecular Sciences, 16:30343–30361, Dec. 2015.

[2] Y.-D. Cai, X.-J. Liu, and K.-C. Chou. Artificial neural network model for predicting protein subcellular location. Computers and Chemistry, 26:179–182, Jan. 2002.

[3] K.-J. Park and M. Kanehisa. Prediction of protein subcellular locations by support vector machines using compositions of amino acids and amino acid pairs. Bioinformatics, 19(13):1656–1663, 2003.

[4] Y. Huang and Y. Li. Prediction of protein subcellular locations using fuzzy k-NN method. Bioinformatics, 20(1):21–28, 2004.

[5] H.-B. Shen and K.-C. Chou. Nuc-PLoc: a new web-server for predicting protein subnuclear localization by fusing PseAA composition and PsePSSM. Protein Engineering, Design & Selection, 20(11):561–567, 2007.

[6] L. Li, S. Yu, W. Xiao, Y. Li, M. Li, L. Huang, X. Zheng, S. Zhou, and H. Yang. Prediction of bacterial protein subcellular localization by incorporating various features into Chou's PseAAC and a backward feature selection approach. Biochimie, 104:100–107, Jun. 2014.

[7] C. Huang and J. Yuan. Using radial basis function on the general form of Chou's pseudo amino acid composition and PSSM to predict subcellular locations of proteins with both single and multiple sites. BioSystems, 113(1):50– 57, Apr. 2013.

[8] L. Li, Y. Zhang, L. Zou, C. Li, B. Yu, X. Zheng, and Y. Zhou. An Ensemble Classifier for Eukaryotic Protein Subcellular Location Prediction Using Gene Ontology Categories and Amino Acid Hydrophobicity. PLoS ONE, 7(1), Jan. 2012.

[9] S. Wan, M.-W. Mak, and S.-Y. Kung. mGOASVM: Multi-label protein subcellular localization based on gene ontology and support vector machines. BMC Bioinformatics, 2012.

[10] C.-S. Yu, C.-W. Cheng, W.-C. Su, K.-C. Chang, S.-W. Huang, J.-K. Hwang, and C.-H. Lu. CELLO2GO: A Web Server for Protein subCELlular LOcalization Prediction with Functional Gene Ontology Annotation. PLOS ONE, 9(6), Jun. 2014.

[11] X. Wang, J. Zhang, and G.-Z. Li. Multi-location gram-positive and gram-negative bacterial protein subcellular localization using gene ontology and multi-label classifier ensemble. BMC Bioinformatics, 16(12), 2015.

[12] X. Qu, Y. Chen, S. Qiao, D. Wang, and Q. Zhao. Predicting the subcellular localization of proteins with multiple sites based on multiple features fusion. presented at the ICIC, 2014, :456–465.

[13] A. Dehzangi, R. Heffernan, A. Sharma, J. Lyons, K. Paliwal, and A. Sattar. Gram-Positive and Gram-Negative Protein Subcellular Localization by Incorporating Evolutionary-based Descriptors into Chou's General PseAAC. Journal of Theoretical Biology, :284–294, 2015.

[14] X. Xiao, Z.-C. Wu, and K.-C. Chou. A Multi-Label Classifier for Predicting the Subcellular Localization of Gram-Negative Bacterial Proteins with Both Single and Multiple Sites. PLoS ONE, 6(6), Jun. 2011.

[15] IUPAC — IUB Commission on Biochemical Nomenclature, A One-Letter Notation for Amino Acid Sequences (Definitive Rules). Pure Appl. Chem., 34(4):639–646, 1971.

[16] K.-C. Chou. Prediction of Protein Cellular Attributes Using Pseudo-Amino Acid Composition. PROTEINS: Structure, Function, and Genetics, 43:246–255, 2001.

[17] S.-W. Zhang, L.-Y. Hao, and T.-H. Zhang. Prediction of Protein–Protein Interaction with Pairwise Kernel Support Vector Machine. International Journal of Molecular Sciences, 15:3220–3233, Feb. 2014.

[18] J. Yanga, J. Yanga, D. Zhangb, and J. Lua. Feature fusion: parallel strategy vs. serial strategy. Pattern Recognition, 36:1369 – 1381, 2003.

[19] W.-R. Qiu, B.-Q. Sun, X. Xiao, Z.-C. Xu, and K.-C. Chou. iPTM-mLys: identifying multiple lysine PTM sites and their different types. Bioinformatics, 32(20):3116–3123, 2016.

[20] Y. Xu, J. Ding, L.-Y. Wu, and K.-C. Chou. iSNO-PseAAC: Predict Cysteine S-Nitrosylation Sites in Proteins by Incorporating Position Specific Amino Acid Propensity into Pseudo Amino Acid Composition. PLOS ONE, 8(2), Feb. 2013.

[21] L. Li, H. Kuang, Y. Zhang, Y. Zhou, K. Wang, and Y. Wan. Prediction of eukaryotic protein subcellular multilocalisation with a combined KNN-SVM ensemble classifier. Journal of Computational Biology and Bioinformatics Research, 3(2):15–24, Feb. 2011.