JBE Journal of Biostatistics and Epidemiology
J Biostat Epidemiol. 2025;11(3): 346-360

ORIGINAL ARTICLE
DOI: *https://doi.org/10.18502/jbe.v11i3.21374/*
*Received **May** 28, 2025; Accepted **July** 08, 2025;*
*Published online **September** 15, 2025.*

Algorithm-Level Data-Guided Correction for Class Imbalance in Biological Machine Learning Predictions: Protein Interactions as a Case

Ebrahim **Barzegari**[1] , Parviz **Abdolmaleki**[2*]

[1]Medical Biology Research Center, Health Technology Institute, Kermanshah University of Medical Sciences, Kermanshah, Iran.
[2]Department of Biophysics, Faculty of Biological Sciences, Tarbiat Modares University, Tehran, Iran.

ABSTRACT

Introduction: In real-world biomedical applications of data mining, machine learning and artificial intelligence, there are situations where the widespread problem of class imbalance cannot be addressed by data-level methods such as over- or under-sampling. Correct and efficient use of algorithm-level methods, on the other hand, needs paying heed to data structure and content. This study aims to devise and examine simple methods for addressing the imbalanced class distribution issue in predicting the protein-protein interaction (PPI) sites in membrane proteins as a biomedical case experiment.

Methods: Using an adopted dataset of membrane protein complexes and a retrieved validation set, a class-weighted random forests (CWRF) classifier model was built for predicting interfacial residues from positional frequencies and an evolutionary index.

Results: Among several class weighting methods, a data imbalance-emulating weighting method for the CWRF model achieved an area under the receiver operating characteristics curve (AUC) of 0.815 (95% CI: 0.805-0.823) in the independent test prediction and 0.802 (95% CI: 0.794-0.809) in the prediction for the external validation set, which outperformed previous similar studies. A case prediction confirmed the practical utility of this method.

Conclusion: The proposed approach implies potential applications in other fields of biomedicine and beyond. It also highlights the role of algorithm-data interplay in addressing the class imbalance.

Key words: Machine learning; Bioinformatics; Statistical bias; Random forests; Protein-protein Interaction domains

*Corresponding Author:
parviz@modares.ac.ir

INTRODUCTION

In relationship modelling tasks that consider classification, the data structure can significantly influence the prediction performance. Real-world implementations of data mining frequently encounter the problem of imbalanced class distribution, which means the situation where the major part of the data points are contained in a majority class while a small proportion belong to an underrepresented minority class. In such cases, the classifying model tends to assign the input data mostly to the larger class, which leads to classification bias. Therefore, adopting prediction strategies suitable for uneven data is indispensable and is currently under active research.[1-3] The methods so far suggested to tackle the imbalance issue, handle the problem either at the data level, such as under- or over-sampling, or at the algorithmic level, such as applying cost to biased predictions.[4, 5] However, there are numerous situations where the data-level methods have not relevance or utility, including highly noisy data, unstructured data (e.g. images), data streams, and high-dimensional data. While algorithm-level methods can be used in such cases, a more precise predictive model would be obtained when the data structure is considered in the modification of the machine learning algorithm or hyperparameters. Adopting this approach, the present study aims to propose simple and novel algorithm-level methods for data-guided correction for data imbalance.

In the biomedical context, the issue of unbalanced classes or its rectification through data manipulation would lead to detrimental consequences such as mischaracterizations and misdiagnoses. Here, we adopt protein-protein interactions (PPIs) among membrane proteins as a basic biological case for examining our novel methods. Membrane proteins contribute to a diverse range of key functions in cells and are targets of more than half of the current drugs.[6] Knowledge about the interactions among these proteins can lay the ground for structure-based design of appropriate therapeutic medicines to manipulate these interactions for treating diseases.[7, 8] Protein chemistry techniques, such as crystallography and nuclear magnetic resonance, represent expensive and laborious methods for structural characterization of PPIs. The task is even more challenging for membrane-integrated proteins, mainly due to their hydrophobic surfaces.[9] Therefore, robust and reliable computational methods such as machine learning prediction of the key residues from sequences are urgently required.[10]

After successful implementation of DeepMind's AlphaFold to predict the protein structures,[11] the next breakthrough would be an efficient artificial intelligence tool for predicting protein complexes and interactions. Typically a small number of key interface residues participate to establish the major portion of the free energy in protein interactions.[12] As a result, the class of minority in their classification is actually the one containing positive instances, i.e. interaction site residues.[13] The imbalance problem should thus be a primary focus in the studies on predicting membrane proteins' interactions. The present study aims to address this concern, by proposing a simple, efficient and accurate data-guided method at the algorithmic level for identifying the key residues in the interaction interfaces of membrane proteins.

METHODS

Dataset

A preprocessed dataset of membrane protein complexes, containing pre-2009 records of PDBTM database[14] was taken from Bordner (2009) as the training set. In addition, another set of complexes was collected as the independent test set, retrieved from PDBTM updates since 2010 onward, with a protocol similar to that used for the training set.[15] This set of 502 complexes was culled to a non-redundant set using PISCES.[16] In the new set, no pair of membrane protein complexes comprised proteins having < 30% sequence similarity. Only surface residues located within the hydrophobic core of the membrane were included in the training and testing sets. Surface residues were defined as those with relative accessible surface area ≥ 0.2, and integration within membrane was predicted using TMDET.[17] Residues with < 4 Å heavy atom distance from opposing protein chain in the complex structure were labelled as interaction interfacial residue (grouped in class I), and otherwise labelled as a non-interfacial residue (grouped in class N).

To have an external validation set for testing our models, the latest update of Orientations of Proteins in Membranes (OPM) database [18] was searched for human membrane proteins with TM subunits ≥ 2, and the list of hits was downloaded (980 entries). PDB IDs existing in training and testing sets were removed, and items containing more than one unique chain were included in the new dataset as the membrane protein complexes. The same pre-processing steps as described above were performed to obtain the final validation dataset.

The Feature Space

Independent variables were defined for each individual position along the sequence, and included the residue frequency for the 20 standard amino acids in the multiple alignment of similar sequences, as well as the evolutionary rate. Sequences were aligned using BLAST with the cutoff of 10^{-2} for E-value; redundant sequences were then removed using CD-HIT.[19] The remaining sequences were used for generating multiple alignments by utilizing MUSCLE.[20] We calculated the residue frequency for each residue type by measuring the proportion of residues of that type in the corresponding column of multiple alignment. The training and testing sets involved only those proteins with a minimum of 20 sequences in the final alignment. For obtaining the evolutionary rate, the REVCOM method was used;[21] the evolutionary conservation values obtained by this method show robustness to the local alignment errors and the sequence set composition.

Class-Weighted Random Forests (CWRF) and Weighting Methods

The sample was split to 70%/30% train/test subsets. For accurate measurement of the prediction efficiency in future applications of the model on novel data, all residue-level data points associated to a particular protein were contained in one of the training or testing sets. The classifier models were trained and tested through a ten-fold cross-validation on the Bordner's dataset, followed by evaluation using the independent test set. Random forest (RF) was used as the main tool for relationship modelling.

It is a promising and widely-used algorithm involving an ensemble of unpruned decision trees, which offers numerous advantages such as efficiency with large datasets, avoiding overfitting, and high accuracy.[22] The model hyperparameters were optimized before implementation.

Dealing with skewed data demands taking the degree of data imbalance into consideration. In that regard, the ratio of the number of instances in the majority class to that in the minority class is defined as the imbalance ratio (IR).[23] One approach to make RFs suitable for learning from uneven data follows the idea of cost sensitive learning. In class-weighted random forests (CWRF), this is realized by assigning weights to classes. Weights in CWRF are an essential model tuning parameter, with the performance for independent test set as a benchmark.[24] We devised and applied several weighting approaches to find the one achieving the highest distinction between interaction classes. Three CWRF models were built this way, including:

o CWRF-1:
Class weight for the large class (i.e. the negative class N) = 1;
Class weight for the small class (i.e. the positive class I) = IR;

o CWRF-2:
Class weight = Total sample size / Number of observations in the class of interest;

o CWRF-3:
Class weights are tuned so that
(Predicted negatives / Predicted positives) = (TN+FN)/(TP+FP) ≈ IR,
where TN is the number of true negative predictions, FN denotes false negative, TP is true positive, and FP shows false positive predictions.

Importance of Residues in Interactions

As an additional criterion of performance, the share of protein residues in the physical interactions was also measured and compared between different models. Our CWRF model estimated the contribution of input variables to the performance of prediction using its specific 'variable importance' measure, which was determined from the mean decrease in accuracy resulting from random scrambling of predictor values.

Statistical Modelling

For comparison, the binomial logistic regression (BLR) model was also applied to classify the residues. The choice of BLR was because it can be used as a standard statistical classifier, which also offers options for dealing with class imbalance, and provides feature weights to compare with RF's importance values. To handle the data imbalance, we chose to change the model cut-off value. In the BLR model, Exp(B) coefficient (odds ratio) provided a measure of how much the corresponding parameter contributes to determining the interaction state of residues.

Model Performance Evaluation

For performance evaluation of the models in the imbalanced class distribution cases, the prediction accuracy (PA) is not a proper criterion as this measure will receive the least impact from the rare class.[25] However, the PA of individual classes may be useful. In this study, we evaluated our models with the area under the receiver operating characteristics (ROC) curve (AUC). Additionally, we used two performance criteria specifically developed for evaluation in class imbalance problems, including F-measure[26] and Geometric mean (GM).[27] PA, F and GM are defined as in Eq. (1)-(3).

$$PA = (TP + TN) / (TP + FP + FN + TN) \quad (1)$$

$$F = 2TP / (2TP + FP + FN) \quad (2)$$

$$GM = ((TP/(TP+FN)).(TN/(FP+TN)))^{1/2} \quad (3)$$

Building, running, evaluation and other analyses of the models and their output were conducted in the R programming language.

RESULTS

Prediction Performance

The three devised class-weighting methods were applied in the CWRF implementation, using the imbalance ratio of 2.62 for our in-house dataset (Non-interface residues (N class) / Interface residues (I class) = 3604/1375). AUC, GM and F measures obtained from the independent test set prediction showed that the highest performance was demonstrated by CWRF-3, i.e. the model which tunes the class weights to make the predictions coincide with the imbalance ratio. Results of performance evaluation of BLR and the CWRF-3 model have been compared in Table 1. The ROC curves were plotted and shown in Figure 1. For CWRF-3, the AUC of training was 0.844 (95% CI: 0.823-0.868). Also, the acceptable AUC values of 0.815 (95% CI: 0.805-0.823) and 0.802 (95% CI: 0.794-0.809) were obtained in the prediction for the independent test set and external validation, respectively. Performance of the BLR model was relatively weak (AUC = 0.713 (95% CI: 0.702-0.727) in ten-fold cross-validation, 0.691 (95% CI: 0.684-0.699) in the independent test and 0.683 (95% CI: 0.675-0.689) using the external validation set; max. PA = 81%; optimal cut-off of 0.663).

Contribution of Residue Types to Membrane Protein-Protein Interfaces

Table 2 presents the importance results of the two models and compares the orders of variable contributions suggested by different models in this research and the Bordner's study.[15] As this table shows, the contribution orders suggested by CWRF-3 and the random forests in the previous study follow approximately the same pattern, while that suggested by the BLR model is in relative accordance with those two. In general, alanine, the evolutionary rate, glycine and leucine can be proposed as the factors mostly contributing in membrane protein interaction interface sites, and glutamic acid, asparagine, cysteine and lysine as the least contributing residues.

Table 1. Performance measures of the utilized predictive models

Model	Testing method	AUC (95% CI)	F (95% CI)	Geometric Mean (95% CI)	Prediction accuracy (%)		
					Total	class I	class N
Binomial logistic regression							
	Ten-fold cross-validation	0.713 (0.702-0.727)	0.7794 (0.7547-0.8011)	0.8242 (0.8014-0.8570)	83.0	78.1	83.9
	Independent test set	0.691 (0.684-0.699)	0.7292 (0.7138-0.7516)	0.7807 (0.7574-0.8032)	79.6	74.2	81.0
	External validation set	0.683 (0.675-0.689)	0.7048 (0.6898-0.7155)	0.7560 (0.7505-0.7612)	77.9	72.2	79.5
Class-weighted random forest using the class weighting method 3 (CWRF-3)							
	Ten-fold cross-validation	0.844 (0.823-0.868)	0.8436 (0.8254-0.8616)	0.8915 (0.8817-0.9026)	91.0	85.2	93.3
	Independent test set	0.815 (0.805-0.823)	0.8072 (0.7786-0.8247)	0.8698 (0.8639-0.8788)	89.1	82.7	91.5
	External validation set	0.802 (0.794-0.809)	0.7900 (0.7815-0.7991)	0.8583 (0.8520-0.8641)	88.2	81.4	90.6

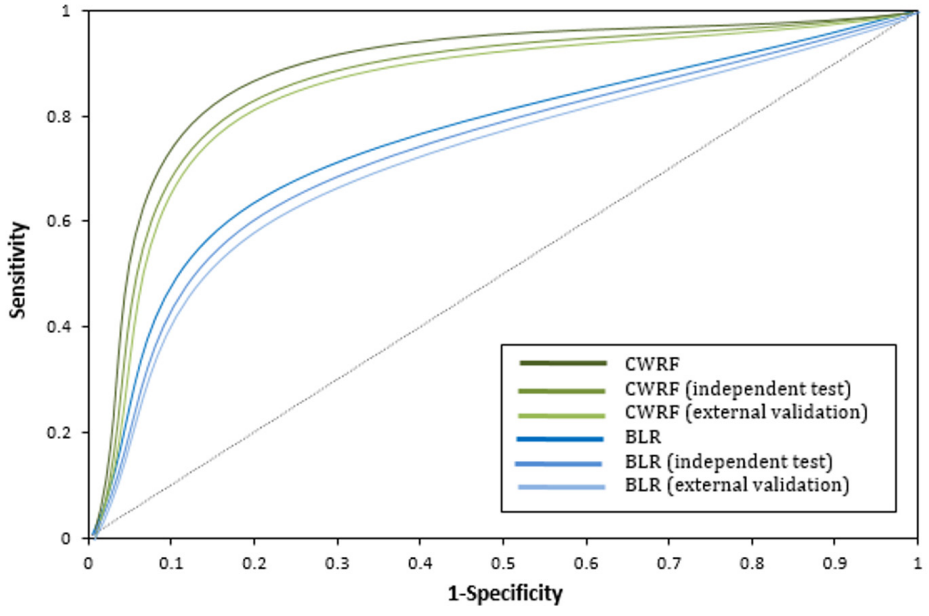

Figure 1. The receiver operating characteristics (ROC) curve plot for the classifications by the Class-Weighted Random Forest using the class weighting method 3 (CWRF-3) and the Binomial Logistic Regression (BLR) models using the training dataset, the independent test set and the external validation set.

Table 2. Order of importance for input parameters, and its comparison between implemented models

Class-weighted random forest using the class weighting method 3 (CWRF-3)	Random forests (Bordner, 2009)	Binomial logistic regression
Ala	Ala	Evolutionary Rate
Evolutionary Rate	Leu	Leu
Gly	Gly	Gly
Leu	Val	Phe
Met	Evolutionary Rate	Val
Val	Met	Met
Ile	Phe	Ala
Phe	Ile	Asp
His	Trp	Trp
Trp	Ser	Gln
Ser	Arg	Ser
Arg	Lys	Ile
Asp	Thr	Lys
Tyr	Asn	Arg
Thr	Cys	Tyr
Pro	Pro	Asn
Gln	His	Pro
Lys	Tyr	Glu
Cys	Gln	His
Asn	Asp	Thr
Glu	Glu	Cys

Case Prediction

To demonstrate the effectiveness and utility of our approach for identifying the interaction interface residues, we drew the CWRF-3 predictions for a protein complex from our collected dataset, and the predictions were scrutinized in terms of accordance/discordance with previous experimental findings. By querying for the structures which have been challenging in the recent years, Glycophorin A (GpA), a sialoglycoprotein in human red blood cell membrane with Protein Data Bank ID 2KPE [28] was chosen as the case. Prediction of its interaction site residues is shown in Figure 2. Of the seven positive residues of GpA (Leu75, Gly79, Val80, Gly83, Val84, Thr87, Ile88), four were correctly predicted as interfacial residues. Prediction of Gly79 and Gly83 is consistent with the well-known role of GxxxG motif in membrane protein interactions.[29] Val80 and Val84, as large hydrophobic beta-branched residues, have proved to perform an essential structural role by the space-filling capacity of their side chains.[29] Regarding the results obtained, this method for prediction of interaction sites offers a valuable first approach for guiding wet-lab investigations on membrane protein-protein interactions and localizing the specific residues at membrane protein interfaces.

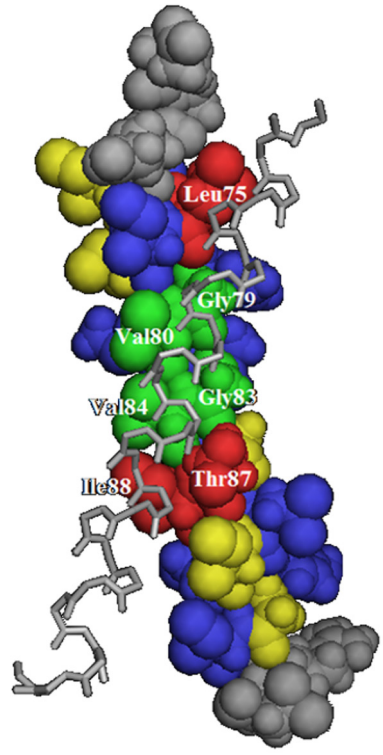

Figure 2. Prediction of the interaction site residues for Chain A of Glycophorin A (GpA) dimeric TM segment. Chain A is shown in space-filling model and the other chain in sticks. Labels show the amino acids from the protein to the back. Colour code of residues: True positives in Green; True negatives in Blue; False positives in Yellow; False negatives in Red. Out-of-membrane residues are shown in grey.

DISCUSSION

Regarding the importance of the data imbalance problem, this study was devoted to address this issue in the prediction of interaction sites for membrane proteins, by proposing a novel method to apply the class-weighted random forests classifier. The reason why the random forest model was chosen instead of other machine learning algorithms is the advantages inherent and established for this model, including high accuracy, low bias, low overfitting, and high speed, making it the first-line choice in many data mining applications. In addition, among other machine learning methods, not all methods allow implementations to handle the class imbalance. An advantage when working with the class-weights feature in the RF algorithm is that the model implementation would involve the tuning of a few numbers, which would be easier and less error-prone in comparison with when the model is supposed to be profoundly modified.

The present study aimed to provide and test the new class-weighting methods. Currently, the class-weights feature is embedded in the Random Forest model, but not available or applicable using other ML algorithms such as SVM, deep NNs or graph-based methods. This is why we chose RF for this work. Nevertheless, we understand that a baseline method is always required for making comparisons. For several reasons, BLR was a suitable comparator for CWRF. Firstly, it has an adjustable hyperparameter (the model cut-off) that can be changed according to the rate of data imbalance; thus, it can be utilized

as an equivalent to class weights. We tuned and optimized this parameter, and the effect on prediction performance was compared between the optimal BLR cut-off and the best-performing CWRF class weights. Secondly, for determining important predictors (e.g. the key residues in protein-protein interfaces), BLR provides coefficients that can be considered as objective weights, while other ML methods do not provide such explainable output with that level of convenience. Thirdly, BLR is a simple and widely-used statistical technique without the complications associated with sophisticated ML algorithms.

The aspect of novelty in this work and its superiority to other approaches lies in defining some strategies for determining the weights when applying the class-weighting technique. This strategy is guided by the data structure and properties, therefore it will be tailored domain-specifically for each type of input data. With this approach, researchers will have a rationale for deciding as to what exact value should be used as the weight for each class in an imbalanced classification. Therefore, blind, baseless and inaccurate determination of weights will be avoided this way. We applied the proposed technique for PPI prediction to show its applicability for an example real-world biological dataset. The performance improvement (although marginal), as well as the identification of previously established residue contributors confirmed that this approach has real utility in biomedical machine learning predictions.

Generally, there are two approaches used to address the class imbalance issue; including the strategies applied at the data level, and those applied at the algorithmic level.[4, 5] In the present study, we devised an algorithmic-level method which utilizes the data structure as a guide for tuning the learning machine. The best model was the one with weights adjusted to correspond with the degree of data imbalance in the dataset. Thus, our solution represents a hybrid approach combining data-level and algorithmic-level methods in handling the imbalance issue. The significance of this methodology is to highlight the interplay of data and algorithm in addressing the class imbalance, as opposed to a CWRF or any ML method blind to properties of the data of interest. The workflow applied in this study provided an exemplar of such interplay.

Despite the existence of dozens of research papers and several webservers for dealing with interaction site prediction,[30] there is still intense ongoing research aiming to develop novel methods to improve the predictions for both cytosolic and membrane proteins.[31-33] Paucity of research for interaction predictions in membrane proteins is evident from a search in the literature. Accordingly, progressive development and improvement of data mining and artificial intelligence modelling algorithms is key to handle this challenge. Our obtained AUC values of 0.815 (95% CI: 0.805-0.823) in the classification for the independent test set and 0.802 (95% CI: 0.794-0.809) in the prediction for the external validation set outperformed the previous similar studies. The research by Bordner reported AUC of 0.75 in its independent test, similarly using all lipid-facing residues.[15] A study by our team on the same datasets using 73 classifiers reached a maximum AUC of 0.786 in the prediction for the independent test set using a tuned support vector machine.[34] For reference, we have also outlined and compared methods, performances and residue importance in a review.[30] In total, comparison of these results indicate the high efficiency of the novel data-guided algorithmic-level method proposed in this study.

It was observed that the improved performance of the CWRF-3 model was significant in most cases, and the few cases of overlap between confidence intervals encompassed small and negligible ranges. The same result was obtained using imbalance-specific measures such as GM and F for CWRF-3, indicating the model's success in making accurate predictions using unbalanced PPI data. Considering the prediction accuracy, it was seen that, though the negative class is predicted with higher accuracy in all cases, the rate of correct predictions for the positive class was also remarkably high, particularly in the classification made by CWRF-3. From biomedical perspective, these improvements indicate that our proposed model is rigorous at distinguishing between the positive and negative classes in a biological or medical problem where the two classes have unequal size, which covers a wide range of applications in the biomedical domain.

Identifying residues such as Ala, Val, Leu, Ile and Gly as the most influential factors in membrane protein interfaces by our model is in complete agreement with previous studies. Extensive investigation has proved the tight packing, specifically through GxxxG motifs in helix lateral bindings, as the main driving force promoting the membrane protein associations.[35] Such van der Waals interactions are mediated by hydrophobic residues. In comparison with non-membrane proteins, the membrane-associated proteins follow a relatively converse pattern in terms of frequent residues in their interfaces. For example, while Lys and Glu are frequent in cytosolic protein interaction sites,[36] they are not preferred in membrane protein binding sites. Ala, Gly and Leu are disfavoured within interaction interfaces of non-membrane proteins,[36] while they have high frequency within membrane protein interaction interfaces, as indicated by our study. Evolutionary conservation was also shown by our model to be of great importance in the interactions, as it has been the case in other studies.[37-40] Conserved residues constitute a key residue set in interacting domains. These residues are clustered as tightly packed regions in the centre of protein interfaces, and play a critical role in maintaining the stability of protein association and preserving the protein function.[41]

The approach proposed in this study has some aspects of limitation and challenge which may restrain its applicability. Firstly, it applies the class-weighting as part of the RF as the machine learning technique, thus its implementation with other machine learning models such as neural networks or support vector machines will require devising specific methodologies. Secondly, the model's generalizability to other biological and medical datasets needs a thorough investigation since extreme rates of data imbalance may reduce the effectiveness of weightings, leading to lower prediction performance. Thirdly, there may be potential problems when the input data suffer a high volume of missing values. To address these issues, we would suggest more research in the computer science field in collaboration with biologists and medical experts to explore novel algorithmic modifications and new approaches for handling missing data.

CONCLUSION

The in silico prediction method proposed in this study is efficient as a novel machine learning-based approach for discerning the usually more valuable but rare instances from the frequent cases of less value in many real-world applications. The use of this method for predicting the putative

binding site residues in membrane proteins is fundamental for detecting the targets of pharmacological therapeutics, obtaining insights on membrane proteins' function, and analysing membrane proteomes. In practice, this method can contribute to faster and more accurate identification of interaction sites in large sets of membrane proteins. Furthermore, it lays the ground for targeted drug design to treat illnesses caused by impaired function of a membrane protein. Our result improvements ensure less erroneous prediction of binding sites on membrane proteins; thus, it could enhance the chance of success in studies exploring the potential targets on the surface of membrane proteins.

Our suggested approach offers a streamlined algorithmic-level technique for improving the performance in imbalanced predictions by emulating the imbalance inherent in the input data. Practical application of this method can be extended to various biological and medical fields and beyond. Nevertheless, further research in the future is warranted to experimentally validate the predicted binding sites and to improve machine learning models for imbalanced data. In addition, utilization of the proposed technique with deep neural networks or more advanced methods based on explainable AI principles is recommended to enable model explainability in real-world applications.

Availability of Data and Materials

The data that support the findings of this study are available through Figshare at https://doi.org/10.6084/m9.figshare.28827632.

Conflict of Interests

None to disclose.

Authors' Contributions

EB: Conceptualization; Data Curation; Formal Analysis; Investigation; Methodology; Software; Validation; Visualization; Writing – Original Draft Preparation. PA: Investigation; Methodology; Project Administration; Resources; Supervision; Writing – Review & Editing.

Abbreviations

PPI: Protein-protein interaction; RF: Random forests; IR: Imbalance ratio; CWRF: Class-weighted random forests; BLR: Binomial logistic regression; PA: Prediction accuracy; ROC: Receiver operating characteristics curve; AUC: Area under the receiver operating characteristics curve; GM: Geometric mean; GpA: Glycophorin A.

ACKNOWLEDGMENT

This work was supported by the Faculty of Biological Sciences, Tarbiat Modares University, and the Deputy for Research and Technology, Kermanshah University of Medical Sciences [grant no. 1403021].

REFERENCES

1. Han K, Kim KZ, Oh JM, Kim IW, Kim K, Park T. Unbalanced sample size effect on genome-wide population differentiation studies. International Journal of Data Mining and Bioinformatics. 2012;6(5):490-504.

2. Fregoso-Aparicio L, Noguez J, Montesinos L, Garcia-Garcia JA. Machine learning and deep learning predictive models for type 2 diabetes: a systematic review. Diabetol Metab Syndr. 2021;13(1):148.

3. Malhotra R, Lata K. Handling class imbalance problem in software maintainability prediction: an empirical investigation. Frontiers of Computer Science. 2022;16(4).

4. Yen S-J, Lee Y-S. Cluster-based under-sampling approaches for imbalanced data distributions. Expert Systems with Applications. 2009;36(3):5718-27.

5. Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP. SMOTE: synthetic minority over-sampling technique. arXiv preprint arXiv:11061813. 2011.

6. Stagljar I, Fields S. Analysis of membrane protein interactions using yeast-based technologies. Trends in biochemical sciences. 2002;27(11):559-63.

7. Ge H, Walhout AJ, Vidal M. Integrating 'omic' information: a bridge between genomics and systems biology. Trends in genetics : TIG. 2003;19(10):551-60.

8. Balit T, Thonabulsombat C, Dharmasaroja P. Moringa oleifera leaf extract suppresses TIMM23 and NDUFS3 expression and alleviates oxidative stress induced by Abeta1-42 in neuronal cells via activation of Akt. Res Pharm Sci. 2024;19(1):105-20.

9. Maurel D, Kniazeff J, Mathis G, Trinquet E, Pin JP, Ansanay H. Cell surface detection of membrane protein interaction with homogeneous time-resolved fluorescence resonance energy transfer technology. Analytical biochemistry. 2004;329(2):253-62.

10. Zhao Z, Gong X. Protein-Protein Interaction Interface Residue Pair Prediction Based on Deep Learning Architecture. IEEE/ACM transactions on computational biology and bioinformatics. 2019;16(5):1753-9.

11. Jumper J, Evans R, Pritzel A, Green T, Figurnov M, Ronneberger O, et al. Highly accurate protein structure prediction with AlphaFold. Nature. 2021;596(7873):583-9.

12. Xia JF, Zhao XM, Song J, Huang DS. APIS: accurate prediction of hot spots in protein interfaces by combining protrusion index with solvent accessibility. BMC bioinformatics. 2010;11:174.

13. Liu GH, Shen HB, Yu DJ. Prediction of Protein-Protein Interaction Sites with Machine-Learning-Based Data-Cleaning and Post-Filtering Procedures. The Journal of membrane biology. 2016;249(1-2):141-53.

14. Kozma D, Simon I, Tusnady GE. PDBTM: Protein Data Bank of transmembrane proteins after 8 years. Nucleic Acids Res. 2013;41(Database issue):D524-9.

15. Bordner AJ. Predicting protein-protein binding sites in membrane proteins. BMC bioinformatics. 2009;10:312.

16. Wang G, Dunbrack RL. PISCES: a protein sequence culling server. Bioinformatics. 2003;19(12):1589-91.

17. Tusnady GE, Dosztanyi Z, Simon I. TMDET: web server for detecting transmembrane regions of proteins by using their 3D coordinates. Bioinformatics. 2005;21(7):1276-7.

18. Lomize MA, Pogozheva ID, Joo H, Mosberg HI, Lomize AL. OPM database and PPM web server: resources for positioning of proteins in membranes. Nucleic Acids Res. 2012;40(Database issue):D370-6.

19. Li W, Godzik A. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. Bioinformatics. 2006;22(13):1658-9.

20. Edgar RC. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res. 2004;32(5):1792-7.

21. Bordner AJ, Abagyan R. REVCOM: a robust Bayesian method for evolutionary rate estimation. Bioinformatics. 2005;21(10):2315-21.

22. Breiman L. Random forests. Machine learning. 2001;45(1):5-32.

23. Orriols-Puig A, Bernadó-Mansilla E. Evolutionary rule-based systems for imbalanced data sets. Soft Computing-A Fusion of Foundations, Methodologies and Applications. 2009;13(3):213-25.

24. Chen C, Liaw A, Breiman L. Using Random Forest to Learn Imbalanced Data. Statistics Department: University of California at Berkeley; 2004. Report No.: 666.

25. Weiss GM. Mining with rarity: a unifying framework. Sigkdd Explorations. 2004;6(1):7-19.

26. Lewis DD, Gale WA, editors. A sequential algorithm for training text classifiers. The 17th annual international ACM SIGIR conference on Research and development in information retrieval; 1994: Springer-Verlag New York, Inc.

27. Steinbach P, Kumar M, Tan V. Introduction to data mining. International Edition–NY: Addison Wesley. 2006.

28. Mineev KS, Bocharov EV, Volynsky PE, Goncharuk MV, Tkach EN, Ermolyuk YS, et al. Dimeric structure of the transmembrane domain of glycophorin A in lipidic and detergent environments. Acta Naturae. 2011;3:90-8.

29. Senes A, Gerstein M, Engelman DM. Statistical analysis of amino acid patterns in transmembrane helices: the GxxxG motif occurs frequently and in association with beta-branched residues at neighboring positions. J Mol Biol. 2000;296:921-36.

30. Barzegari Asadabadi E, Abdolmaleki P. A review and comparative assessment of machine learning approaches for interaction site prediction in membrane proteins. Current Bioinformatics. 2015;10(3):284-91.

31. Liu L, Zhu X, Ma Y, Piao H, Yang Y, Hao X, et al. Combining sequence and network information to enhance protein-protein interaction prediction. BMC bioinformatics. 2020;21(Suppl 16):537.

32. Xie Z, Deng X, Shu K. Prediction of Protein-Protein Interaction Sites Using Convolutional Neural Network and Improved Data Sets. International journal of molecular sciences. 2020;21(2).

33. Zhong X, Rajapakse JC. Graph embeddings on gene ontology annotations for protein-protein interaction prediction. BMC bioinformatics. 2020;21(Suppl 16):560.

34. Barzegari Asadabadi E, Abdolmaleki P. Predictions of protein-protein interfaces within membrane protein complexes. Avicenna Journal of Medical Biotechnology. 2013;5(3):148-57.

35. Lemmon MA, Flanagan JM, Treutlein HR, Zhang J, Engelman DM. Sequence specificity in the dimerization of transmembrane alpha-helices. Biochemistry. 1992;31:12719-25.

36. Gallet X, Charloteaux B, Thomas A, Brasseur R. A fast method to predict protein interaction sites from sequences. J Mol Biol. 2000;302(4):917-26.

37. Guo L, Wang S, Li M, Cao Z. Accurate classification of membrane protein types based on sequence and evolutionary information using deep learning. BMC bioinformatics. 2019;20(Suppl 25):700.

38. Li BQ, Feng KY, Chen L, Huang T, Cai YD. Prediction of protein-protein interaction sites by random forest algorithm with mRMR and IFS. PloS one. 2012;7(8):e43927.

39. Wang X, Yu B, Ma A, Chen C, Liu B, Ma Q. Protein-protein interaction sites prediction by ensemble random forests with synthetic minority oversampling technique. Bioinformatics. 2019;35(14):2395-402.

40. Nicoludis JM, Gaudet R. Applications of sequence coevolution in membrane protein biochemistry. Biochimica et biophysica acta Biomembranes. 2018;1860(4):895-908.

41. DeLano WL. Unraveling hot spots in binding interfaces: progress and challenges. Current opinion in structural biology. 2002;12(1):14-20.