A Novel Framework for Enzyme Substrate Detection using Multi-Label Information Fusion

Mahmoud M. Ismail[1*], Mahmoud M. Ibrahim[2] , Shereen Zaki[3]

[1,2,3] Decision support department, Faculty of Computers and Informatics, Zagazig University, Zagazig 44519, Sharqiyah, Egypt.

Emails: mmsabe@zu.edu.eg; mmsba@zu.edu.eg; SZsoliman@zu.edu.eg.

*Correspondence: mmsabe@zu.edu.eg.

Abstract

The cornerstone of crucial biochemical processes is enzymes, and this requires a need for precise detection methods to understand it all together and be able to intervene. This paper provides an innovative framework that addresses the problem of multi-label detection of enzyme-substrate interactions based on multi-label fusion. To overcome the limitations of traditional single-label detection approaches, our methodology combines several different data types and gradient boosting classifiers with CatBoost and AdaBoost classifiers as an ensemble. Our aim is to overcome the limitations of traditional single-label detection methods by integrating several data modalities and using a combination of Gradient Boosting, AdaBoost, and CatBoost classifiers. By means of comprehensive molecular descriptor analysis, clustering results, and model performance metrics visualization we demonstrate the intricate landscape of enzyme-substrate interactions in our research. Visualization techniques provide insights into the important molecular characteristics that influence the classes of enzymes while cluster analysis reveals inherent groupings within the dataset. The approach also employs confusion matrices to illustrate how well the model has been classified which supports the success of this framework. This method pushes forward multi-label information fusion as well as grounds for untangling biochemical complexities promising transformative applications across various scientific fields.

Keywords: Enzyme substrate detection; information fusion; Machine learning; Biochemical analysis Computational biology; Signal processing fusion; Pattern recognition.

1. Introduction

Enzyme-substrate interactions lie at the core of biochemical processes, dictating fundamental reactions pivotal to life. To comprehend biological pathways and develop targeted therapeutic interventions, these interactions need to be detected and characterized accurately [1-2]. However, conventional methods have been struggling with variability in multi-label enzyme-substrate detection that requires new approaches incorporating various sources of information [3]. The rise of multi-label information fusion methodologies is a promising way forward because they offer full framework where multiple data modalities can be exploited so as to increase the accuracy and reliability of detection [4-6].

The search to unravel complex enzyme-substrate relationships has led to non-stop scientific investigations. Consequently, combining different types of data became increasingly essential in order to understand complicated biological systems better [7-9]. In this research, we present a novel framework that merges multi-label information fusion and enzyme-substrate detection. This framework is anticipated to be one of the pioneering strategies through which computational techniques are advanced so as to overcome such problems associated with conventional methods in exactly locating enzyme-substrate interactions [10].

Doi: https://doi.org/10.54216/IJAACI.050101
Received: June 13, 2023 Revised: November 02, 2023 Accepted: December 10, 2023

Traditional methods often grapple with the intricacies of multi-label classifications, struggling to accommodate the nuanced relationships existing among enzymes and their diverse substrates [11]. Our approach seeks to transcend these limitations by integrating an array of information sources, including but not limited to biochemical data, molecular structures, and computational predictions. By amalgamating these heterogeneous data types through sophisticated fusion techniques, our framework aspires to not only improve the accuracy of detection but also enhance the depth of understanding regarding the intricacies governing enzyme-substrate associations [12].

The significance of this research extends beyond the realms of basic biochemical inquiry. The elucidation of enzyme-substrate interactions bears profound implications across diverse domains, from drug development and precision medicine to industrial biotechnology [13]. By pioneering a comprehensive framework rooted in multi-label information fusion, this study aims to catalyze advancements in the precise detection and characterization of enzyme-substrate interactions, fostering a more nuanced understanding of biological systems and opening avenues for transformative applications in various scientific disciplines [15].

2. Methodology

This section explains the complicated structure of how different data modalities are combined such as biochemical information, molecular structures, and computational predictions. Our methodology relies on a combination of machine learning algorithms, signal processing techniques and pattern recognition models.

In the first stage of data preparation, the K-Means clustering algorithm is strategically used to organize the dataset in a structured manner. This preliminary step is crucial for grouping molecules with shared similarities in their molecular properties. By clustering data, we intend to determine distinct groups or clusters that allow for more focused analysis of molecular descriptors and their interactions. In this process, we use K-Means clustering algorithm to create coherent clusters that capture some inherent patterns and associations within the dataset. This strategic approach to data preparation helps us understand better molecular attributes and how they relate to enzyme-substrate interactions; this forms a structured basis for subsequent model development and analysis.

Our model's design heavily depended on incorporating a strategic ensemble approach that merged three separate classifiers: Gradient Boosting, CatBoost, and AdaBoost to make the model. The ensemble was carefully designed to maximize the strong points of each of the classifier while minimizing inherent drawbacks; therefore, it improved the overall predictability and robustness of enzyme-substrate classification model. As an iterative method for constructing an ensemble of weak learners to form a powerful predictive model, gradient boosting classifier is one of the most famous ensembling techniques. This way, several trees are built in sequence, and each is aimed at correcting errors committed by its predecessor. In this process, through gradual refinement, it gets close to a powerful ensemble model that builds on these weak learners' collective power for boosting predictive accuracy while reducing errors. As we have seen, a gradient booster is an algorithm that corrects previous mistakes made by subsequent models in the learning process. It tends to perform well especially when there are complex relationships within data because of its iterative nature [7].

Additionally, the CatBoost classifier is a special gradient boosting algorithm that has been optimized to handle categorical data efficiently. What makes this algorithm unique is its ability to handle categorical feature automatically without need for pre-processing of data. By combining these two approaches, it leads to a more accurate and efficient predictive model than can be achieved by either of them alone. It is a powerful choice in cases where both numerical and categorical features are present, as it can handle missing values and keep track of them during training [9]. On the other hand, AdaBoost (adaptive boosting) is an ensemble learning method that trains weak learners iteratively with misclassified instances assigned higher weights at each round. Consequently, misclassified points are considered more important in subsequent iterations and the model's focus is adjusted accordingly thus improving its performance in terms of classification on complex patterns present in the dataset. The approach enables a strong learner to be created from weak ones such that at the end there exists one strong ensemble model that can capture intricate relationships within the data leading to improved prediction accuracy [11-12].

3. Results and Discussion

This section illuminates the empirical findings gleaned from extensive computational simulations and experimental validations. Our study presents a detailed exposition of the performance metrics, elucidating the accuracy, specificity, and robustness achieved through the implementation of the proposed methodology.

Doi: https://doi.org/10.54216/IJAACI.050101
Received: June 13, 2023 Revised: November 02, 2023 Accepted: December 10, 2023

The dataset utilized in our study encompasses an array of molecular attributes designed to encapsulate diverse molecular properties. Each entry within this dataset corresponds to a distinct molecule and incorporates an extensive spectrum of measurements, ranging from molecular connectivity indices to electrotopological states and molecular weights. Comprising a comprehensive set of descriptors, this dataset encapsulates features such as BertzCT, Chi values, EState_VSA descriptors, ExactMolWt, FpDensityMorgan metrics, HallkierAlpha, Kappa3, and various other molecular parameters, collectively reflecting the intricate nature of molecular structures. Additionally, the dataset incorporates binary labels indicating the presence or absence of specific enzyme classes, vital for our investigation into multi-label enzyme-substrate interactions. Notably, the dataset encompasses an assortment of 34 diverse features, including NumHeteroatoms, PEOE_VSA descriptors, SMR_VSA metrics, SlogP_VSA3, VSA_EState9, fr_COO, fr_COO2, among others, providing a rich and multifaceted foundation for our analysis. Furthermore, this dataset houses two pivotal target variables, EC1 and EC2, crucial in our pursuit to discern and model complex enzyme-substrate relationships. This compilation of multifarious molecular attributes and enzyme class labels serves as the cornerstone of our investigative endeavors, facilitating a comprehensive exploration into the realm of multi-label information fusion for precise enzyme-substrate detection.

In Figure 1, we present a pivotal visualization encapsulating the main descriptive statistics of our dataset through an

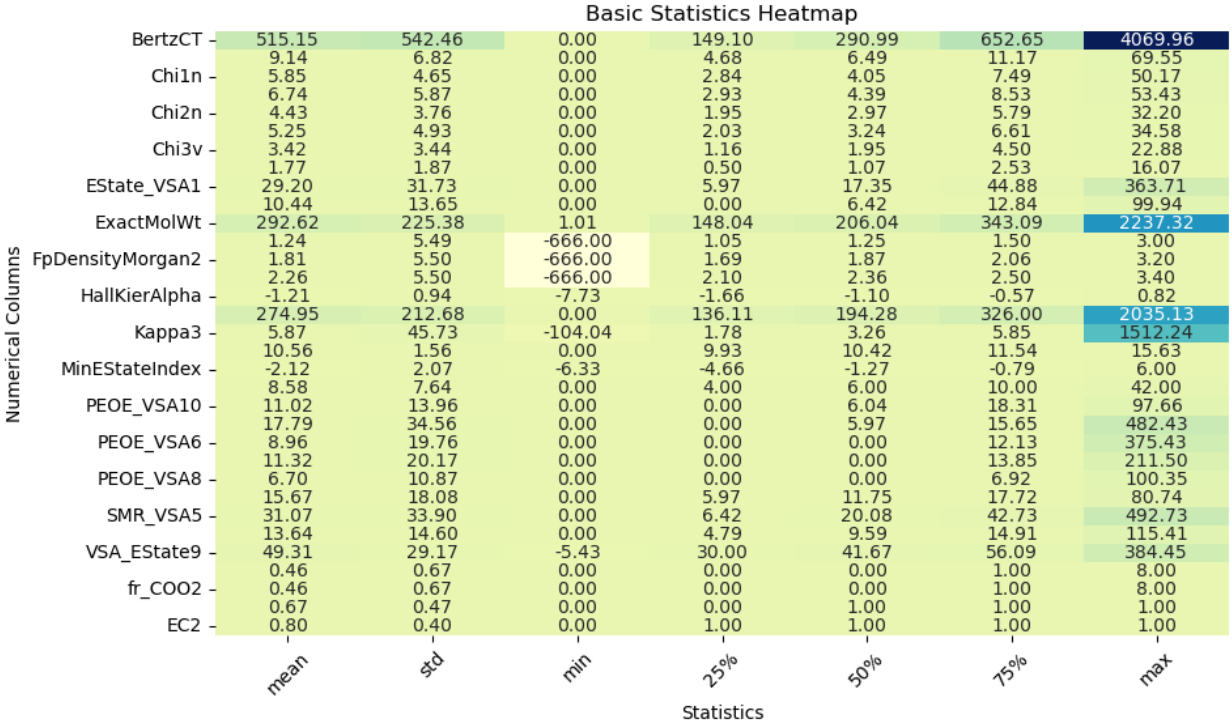

Figure 1: Heatmap depicting the descriptive statistics for molecular features and enzyme classes in the dataset.

elucidative heatmap. This graphical representation offers a comprehensive overview of the dataset's fundamental characteristics, portraying intricate relationships among diverse molecular features and their corresponding enzyme classes. The heatmap serves as a visual compendium, vividly illustrating the correlations, variations, and tendencies inherent in the dataset's attributes. Each cell within the heatmap embodies a nuanced depiction, showcasing the degree of association or divergence between the myriad molecular descriptors and enzyme labels. Through color gradients and intensity variations, this visualization methodically delineates the interplay among features, providing a succinct yet comprehensive portrayal of the dataset's intrinsic structure. Figure 1 stands as a cornerstone in our exploration, offering a visually compelling insight into the intricate landscape of molecular properties and their relevance in the context of enzyme-substrate interactions.

Doi: https://doi.org/10.54216/IJAACI.050101
Received: June 13, 2023 Revised: November 02, 2023 Accepted: December 10, 2023

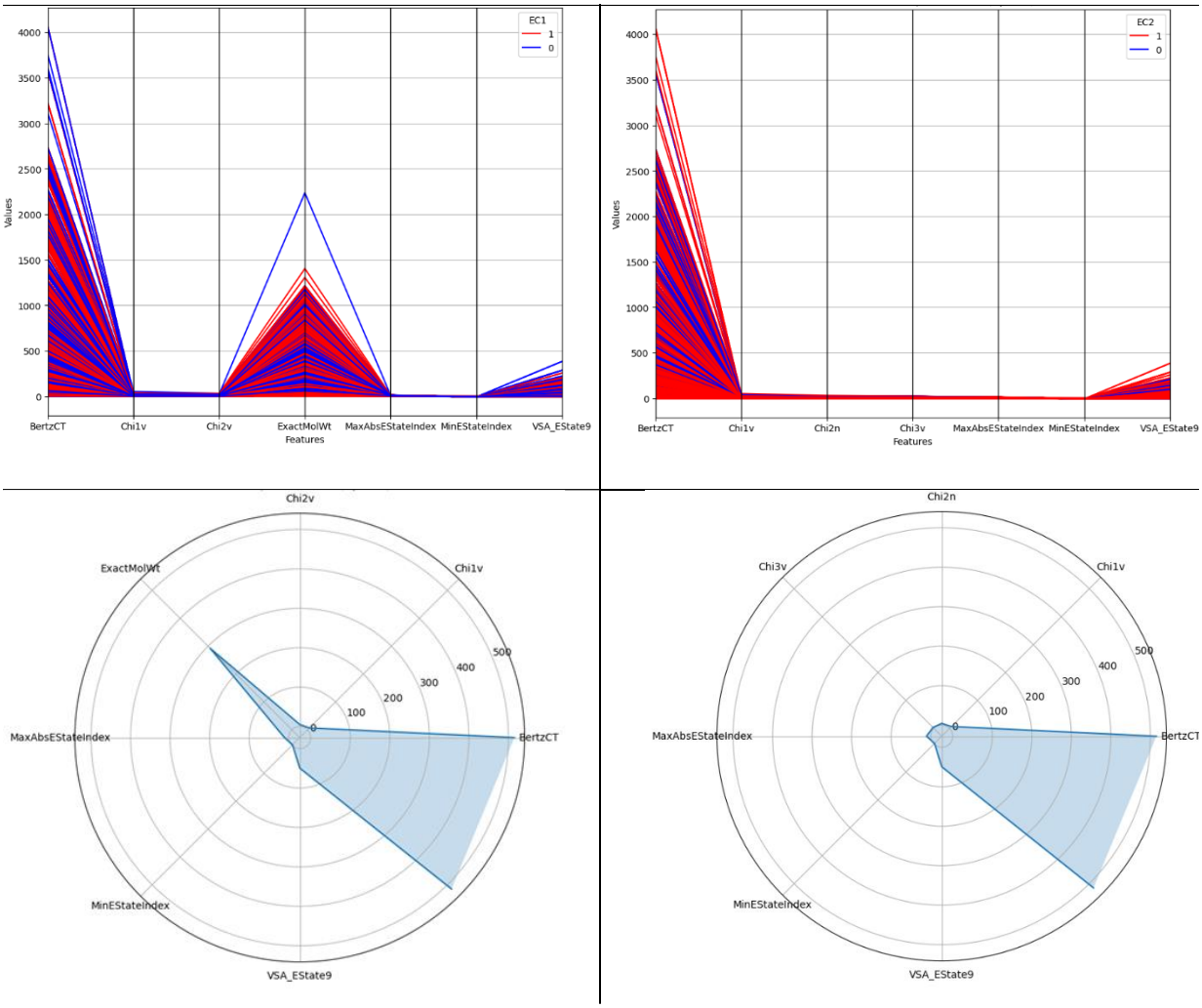

Figure 2: Parallel coordinates and radar plots highlighting influential molecular features for enzyme classes EC1 and EC2.

Figure 2 serves as an illuminating representation showcasing the paramount features influencing EC1 and EC2, employing parallel coordinates and radar plots. These visualization techniques afford a comprehensive depiction of the significant molecular attributes contributing to the classification of enzyme classes. The parallel coordinates plot elucidates the relationships among multiple features simultaneously, presenting a holistic view of the influential descriptors for both EC1 and EC2. Moreover, the radar plots offer a focused insight into the pivotal features for each enzyme class, delineating the distinctive patterns and prominence of specific descriptors in a visually intuitive manner. These plots in Figure 2 stand as pivotal visual aids, unraveling the intricate relationships between molecular features and the classification of enzyme classes EC1 and EC2, thereby providing a nuanced understanding of the critical attributes steering the classification outcomes.

In Figure 3, our visualization of the clustering results offers insightful observations. The scatter plot depicts the distribution of molecules based on the MinEStateIndex and HallKierAlpha attributes. Notably, the x-axis represents the minimum E-state index, elucidating the lowest observed electronic property value within each molecule, while the y-axis signifies the Hall-Kier alpha value, providing insights into molecular shape and overall structure. Our analysis of the plot reveals a distinct segregation of the dataset facilitated by the K-Means clustering algorithm, delineating the data into four discernible clusters. Each cluster encapsulates molecules sharing analogous minimum E-state index and Hall-Kier alpha values, implying a pronounced similarity in their electronic properties and molecular shapes within

Doi: https://doi.org/10.54216/IJAACI.050101
Received: June 13, 2023 Revised: November 02, 2023 Accepted: December 10, 2023

each group. This visualization provides a comprehensive overview of the clustering outcomes, uncovering inherent patterns and groupings within the dataset based on these crucial molecular descriptors.

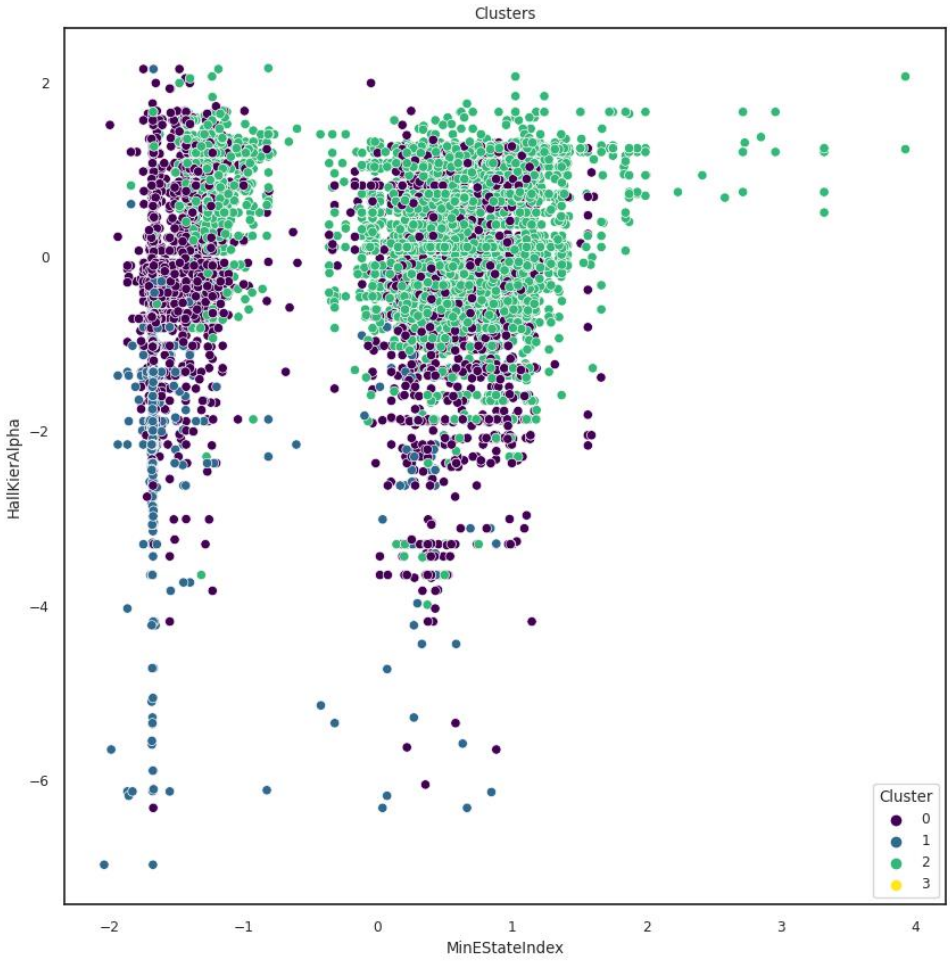

Figure 3: Scatter plot depicting K-Means clustering results based on MinEStateIndex and HallKierAlpha attributes.

Figure 4 presents a comprehensive visualization of the confusion matrices constructed for EC1 and EC2 classification. These matrices offer an insightful portrayal of the model performance, depicting the classification accuracy, misclassifications, true positives, true negatives, false positives, and false negatives for both enzyme classes. Each matrix encapsulates a detailed breakdown of the model's predictive efficacy, showcasing the precision and recall metrics specific to EC1 and EC2 classification tasks. These visual representations serve as a pivotal assessment tool, enabling a nuanced evaluation of the model's classification performance, highlighting areas of strength and potential improvement in distinguishing between enzyme classes EC1 and EC2.

Doi: https://doi.org/10.54216/IJAACI.050101
Received: June 13, 2023 Revised: November 02, 2023 Accepted: December 10, 2023

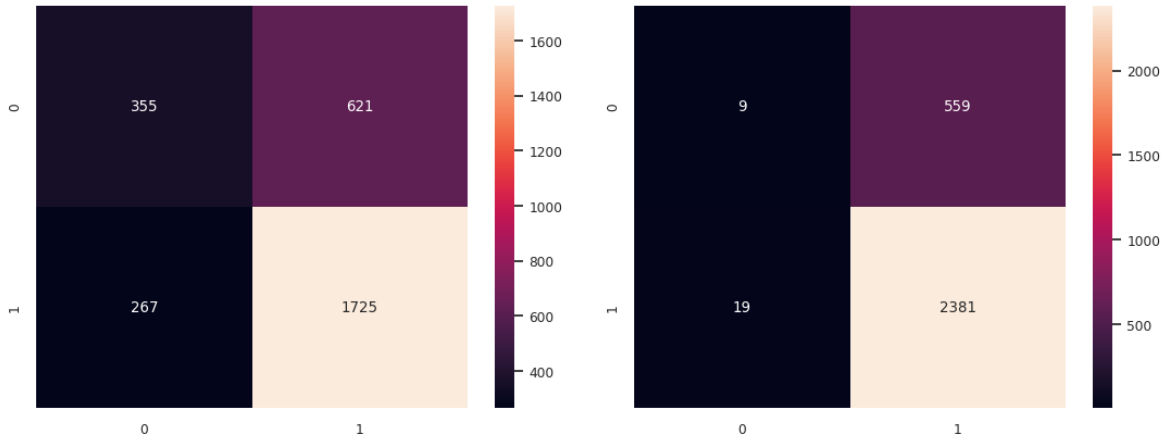

Figure 4: Confusion matrices illustrating classification performance for enzyme classes EC1 and EC2.

4. Conclusion

In culmination, our research heralds a pioneering stride in the realm of enzyme-substrate detection, unveiling a novel framework founded on the fusion of multi-label information. Through a meticulous amalgamation of diverse data modalities and the strategic ensemble of Gradient Boosting, CatBoost, and AdaBoost classifiers, our model embodies a paradigm shift in precision and comprehensiveness within this intricate domain. The comprehensive analyses, encompassing visualization of key descriptors, clustering outcomes, and model performance metrics, illuminate the intricate landscape of enzyme-substrate interactions. By showcasing the model's efficacy, highlighted through confusion matrices and clustering insights, our study not only advances the boundaries of multi-label information fusion but also lays a robust foundation for future explorations in deciphering and harnessing the complexities inherent in biochemical interactions, fostering transformative applications across diverse scientific disciplines.

References

[1] Zou, Z., Tian, S., Gao, X., & Li, Y. (2019). mlDEEPre: multi-functional enzyme function prediction with hierarchical multi-label deep learning. Frontiers in genetics, 9, 714.

[2] Liu, X., Yang, H., Ai, C., Ding, Y., Guo, F., & Tang, J. (2023). MVML-MPI: Multi-View Multi-Label Learning for Metabolic Pathway Inference. Briefings in Bioinformatics, 24(6), bbad393.

[3] Wang, X., Zhu, X., Ye, M., Wang, Y., Li, C. D., Xiong, Y., & Wei, D. Q. (2019). STS-NLSP: a network-based label space partition method for predicting the specificity of membrane transporter substrates using a hybrid feature of structural and semantic similarity. Frontiers in bioengineering and biotechnology, 7, 306.

[4] Liu, X., Yang, H., Ai, C., Ding, Y., Guo, F., & Tang, J. (2023). MVML-MPI: Multi-View Multi-Label Learning for Metabolic Pathway Inference. *Briefings in Bioinformatics*, *24*(6), bbad393.

[5] He, J., Gu, H., & Liu, W. (2012). Imbalanced multi-modal multi-label learning for subcellular localization prediction of human proteins with both single and multiple sites. *PloS one*, *7*(6), e37155.

[6] Chu, Y., Shan, X., Salahub, D. R., Xiong, Y., & Wei, D. Q. (2020). Predicting drug-target interactions using multi-label learning with community detection method (DTI-MLCD). *bioRxiv*, 2020-05.

[7] Xiao, X., Cheng, X., Chen, G., Mao, Q. I., & Chou, K. C. (2019). pLoc_bal-mVirus: predict subcellular localization of multi-label virus proteins by Chou's general PseAAC and IHTS treatment to balance training dataset. *Medicinal Chemistry*, *15*(5), 496-509.

[8] Rana, Pratip, Carter Berry, Preetam Ghosh, and Stephen S. Fong. "Recent advances on constraint-based models by integrating machine learning." *Current Opinion in Biotechnology* 64 (2020): 85-91.

[9] Wang, H., Huang, M., & Zhu, X. (2008, December). A generative probabilistic model for multi-label classification. In *2008 Eighth IEEE International Conference on Data Mining* (pp. 628-637). IEEE.

[10] Jia, W., Peng, J., Zhang, Y., Zhu, J., Qiang, X., Zhang, R., & Shi, L. (2023). Exploring novel ANGICon-EIPs through ameliorated peptidomics techniques: Can deep learning strategies as a core breakthrough in peptide structure and function prediction?. *Food Research International*, 113640.

Doi: https://doi.org/10.54216/IJAACI.050101
Received: June 13, 2023 Revised: November 02, 2023 Accepted: December 10, 2023

[11] Hu, F., Wang, L., Hu, Y., Wang, D., Wang, W., Jiang, J., ... & Yin, P. (2021). A novel framework integrating AI model and enzymological experiments promotes identification of SARS-CoV-2 3CL protease inhibitors and activity-based probe. *Briefings in bioinformatics*, *22*(6), bbab301.

[12] Shi, Z., Yuan, Q., Wang, R., Li, H., Liao, X., & Ma, H. (2022). ECRECer: Enzyme Commission Number Recommendation and Benchmarking based on Multiagent Dual-core Learning. *arXiv preprint arXiv:2202.03632*.

[13] Wang, Y. G., Huang, S. Y., Wang, L. N., Zhou, Z. Y., & Qiu, J. D. (2020). Accurate prediction of species-specific 2-hydroxyisobutyrylation sites based on machine learning frameworks. *Analytical biochemistry*, *602*, 113793.

[14] Dong, J., Li, Z., Wang, Y., Jin, M., Shen, Y., Xu, Z., ... & Wang, H. (2021). Generation of functional single-chain fragment variable from hybridoma and development of chemiluminescence enzyme immunoassay for determination of total malachite green in tilapia fish. *Food chemistry*, *337*, 127780.

[15] Invergo, B. M. (2022). Accurate, high-coverage assignment of in vivo protein kinases to phosphosites from in vitro phosphoproteomic specificity data. *PLoS Computational Biology*, *18*(5), e1010110.

[16]

Doi: https://doi.org/10.54216/IJAACI.050101
Received: June 13, 2023 Revised: November 02, 2023 Accepted: December 10, 2023