DEEP_TPPRED: IMPROVED PREDICTION OF PROTEIN TOXICITY USING FEATURE FUSION AND HYBRID NEURAL NETWORK APPROACH

Md. Mustahid Hasan, Md. Ashikur Rahman, Md Mamun Ali, Student Member, IEEE, Kawsar Ahmed, Graduate Student Member, IEEE, Francis M. Bui, Member, IEEE, Sobhy M. Ibrahim, Imran Mahmud, Kowshik Kumer and Mohammad Ali Moni

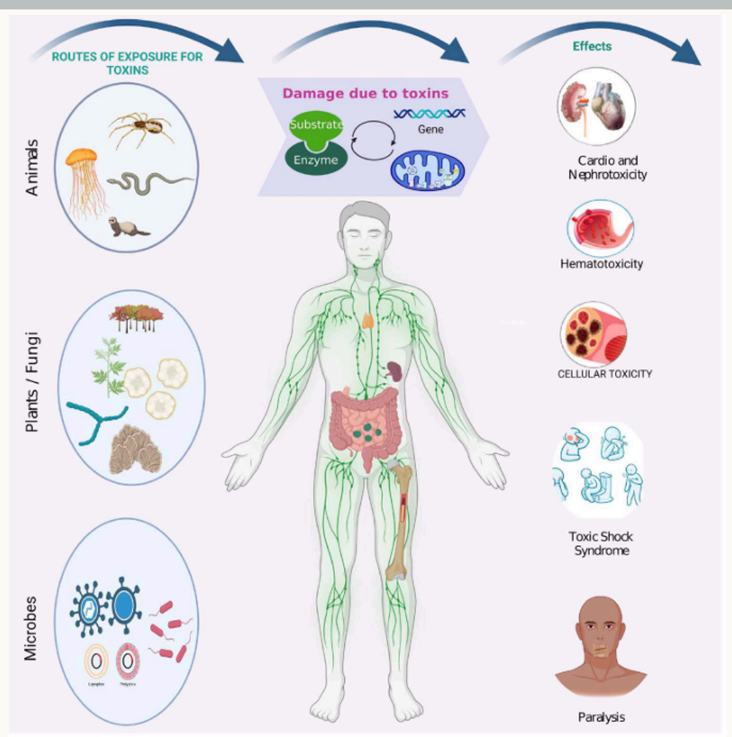


Figure 1: Graphical representation of sources of toxic proteins in nature and their effect on human health.

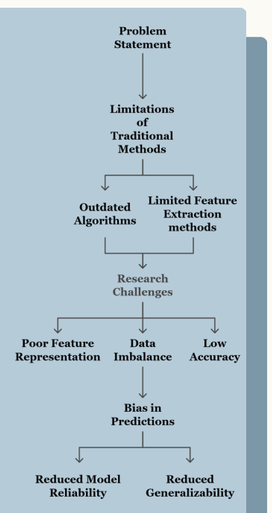
Abstract

Protein toxicity prediction is crucial in computational biology, with significant implications for drug discovery, safety assessment, and toxicological research. This study introduces Deep TPPred, a novel hybrid deep learning model that integrates Convolutional Neural Networks (CNN) and Recurrent Neural Networks (RNN) for accurate protein toxicity prediction. The model effectively combines diverse protein sequence descriptors to capture complex sequence relationships by leveraging a feature fusion technique. The methodology involved advanced feature extraction, rigorous training, and performance evaluation using benchmark datasets. Deep TPPred demonstrates state-of-the-art performance with an accuracy of 0.9983, specificity of 0.9988, sensitivity of 0.9975, and Kappa and MCC values of 0.9963. These results underscore the proposed model's robustness, reliability, and generalization capability, surpassing existing models across all metrics. The study highlights the potential of hybrid deep learning and feature fusion techniques to significantly enhance protein toxicity prediction, providing valuable insights and tools for bioinformatics research and applications.

Keywords – Protein Toxicity, Neural Network, Feature Fusion, Feature Extraction, Recurrent Neural Networks, Convolutional Neural Networks.

Problem Statement

Predicting toxic proteins is difficult due to their complex structure and the limits of current models. Many existing methods use limited features and struggle with imbalanced data, leading to poor results. Therefore, there's a need for a better model that uses diverse features and handles data imbalance to improve prediction accuracy.



Objectives

Diverse Feature Extraction: To apply various encoding methods that capture the biological and chemical properties of protein sequences.
Data Balancing: To use effective techniques that address class imbalance and ensure fair, accurate predictions.
Feature Fusion: To combine multiple feature sets into a unified representation for richer and more informative sequence analysis.
Hybrid Deep Learning Model: To develop a hybrid model that enhances accuracy in protein toxicity prediction over current state-of-the-art methods.

Methodology

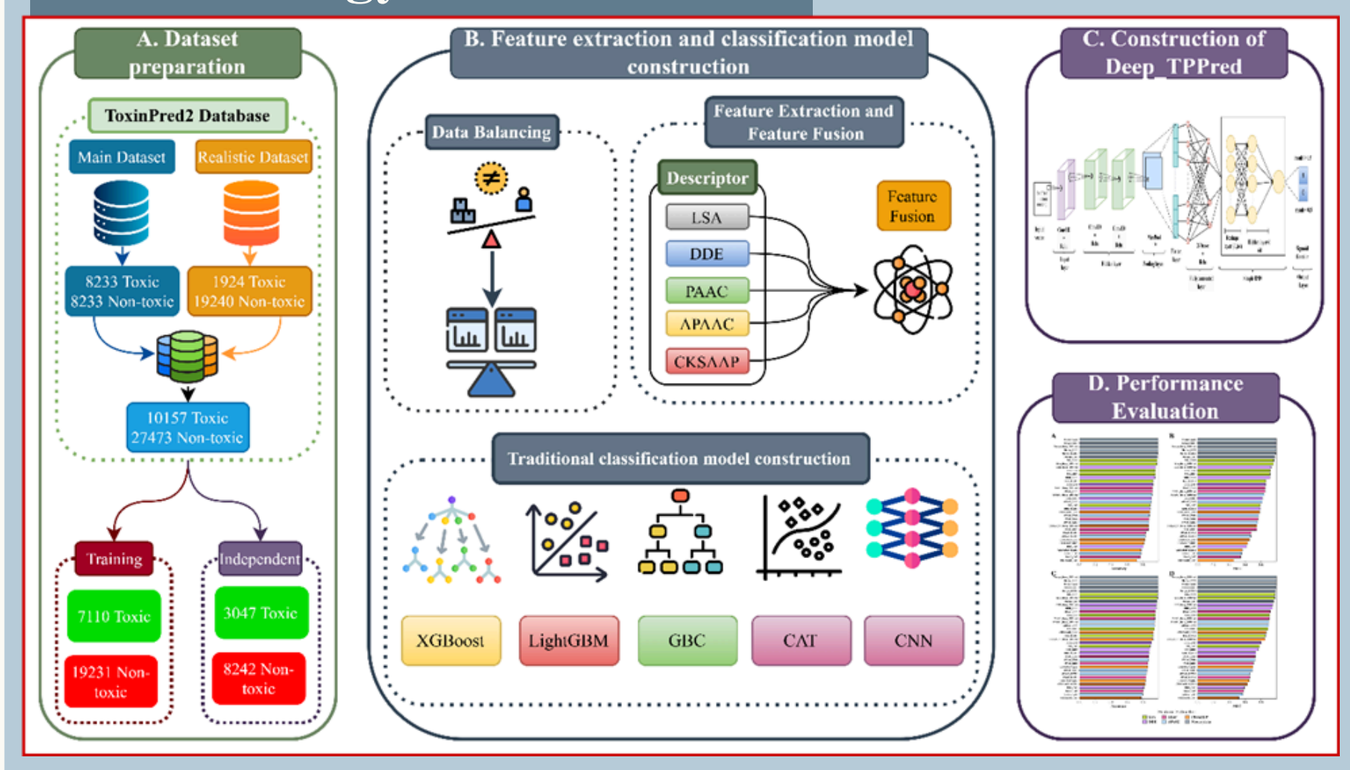


Figure 2: Research methodology and structural architecture used to develop the proposed model

This study used the ToxinPred2 dataset, combining the primary (8,233 toxic and 8,233 non-toxic sequences) and realistic (1,924 toxic and 19,240 non-toxic sequences) datasets into a merged dataset of 10,157 toxic and 27,473 non-toxic sequences. The dataset was split into 70% for training and 30% for evaluation, with Adaptive Synthetic Sampling (ADASYN) used to address class imbalance. Five feature extraction methods (LSA, DDE, PAAC, APAAC, and CKSAAGP) were applied to capture various protein characteristics, and the features were merged through feature fusion to build a robust dataset for toxicity prediction.

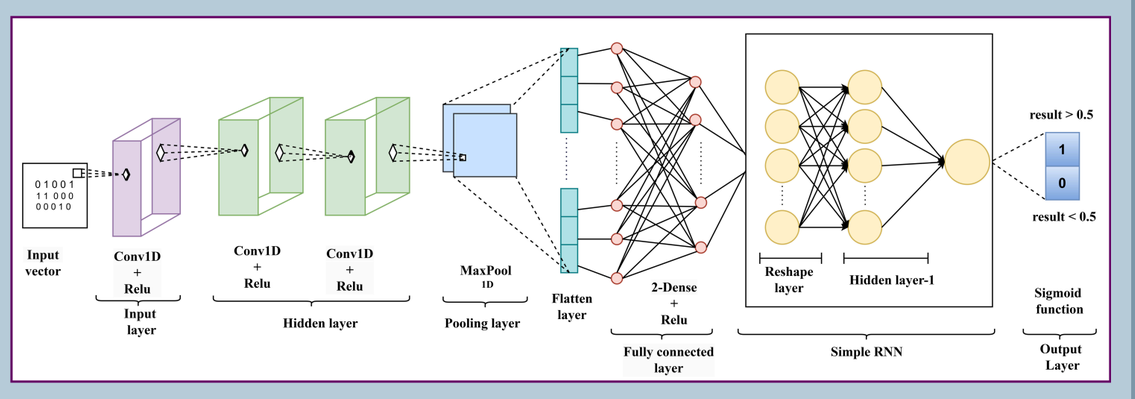


Figure 3: Structural architecture of the proposed model Deep_TPPred.

Results

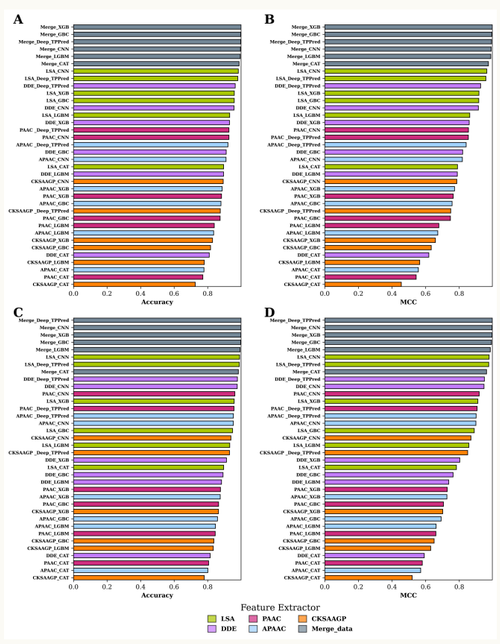


Figure 4 : Comparison of the Accuracy and MCC of the different feature extractors and applied ML algorithm. (A) and (B) for the 5-fold CV result. (C) and (D) for the independent test result.

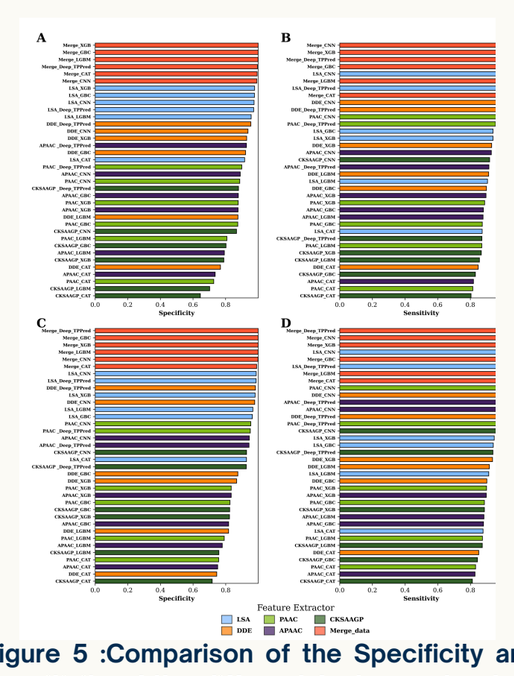


Figure 5 :Comparison of the Specificity and Sensitivity of the different feature extractors and applied ML algorithm. (A) and (B) for the 5-fold CV result. (C) and (D) for the independent test result.

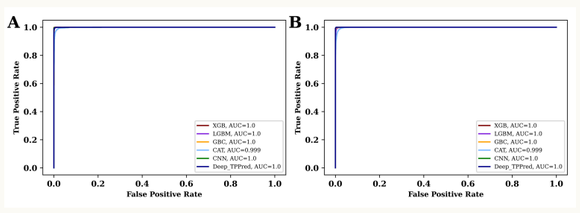


Figure 6 : ROC curve for different ML classifiers on merge dataset. (A) shows the result of a 5-fold CV, and (B) shows the outcome of the independent test.

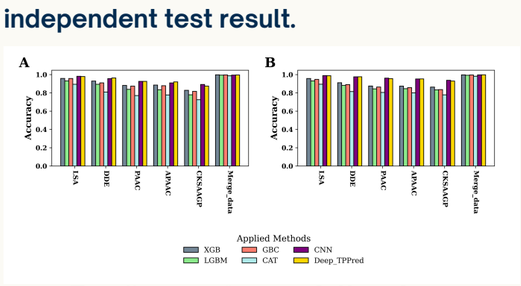


Figure 7 : Comparison of the Accuracy of different feature extractors and applied ML algorithm. (A) for the 5-fold CV result. (B) for the independent test result.

References

1.Gacesa, R., Barlow, D.J. and Long, P.F., 2016. Machine learning can differentiate venom toxins from other proteins having non-toxic physiological functions. PeerJ Computer Science, 2, p.e90.
2.Jain, A. and Kihara, D., 2019. NNTox: gene ontology-based protein toxicity prediction using neural network. Scientific reports, 9(1), p.17923.
3.Pan, X., Zuallaert, J., Wang, X., Shen, H.B., Campos, E.P., Marushchak, D.O. and De Neve, W., 2020. ToxDL: deep learning using primary structure and domain embeddings for assessing protein toxicity. Bioinformatics, 36(21), pp.5159-5168.
4.Sharma, N., Naorem, L.D., Jain, S. and Raghava, G.P., 2022. ToxinPred2: an improved method for predicting toxicity of proteins. Briefings in bioinformatics, 23(5), p.bbac174
5.Shi, H., Li, Y., Chen, Y., Qin, Y., Tang, Y., Zhou, X., Zhang, Y. and Wu, Y., 2022. ToxMVA: An end-to-end multi-view deep autoencoder method for protein toxicity prediction. Computers in Biology and Medicine, 151, p.106322.
6.Morozov, V., Rodrigues, C.H. and Ascher, D.B., 2023. CSM-Toxin: a web-server for predicting protein toxicity. Pharmaceutics, 15(2), p.431.
7.Mall, R., Singh, A., Patel, C.N., Guirimand, G. and Castiglione, F., 2024. VISH-Pred: an ensemble of fine-tuned ESM models for protein toxicity prediction. Briefings in Bioinformatics, 25(4).
8.Vishnoi, S., Matre, H., Garg, P. and Pandey, S.K., 2020. Artificial intelligence and machine learning for protein toxicity prediction using proteomics data. Chemical Biology & Drug Design, 96(3), pp.902-920.

Achievements

This research paper has been officially submitted to the prestigious journal IEEE/ACM Transactions on Computational Biology and Bioinformatics (TCBB). The paper was submitted on January 27, 2025, and is currently under review with the status Major Revisions Required as of April 15, 2025.