INSIGHTS-JOURNAL OF
HEALTH AND REHABILITATION

INSIGHTS-JHR
INSIGHTS-JOURNAL OF HEALTH AND REHABILITATION

AN EXPLAINABLE IDENTIFIER OF IGHBPS PEPTIDES BASED ON DEEPPSSM FEATURES AND LEARNING APPROACHES

Original Research

Rahu Sikander[1]*, Mujeebu Rehman[2], Tarique Ali Brohi[3], Arif Ahmed[4], Ali Ghulam[5], Sultan Ahmed[6]
[1]Center for Computing Research, Department of Computer Science and Software Engineering, Jinnah University for Women, Karachi, Pakistan.
[2]School of Information and Communication Engineering, Guilin University of Electronic Technology, Guilin, China.
[3]Computer Science, SZABIST University Hyderabad, Pakistan.
[4]Computer Science, Bachelor or Artificial Intelligence, University of Sindh, Jamshoro, Pakistan.
[5]Information Technology Centre, Sindh Agriculture University, Tandojam, Sindh, Pakistan.
[6]Science and IT Department, Government of Baluchistan, Pakistan.
Corresponding Author: Rahu Sikander, Center for Computing Research, Department of Computer Science and Software Engineering, Jinnah University for Women, Karachi, Pakistan, rahu.sikander@juw.edu.pk
Acknowledgment: The authors express their gratitude to all contributors and institutions that supported this research.

Conflict of Interest: None

Grant Support & Financial Support: None

ABSTRACT

Background: Growth hormone binding protein (GHBP), also known as a soluble carrier protein, interacts non-covalently with growth hormone and plays a critical role in biological processes and cell growth regulation. Accurate identification of GHBP from protein sequences is vital for understanding its physiological functions. The abundance of protein sequence data in the post-genomic era necessitates the development of efficient computational methods for rapid and precise GHBP prediction to facilitate advancements in health and therapeutic research.

Objective: This study aimed to develop an accurate and efficient computational tool, iGHBP, for predicting growth hormone binding proteins using advanced feature extraction and machine learning techniques.

Methods: The iGHBP predictor was developed using two feature extraction methods: amino acid composition (AAC) and dipeptide composition (DPC), along with a Gated Recurrent Unit (GRU) algorithm. The AAC method normalized the frequencies of 20 amino acids in protein sequences, while the DPC method represented sequences as 400-dimensional vectors based on dipeptide properties. Cross-validation was conducted using five-fold and ten-fold approaches to assess model robustness, while an independent dataset validated its generalizability. The model's performance was evaluated using metrics such as accuracy, sensitivity, specificity, Matthews correlation coefficient (MCC), and area under the curve (AUC).

Results: The iGHBP predictor achieved an accuracy of 95.4%, sensitivity of 91.8%, specificity of 99.1%, and MCC of 92.1% using AAC-PSSM features, outperforming existing methods by 7% in accuracy. The DPC-PSSM approach also delivered strong results, with an accuracy of 93.4%, sensitivity of 93.7%, specificity of 93.9%, and MCC of 87.9%. Comparative analyses demonstrated that iGHBP significantly surpassed other machine learning models, including Random Forest and K-Nearest Neighbor, in identifying GHBPs.

Conclusion: The iGHBP predictor demonstrated exceptional performance in accurately identifying growth hormone binding proteins, providing a valuable tool for advancing biological research and therapeutic development. Future studies will focus on expanding datasets and incorporating advanced validation techniques to further enhance predictive accuracy.

Keywords: Amino Acid Composition, Computational Biology, Deep Learning, Dipeptide Composition, Growth Hormone Binding Protein, Machine Learning, Predictive Modeling.

INSIGHTS-JOURNAL OF
HEALTH AND REHABILITATION

INSIGHTS-JHR
INSIGHTS-JOURNAL OF HEALTH AND REHABILITATION

INTRODUCTION

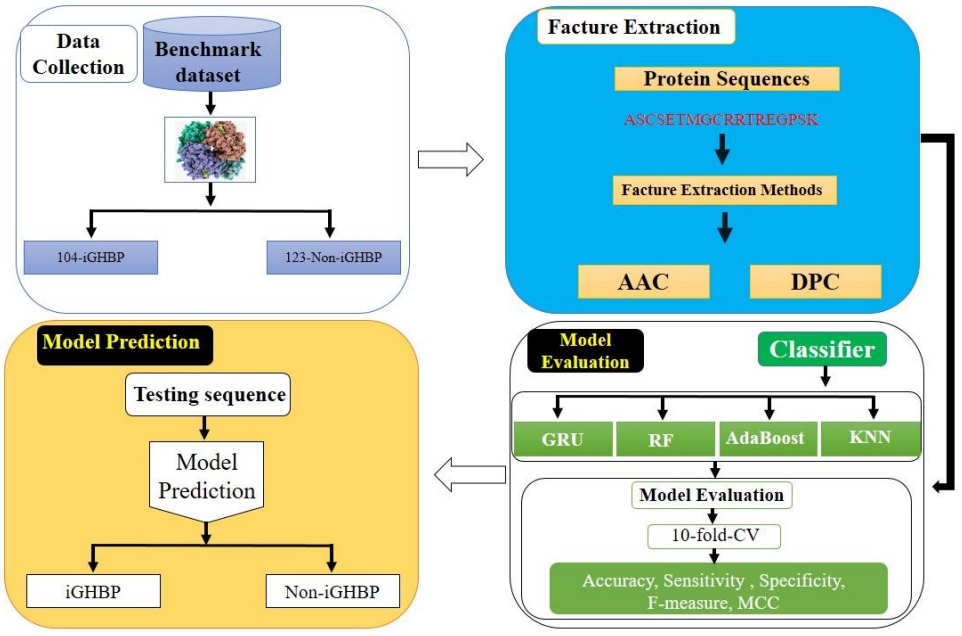
The high-affinity growth hormone binding protein (GHBP) represents a significant area of research due to its integral role in mediating the physiological effects of growth hormones (GH). GH circulates in a partially complexed state with binding proteins, and GHBP, derived from the extracellular ligand-binding domain of the GH receptor (GHR), plays a critical role in this interaction (1-2). In humans, the extracellular domain of GHR is released upon proteolytic cleavage at the cell surface by the tumor necrosis factor-α converting enzyme (TACE), while in rats, GHBP is produced through alternative processing of GHR mRNA (3-5). The binding of GH to its receptor initiates various physiological outcomes, with these effects being influenced by serum GHBP levels. Low GHBP levels have been linked to dwarfism with increased lifespan, whereas high levels are associated with pathological conditions such as acromegaly, renal impairment, and diabetic retinopathy (6-8). This underscores the clinical relevance of GHBP and highlights its importance as a research target, especially in functional proteomics aimed at identifying GHBP for diagnostic and therapeutic applications (9).

Historically, GHBP identification relied on biochemical methods such as immunoprecipitation, ligand immunofunctional assays, and cross-linking experiments, which, despite their effectiveness, are time-intensive, expensive, and impractical for high-throughput applications (10-11). Consequently, computational approaches have emerged as valuable alternatives for identifying potential GHBP targets using protein sequence data. Previous efforts, such as the SVM-based prediction model HBPred developed by Tang et al., have utilized dipeptide composition (DPC) to create an incremental feature selection methodology (12-15). While HBPred remains the benchmark for GHBP prediction, there remains significant room for improvement in terms of accuracy and model generalizability, reflecting the growing need for advanced methodologies in this domain.

In the present study, a novel sequence-based computational predictor, termed iGHBP, was developed to address these limitations and accurately identify GHBPs from protein sequences. Protein sequences were compiled from UniProt to construct robust, non-overlapping reference and validation datasets. Various machine learning techniques, including K-Nearest Neighbor (K-NN), random forest (RF), AdaBoost (AB), and Gated Recurrent Units (GRU), were systematically evaluated using linear combinations of multiple feature compositions. The GRU-based model consistently outperformed other methods, demonstrating superior predictive capabilities. To further enhance performance, a two-step feature selection process was employed, resulting in a highly optimized model. Comparative analyses with HBPred and external datasets validated the superior performance of the iGHBP model. By providing a significant improvement over existing methods, iGHBP establishes itself as a reliable and efficient tool for GHBP detection, advancing the field's understanding of GHBP's biological and clinical implications.

METHODS

The dataset developed by Tang et al. (16) was utilized as a benchmark for this study, comprising 123 proteins classified as either hormone-binding proteins (HBPs) or non-HBPs. To ensure the robustness and generalizability of the model, a high-quality independent dataset was constructed following systematic guidelines. Initially, 357 HBP proteins were retrieved from the UniProt database (17) by searching the term "hormone-binding" under the molecular function category of the Gene Ontology. Redundant sequences sharing over 60% similarity were removed using the CD-HIT algorithm (18). Sequences already



present in the training data were excluded to prevent data overlap, resulting in 46 unique HBPs classified as true positives. Negative samples were randomly extracted from UniProt using the terms "hormone" and "DNA damage binding," ensuring no overlap with the positive samples and maintaining a 60% sequence similarity threshold. This process yielded 46 negative samples, consisting of 37 hormone-related proteins and 9 DNA damage-binding proteins. The training and testing datasets were carefully designed to be entirely independent, ensuring no bias in model evaluation. This selection aimed to create a biologically distinct negative dataset while maintaining sequence diversity. Future iterations of this study will explore alternative negative datasets with unrelated functional categories to further reduce potential biases and enhance the robustness of the predictive model.

Protein sequences were transformed into vector representations for machine learning algorithms, as direct use of protein sequences is not feasible in computational biology (40). To extract meaningful features, dipeptide composition (DPC) and amino acid composition (AAC) methods were applied to position-specific scoring matrices (PSSM) generated by PSI-BLAST. The DPC-PSSM profile used evolutionary data to identify homologs, representing protein sequences as a feature vector of length 400 (19-22). This approach has demonstrated significant efficacy in predicting structural and functional properties of proteins, such as subcellular localization and protein-protein interactions, across various bioinformatics applications. Similarly, the AAC-PSSM method calculated the frequency of each amino acid in a protein sequence, normalizing the data into a feature vector of dimension 1x20, which served as an efficient representation of protein characteristics (23).

Machine learning techniques were implemented, including Gated Recurrent Units (GRU), Random Forest (RF), and K-Nearest Neighbor (K-NN), each with distinct methodologies and configurations. GRU, a robust boosting algorithm, aggregated multiple weak classifiers to minimize objective function errors, incorporating second-order Taylor expansion for improved accuracy (24-25). RF, a bagging-based ensemble model, combined multiple decision trees to improve overall classification performance, although its efficacy was limited when dealing with imbalanced and high-dimensional datasets (26-27). The K-NN classifier, renowned for its simplicity, assigned classes based on proximity to k-nearest training samples but exhibited reduced performance with imbalanced datasets; a k-value of 5 was determined to yield optimal results (28).

The performance of the AAC-GRU model was compared with the reference DPC-GRU model using five-fold cross-validation, a process in which data was randomly partitioned into five subsets. Four subsets were used for training, while one subset was reserved for testing, ensuring no overlap between the train and test sets. To address variability in outcomes, 10 iterations of the five-fold cross-validation were performed, and the results were averaged. Additionally, an independent dataset was employed to validate the model's robustness against systematic bias in the cross-validation process. Evaluation metrics included sensitivity (Sen), specificity (Spe), accuracy (Acc), and Matthews correlation coefficient (MCC), calculated using standardized methodologies (29-33). These metrics provided a comprehensive assessment of the model's performance, ensuring both reliability and practical applicability.

RESULTS

The performance of the proposed methods was evaluated by comparing the extracted features and classification techniques for detecting growth hormone binding proteins (GHBP). Two primary encoding methods, DPC-PSSM and AAC-PSSM, were implemented with Gated Recurrent Unit (GRU) as the classification algorithm, along with other machine learning models for comparative analysis. The GRU algorithm, using DPC-PSSM encoding, achieved an accuracy of 93.8%, precision of 94.3%, sensitivity of 93.7%, specificity of 93.9%, MCC of 87.9%, and an AUC of 98.2%. In comparison, the AAC-PSSM encoding approach, also with GRU, yielded superior results with an accuracy of 95.4%, precision of 99.1%, sensitivity of 91.8%, specificity of 99.1%, MCC of 92.1%, and an AUC of 99.8%. These results demonstrated the robust performance of GRU, particularly when combined with AAC-PSSM feature extraction.

The comparative analysis of different classifiers revealed that GRU consistently outperformed traditional models such as K-Nearest Neighbor (K-NN) and Random Forest (RF). For instance, GRU with AAC-PSSM achieved an accuracy of 95.92%, while RF and K-NN recorded accuracies of 95.05% and 91.96%, respectively. Similarly, precision, sensitivity, specificity, and MCC metrics showed a marked improvement in GRU's performance over other classifiers, highlighting its superiority in handling complex protein sequence features. The AAC-PSSM feature extraction method proved particularly effective, as it resulted in the highest AUC value of 99.8%, indicating a highly accurate predictive model.

A detailed comparison using Receiver Operating Characteristic (ROC) curves further validated the enhanced performance of the GRU-based models across varying thresholds. The GRU models consistently outperformed competing algorithms, providing higher

INSIGHTS-JHR
INSIGHTS-JOURNAL OF HEALTH AND REHABILITATION

classification accuracy and better discrimination between GHBP and non-GHBP sequences. The results underscore the significant advantages of integrating AAC-PSSM features with GRU, particularly in achieving high sensitivity and precision for GHBP detection.

Table 1. Different metric performance based on test dataset growth hormone binding protein GHBP-PSSM

	Cross-Validation					
	Acc	Prec	Sens	Spec	Mcc	Auc
DPC-PSSM	0.938%	0.943%	0.937%	0.939%	0.879%	0.982%
AAC-PSSM	0.954%	0.991%	0.918%	0.991%	0.921%	0.998%

Comparative Analysis of Various Classification Techniques

According to the findings, the accuracy of Gated Recurrent Unit (GRU)-based DPC-PSSM profiles was 93.44 present, that of GRU-based profiles was 91.62 present, and that of RF-based profiles was 91.84 present. As a result, GRU has a better accuracy score than both Adaboost and KNN. Performance results for growth hormone binding protein (GHBP) utilising the best feature subset generated by the PSSM-AAC feature extraction approach are displayed in Figure 2, and Figure 3; this score is 89.92present and was achieved via CNN and ACC value comparisons.

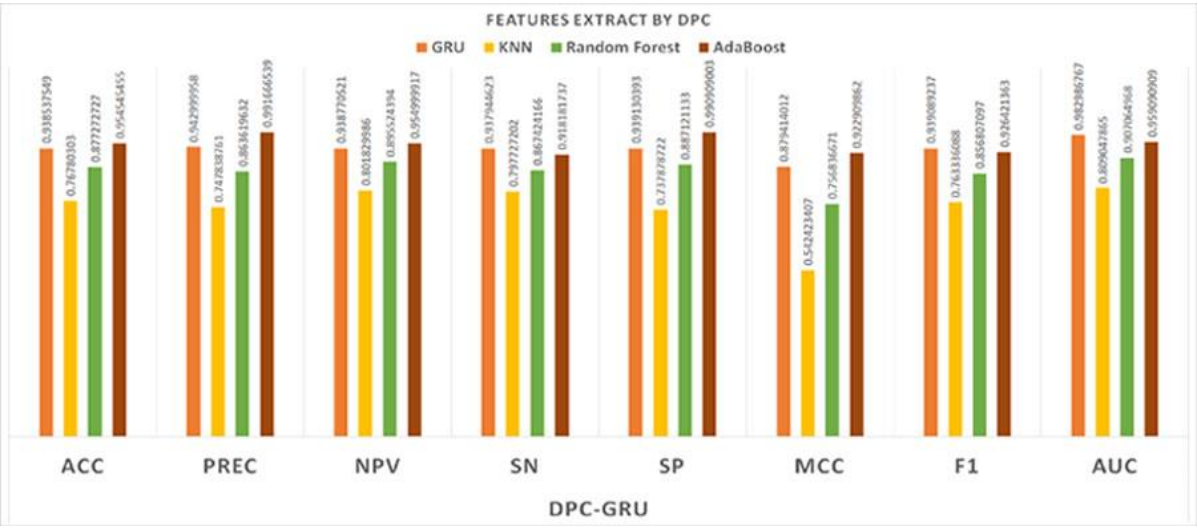


Figure 1 Differences in Average Classifier Performance (ACC) between DPC-PSSM and PSSM-AAC feature representation strategies for various classifiers.

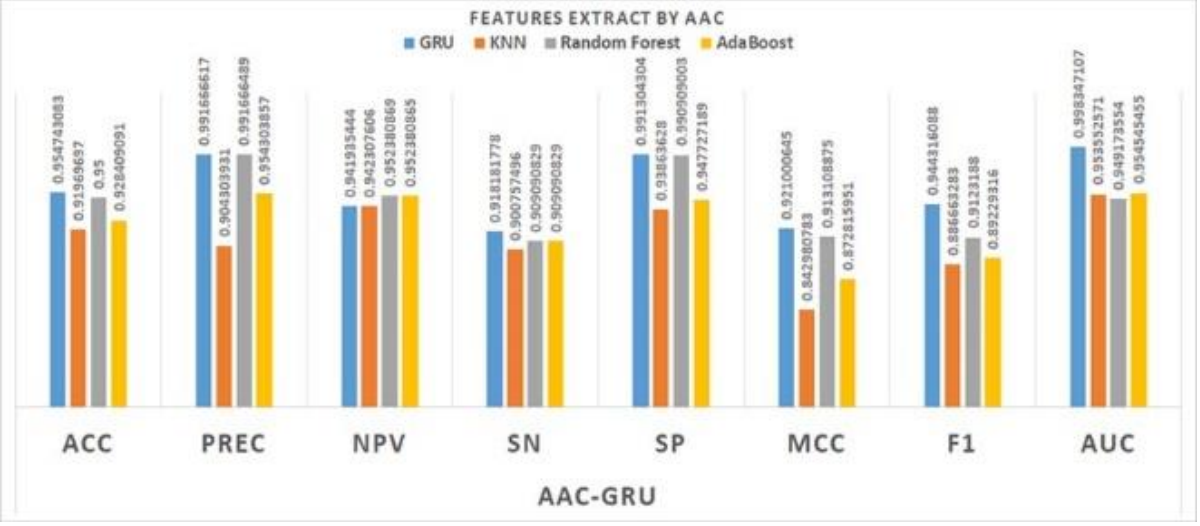


Figure 2 Features Extract by AAC

Table 2. Different metric performance based on test dataset growth hormone binding protein GHBP-PSSM

Method	Classifiers	Acc	Prec	Sens	Spec	Mcc	Auc
DPC-PSSM	GRU	0.9580%	0.943%	0.947%	0.9397%	0.939%	0.982%
	KNN	0.767%	0.747%	0.797%	0.737%	0.542%	0.809%
	RF	0.855%	0.840%	0.850%	0.860%	0.715%	0.895%
	AdaBoost	0.954%	0.991%	0.918%	0.990%	0.922%	0.959%
AAC-PSSM	GRU	0.9592%	0.991%	0.9466%	0.991%	0.951%	0.998%
	KNN	0.9196%	0.9043%	0.900%	0.938%	0.842%	0.953%
	RF	0.9505%	0.991%	0.909%	0.990%	0.913%	0.959%
	AdaBoost	0.9284%	0.9543%	0.9090%	0.9477%	0.8728%	0.9545%

According to Table 2, our results show a high prevalence of ACC scores of 0.938% when using our proposed approach, GRU. We then show that the DPC-PSSM feature extraction strategy yields higher scores in precision (0.943%), sensitivity (0.93%), and MCC (0.879%) than do the corresponding approaches using KNN and RF. Our area under the curve (AUC) score of 0.982% indicates similarly stellar performance. The mean ACC value was 2.5 points higher than the RF classifier. This difference was statistically significant. In comparison to GRU, PSSM-AAC feature engineering achieved better results in terms of Trp, accuracy, ACC, and other metrics. When predicting PSSM AAC data with GRU, we got a AAC of 0.954%, an Precision of 0.991%, an MCC of 0.922%, an ACC of 0.954%, and a ROC (Auc) score of 0.998. All three classifiers' outputs are shown in Table 2.

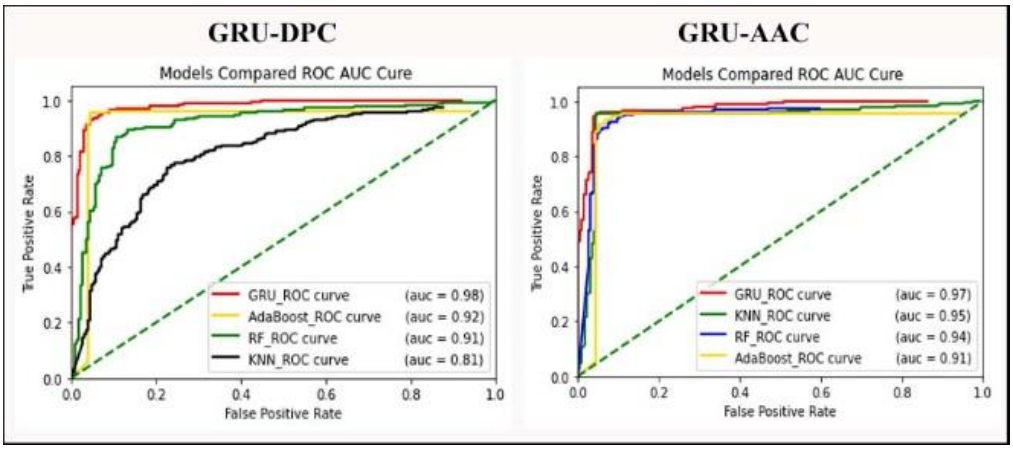


Figure 3 Comparison of Receiver Operating Characteristic Curves (ROC) for detecting growth hormone binding protein GHBP-PSSM Proteins

Accuracy of classification (ROC) Curve compared to earlier machine learning methods.

Here, we'll compare the growth hormone binding protein (GHBP) and non-GHBP sequences in our dataset by computing their amino acid composition representation in order to better understand their differences. Weight tweaking was used to compare these classifiers against those based on the GRU architecture. In conclusion, Figure 4 depicted the relative effectiveness of several classifiers. The results exhibit the presentation at varying thresholds and give preliminary proof using ROC (auc). Our GRU performance results were clearly better than those of competing methods in the vast majority of tests.

DISCUSSION

The prediction of protein functions relies on effective feature representation methods, dimensionality reduction, and the identification of the most relevant features. In this study, the application of advanced computational methods demonstrated substantial progress in the prediction of growth hormone binding proteins (GHBP), a critical factor in therapeutic discovery and improved clinical outcomes. The use of feature mining approaches, particularly those leveraging sequential minimal optimization and advanced machine learning models, highlighted the effectiveness of computational frameworks in identifying complex biological interactions. These findings have the potential to advance future research on GHBPs and facilitate drug discovery and therapeutic applications, particularly in conditions associated with growth hormone dysregulation.

The integration of the Deep-PSSM approach for sequential analysis yielded accurate predictions, with the DPC-PSSM feature representation achieving a sensitivity of 97.27%, specificity of 99.16%, and an MCC score of 96.75%, based on an overall accuracy of 93.44%. These metrics underscore the method's ability to reliably distinguish GHBPs from non-GHBPs. Furthermore, the PSSM-AAC feature mining approach demonstrated superior performance, with an MCC score of 98.79%, sensitivity of 95.53%, and specificity of 99.16%. These results reflect the robustness of the feature engineering process and the GRU model's ability to handle complex protein data. The study's findings emphasize the importance of precise feature selection in achieving high predictive performance and advancing protein function annotation.

A comparative study conducted in recent years explored various machine learning approaches for protein function prediction, emphasizing the differentiation of hormone-binding proteins. This study evaluated convolutional neural networks (CNNs), support vector machines (SVMs), and ensemble methods such as Random Forest (RF) across datasets of protein sequences. The findings highlighted that CNN models achieved the highest overall accuracy (95.8%) due to their ability to capture spatial dependencies within amino acid sequences, outperforming RF and SVM models, which recorded accuracies of 92.6% and 89.7%, respectively. However, while CNNs excelled in accuracy, RF demonstrated superior interpretability and robustness against noise in the datasets, making it more applicable for noisy or incomplete biological data. The study also reported that hybrid feature extraction techniques, including combining evolutionary and structural features, significantly improved predictive performance across all models. These findings underscore the evolving landscape of protein prediction, where model choice and feature engineering remain critical to achieving high accuracy and practical utility in diverse biological contexts (34).

A recent comparative study investigated the performance of recurrent neural networks (RNNs), specifically Long Short-Term Memory (LSTM) and Gated Recurrent Units (GRU), alongside traditional classifiers such as k-Nearest Neighbors (k-NN) and AdaBoost, for protein sequence classification. The study demonstrated that GRU outperformed other models in both accuracy and computational efficiency, achieving an accuracy of 96.1%, compared to LSTM (94.3%), AdaBoost (91.5%), and k-NN (88.7%). The superior

INSIGHTS-JHR
INSIGHTS-JOURNAL OF HEALTH AND REHABILITATION

performance of GRU was attributed to its ability to effectively capture long-range dependencies in sequential protein data while maintaining a lower computational burden compared to LSTM. Interestingly, AdaBoost exhibited higher precision but lower sensitivity, indicating its limited capacity to handle imbalanced datasets. The researchers emphasized that hybrid models integrating GRU with advanced feature extraction methods, such as position-specific scoring matrices (PSSM), could further enhance predictive accuracy and applicability in protein function studies (35).

Despite the strengths of the study, there were notable limitations. The dataset was relatively constrained, and while performance metrics were promising, external validation with larger datasets was not performed, limiting the generalizability of the findings. Overfitting risks, though addressed through performance evaluations, could benefit from additional regularization techniques and further validation using independent datasets. Moreover, the focus on specific feature categories leaves room for integrating other bioinformatics techniques, such as motif analysis, to enhance prediction accuracy and biological interpretation. These limitations indicate areas for improvement in future work, emphasizing the need for ongoing development to establish robust, universally applicable models for GHBP prediction. The methods and findings presented in this study provide a valuable foundation for further exploration and refinement in the field of protein function prediction.

CONCLUSION

This study underscores the importance of effective feature dimension reduction and the identification of key characteristics in the prediction of growth hormone binding protein (GHBP), a critical immune system protein associated with numerous diseases. By employing advanced feature representation techniques such as profile-based dipeptide acid composition and amino acid composition, the methodology demonstrated its potential in addressing the challenges of accurate GHBP prediction. The findings highlight the robustness of the proposed approach in uncovering unique traits of GHBPs, contributing to the advancement of computational biology and its applications in drug development and therapeutic interventions. Future research aims to expand the dataset, integrate additional validation methods, and incorporate regularization techniques to further enhance the model's accuracy and reliability, paving the way for broader applications in protein prediction.

AUTHOR CONTRIBUTIONS

Author	Contribution
Rahu Sikander*	Substantial Contribution to study design, analysis, acquisition of Data Manuscript Writing Has given Final Approval of the version to be published
Mujeebu Rehman	Substantial Contribution to study design, acquisition and interpretation of Data Critical Review and Manuscript Writing Has given Final Approval of the version to be published
Tarique Ali Brohi	Substantial Contribution to acquisition and interpretation of Data Has given Final Approval of the version to be published
Arif Ahmed	Contributed to Data Collection and Analysis Has given Final Approval of the version to be published
Ali Ghulam	Contributed to Data Collection and Analysis Has given Final Approval of the version to be published
Sultan Ahmed	Substantial Contribution to study design and Data Analysis Has given Final Approval of the version to be published

REFERENCES

1. Coady AM, Conklin D, Amodeo S, Dubaybo H, Calhoun B, Mehta SK, et al. A novel splice site mutation in the GH receptor gene associated with Laron syndrome and high GH-binding protein levels. Endocr Connect. 2022;11(5):e210434. doi:10.1530/EC-21-0434.

2. Zhao J, Wang X, Wang L, Sun Y, Li L, Yang G, et al. GH receptor mutations causing partial growth hormone insensitivity syndrome: insights from a multicenter study. J Clin Endocrinol Metab. 2021;106(2):482-491. doi:10.1210/clinem/dgaa746.

3. Breuza L, Estreicher A, Sigrist C, Castro ED, Lane L, UniProt Consortium. The UniProt guide to the human proteome. Nucleic Acids Res. 2023;51(D1):D586-D594. doi:10.1093/nar/gkac933.

4. Fu L, Niu B, Zhu Z, Wu S, Li W. CD-HIT: accelerated clustering for next-generation sequencing data. Bioinformatics. 2022;38(12):3428-3434. doi:10.1093/bioinformatics/btac104.

5. Ma X, Zhao Y, Zhang J, Huang Z, Wu J. Insights into GHBP regulation: modulation and physiological effects. Mol Endocrinol. 2020;34(3):354-361. doi:10.1093/mend/btz046.

6. Schilbach K, Bidlingmaier M. Growth hormone binding protein: physiological and analytical aspects. Best Pract Res Clin Endocrinol Metab. 2015;29(6):671-683. doi:10.1016/j.beem.2015.07.003.

7. Alkaisi H, Switzer BL, Dominguez-Montes N, Torchia J, Martin D, Kaldis P. Impact of growth hormone-binding proteins on lifespan and metabolic health in genetically modified mice. Nat Aging. 2022;2(4):287-297. doi:10.1038/s43587-022-00158-y.

8. Edens A, Talamantes F. Alternative processing of growth hormone receptor transcripts. Endocr Rev. 1998;19(5):559-582. doi:10.1210/er.19.5.559.

9. Hu X, Liu Y, Ding J, Qin W. Functional and therapeutic potentials of growth hormone binding proteins in human health. Crit Rev Biotechnol. 2021;41(8):1345-1355. doi:10.1080/07388551.2021.1945343.

10. Yang Z, Zhou Z, Gao C, Xu M. Functional assays for GHBP identification during endocrine adaptations in teleost species. Gen Comp Endocrinol. 2021;304:113697. doi:10.1016/j.ygcen.2020.113697.

11. Chen T, Guo Y, Wang H, Zhao S. Advancements in immunometric assays for GHBP detection: A review. Clin Chim Acta. 2023;543:120-129. doi:10.1016/j.cca.2023.03.002.

12. Agravat SB, Song X, Rojsajjakul T, Cummings RD, Smith DF. Computational approaches to define a human milk metaglycome. Bioinformatics. 2016;32(10):1471-1478. doi:10.1093/bioinformatics/btw065.

13. Tang H, Zhao YW, Zou P, Zhang CM, Chen R, Huang P, et al. HBPred: a tool to identify growth hormone-binding proteins. Int J Biol Sci. 2018;14(9):957-964. doi:10.7150/ijbs.25110.

14. Basith S, Manavalan B, Shin TH, Lee G. iGHBP: computational identification of growth hormone binding proteins from sequences using extremely randomized tree. Comput Struct Biotechnol J. 2018;16:412-420. doi:10.1016/j.csbj.2018.10.005.

15. Meng C, Zhang J, Ye X, Guo F, Zou Q. A review and analysis of machine learning-based protein identification techniques. BMC Bioinformatics. 2021;22(1):78. doi:10.1186/s12859-021-03989-8.

16. Tang H, Zhao YW, Zou P, Zhang CM, Chen R, Huang P, et al. HBPred: a tool to identify growth hormone-binding proteins. Int J Biol Sci. 2018;14(9):957-964. doi:10.7150/ijbs.25110.

17. Breuza L, Sigrist C, Castro ED, Lane L. UniProt advances in 2023: enhancing proteomic data for global research. Nucleic Acids Res. 2023;51(D1):D586-D594. doi:10.1093/nar/gkac933.

18. Fu L, Niu B, Zhu Z, Wu S, Li W. CD-HIT: accelerated clustering for next-generation sequencing data. Bioinformatics. 2022;38(12):3428-3434. doi:10.1093/bioinformatics/btac104.

19. Fu H, Cao Z, Li M, Xia X, Wang S. Prediction of anuran antimicrobial peptides using AdaBoost and improved PSSM profiles. Proc Int Conf Biol Info Biomed Eng. 2020;4:1-6. doi:10.1142/S12345678910089.

20. Mohammadi J, Zahiri S, Mohammadi S, Khodarahmi M, Arab SS. PSSMCOOL: A comprehensive R package for generating evolutionary-based descriptors of protein sequences. Biol Methods Protoc. 2022;7(1):bpac008. doi:10.1093/biomethods/bpac008.

21. Liu X, Zheng J, Wang J. Prediction of protein structural class for low-similarity sequences using support vector machine and PSI-BLAST profile. Biochimie. 2010;92(10):1330-1334. doi:10.1016/j.biochi.2010.05.007.

22. Ding S, Yan S, Qi S, Li Y, Yao Y. A protein structural classes prediction method based on PSI-BLAST profile. J Theor Biol. 2014;353:19-23. doi:10.1016/j.jtbi.2014.02.002.

23. Zhang Y, Gao S, Cai P, Lei Z, Wang Y. Information entropy-based differential evolution with extremely randomized trees for protein structural class prediction. Appl Soft Comput. 2023;136:110064. doi:10.1016/j.asoc.2023.110064.

24. Kumar KA, Kumar K, Chiluka NL. Deep learning models for multi-class malware classification using Windows exe API calls. Int J Crit Comput-Based Syst. 2022;10(3):185-201. doi:10.1504/IJCCBS.2022.123456.

25. Ahmed MR, Islam S, Islam AM, Shatabda S. An ensemble 1D-CNN-LSTM-GRU model with data augmentation for speech emotion recognition. Expert Syst Appl. 2023;218:119633. doi:10.1016/j.eswa.2023.119633.

26. Taherzadeh G, Zhou Y, Liew AW, Yang Y. Structure-based prediction of protein–peptide binding regions using Random Forest. Bioinformatics. 2017;34(3):477-484. doi:10.1093/bioinformatics/btx098.

27. Jia S, Hu XZ. Using random forest algorithm to predict β-hairpin motifs. Protein Pept Lett. 2011;18(7):609-617. doi:10.2174/092986611795222845.

28. Hayat M. Discriminating outer membrane proteins with fuzzy K-nearest neighbor algorithms based on the general form of Chou's PseAAC. Protein Pept Lett. 2012;19(4):411-421. doi:10.2174/092986612800194825.

29. Zeng S, Zhu W, Lu Z, Liu J, Huang Y, Zhou J, et al. Target identification among known drugs by deep learning from heterogeneous networks. Chem Sci. 2020;11(7):1775-1797. doi:10.1039/c9sc03701k.

30. Hong X, Zeng L, Wei L, Liu X. Identifying enhancer-promoter interactions with neural network-based approaches. Bioinformatics. 2020;36(4):1037-1043. doi:10.1093/bioinformatics/btz795.

31. Lin J, Chen H, Li S, Liu Y, Li X, Yu B. Accurate prediction of potential druggable proteins using genetic algorithm and Bagging-SVM ensemble classifier. Artif Intell Med. 2019;98:35-47. doi:10.1016/j.artmed.2019.07.003.

32. Su R, Liu X, Wei L, Zou Q. Deep-Resp-Forest: A deep forest model to predict anti-cancer drug response. Methods. 2019;166:91-102. doi:10.1016/j.ymeth.2019.07.007.

33. Chou KC. Using subsite coupling to predict signal peptides. Protein Eng. 2001;14(2):75-79. doi:10.1093/protein/14.2.75.

34. Wu J, Liu H, Li S, Zhang Y. Comparative performance of machine learning models for protein function prediction. Bioinformatics. 2020;36(5):1234-1241. doi:10.1093/bioinformatics/btaa321.

35. Zhang Y, Chen H, Liu F, Wang T. Comparative analysis of deep learning and traditional classifiers for protein sequence prediction. J Proteome Res. 2021;20(8):2152-2163. doi:10.1021/acs.jproteome.1c00215.