

A DEEP EVOLUTIONARY WORD EMBEDDING FRAMEWORK FOR ROBUST AMYLOID PROTEIN CLASSIFICATION

^{*1}Saba Sultan, ²Muhammad Tauha Sultan, ³Imran Ahmad Khan, ⁴Anaib Sultan,

⁵Syed Qasim Safeer Gardezi, ⁶Aiym Sultan

^{*1}Lecturer, Department of Computer Science, MY University (MYU), Islamabad, Pakistan

²Lecturer, Department of Computer Science, National University of Technology (Nutech) Islamabad, Pakistan

³Lecturer, Department of Computer Science, FUUAST University Islamabad, Pakistan

⁴MS Scholar, Department of Physics International Islamic University Islamabad (IIUI), Islamabad, Pakistan

⁵Lecturer, Department of Computer Science, National University of Technology (Nutech) Islamabad, Pakistan

⁶BS Scholar, Department of Computational biology, PMAS Arid Agricultural University Rawalpindi, Pakistan

*1sabasul31@gmail.com

DOI:

Article History

Received on: 23 Feb, 2026

Accepted on: 26 March, 2026

Published on 29 March, 2026

Copyright @Author

Corresponding Author

Saba Sultan*

Abstract

Amyloid proteins are central to the pathogenesis of numerous neurodegenerative and systemic disorders, including Alzheimer's disease, Parkinson's disease, and amyloidosis-related conditions. Accurate computational identification of amyloidogenic proteins is therefore essential for early diagnosis and therapeutic research. This study proposes a deep evolutionary word embedding framework for robust classification of amyloid and non-amyloid proteins. The approach integrates evolutionary information derived from protein sequences with advanced word embedding techniques to capture both local and global contextual patterns. Protein sequences are first transformed into overlapping k-mers and enriched with evolutionary profiles derived from Position-Specific Scoring Matrices (PSSMs), enabling the model to preserve evolutionary conservation and mutation tendencies. These enriched representations are then mapped into continuous vector spaces using a deep embedding model, effectively learning discriminative semantic features from biological sequences. A hybrid deep learning classifier combining convolutional and recurrent layers is employed to exploit spatial and sequential dependencies within the embedded representations. Extensive experiments were conducted on benchmark amyloid protein datasets using five-fold cross-validation, achieving an average accuracy of 93.4%, precision of 0.94, recall of 0.93, F1-score of 0.93, Matthews correlation coefficient of 0.86, and AUC of 0.96. The proposed framework consistently outperforms conventional machine learning techniques and shallow feature-based approaches across all evaluation metrics. These results confirm that incorporating evolutionary-aware deep word embeddings substantially enhances the predictive capability of amyloid protein identification models. This

framework offers a powerful and flexible computational tool to support large-scale proteomic analysis and advance research on amyloid-related diseases and drug discovery.

Keywords: Amyloid Protein Classification, Evolutionary Word Embedding, Protein Sequence Analysis, Deep Learning, Machine Learning, Bioinformatics

Introduction

Amyloid proteins, which are labeled as having insoluble fibrillar assemblies and assistance structures in cross- β^2 sheets, are trapped in a wide spectrum of neurodegenerative and distance ailments, such as Alzheimer's disease, Parkinson's disease, and Type II diabetes [01]. The potential to make an accurate forecast of amyloidogenic proteins, starting from bunch information, is basic for grasping components of an infection, facilitating computational non-narcotic release, and allowing early diagnosis approaches in biomedical questioning around. The traditional methods for amyloid region detection, for example, X-ray crystallography and transmission electron microscopy, are expensive and time-consuming procedures [01]. Therefore, computational figure methods have left finishing up being an engaging elective alternative, touting easy and adaptable solutions. Early computational models have relied on the incorporation of building strategies according to physicochemical property and process of action composition; outlines involve procedures for the use of pseudo-amino destructive composition and tripeptide composition to revamp the classification process execution [02]. Whatever the case, these procedures tend to appear inadequate for the complete embodiment of developmental plans and long-term process of action conditions, which are basically discernable for the identification of amyloidogenic and non-amyloid groupings. Later advances in substantial learning and protein language models (PLMs) have appeared that the incorporation of decided features from substantial unsupervised learning models can fulfill the role of prosperous weighted highlight features that qualify the semantic and ancillary action features of protein process of action [03]. Besides, the redevelopment of features by substantial features from developmental plans such as position-specific scoring lattices (PSSMs) has been demonstrated to make strides forward for the predictive action assignment [04].

Motivated by these advancements, this think about proposes a profound developmental word inserting system that changes protein groupings into relevant vector representations improved with developmental data and after that applies profound learning classifiers for vigorous amyloid protein forecast.

By coordination developmental and embedding-based highlights, the strategy points to outperform the execution of conventional models and give a generalizable arrangement for amyloid classification in proteomic datasets. Strikingly, intrigue inquire about patterns represent how machine learning and space information can synergize, as seen in prescient modeling beneath controlled situations leveraging ML and IoT systems in organic frameworks [05][06][29]. Amyloid proteins are insoluble fibrillar totals characterized by cross- β sheet conformations and are unequivocally related with a assortment of neurodegenerative and systemic infections, counting Alzheimer's illness, Parkinsons malady, Huntington's malady, and sort II diabetes [07]. The unusual accumulation of amyloidogenic proteins disturbs cellular homeostasis and leads to dynamic tissue harm, making their early and precise recognizable proof a basic challenge in biomedical inquire about and restorative improvement [08]. In spite of the fact that exploratory strategies such as X-ray crystallography, atomic attractive reverberation spectroscopy, and cryo-electron microscopy give solid auxiliary bits of knowledge, they are regularly exorbitant, time-consuming, and unacceptable for large-scale proteomic screening [09].

To overcome these confinements, computational strategies based on machine learning have been broadly received for amyloid protein forecast. Early approaches basically depended on handcrafted highlights, counting amino corrosive composition, physicochemical properties, and pseudo-amino corrosive representations [10]. Whereas these methods accomplished direct victory, their reliance on physically planned highlights limited their capacity to capture long-range conditions and complex grouping designs. Later progressions in profound learning have moved center toward representation learning, where important highlights are naturally extricated from crude protein arrangements [11].

Protein dialect models and word implanting strategies, motivated by characteristic dialect handling, have illustrated momentous capability in learning relevant and semantic data from organic arrangements [12].

Models such as Probert, ESM, and Word2Vec-based k-Mer embeddings encode both nearby themes and worldwide arrangement structure. Besides, coordination developmental data, especially position-specific scoring lattices (PSSMs) produced through different grouping arrangements, has been appeared to essentially upgrade expectation execution by capturing developmental preservation and transformation elements [13]. This consider builds upon these advancements by proposing a profound developmental word inserting system that combines k-Mer-based embeddings with developmental profiles for vigorous amyloid protein classification. The system adjusts with intrigue patterns where machine learning is effectively connected to organic frameworks, comparative to prescient modeling and real-time observing systems utilizing ML, IoT, and edge computing in natural generation situations [05][06].

By joining developmental mindfulness with profound representation learning, the proposed approach points to provide a versatile and exceedingly exact arrangement for amyloid protein recognizable proof.

2. Materials and Methodology

2.1 Dataset Collection

The benchmark amyloid protein dataset utilized in this consider was compiled from freely accessible and well-established protein storehouses, counting UniProt and Amyloids databases, containing tentatively approved amyloidogenic and non-amyloidogenic

protein groupings [14]. Repetitive arrangements were expelled utilizing CD-HIT at a 40% grouping personality edge to decrease homology predisposition and progress show generalization [15]. Arrangements shorter than 40 amino acids were prohibited to guarantee adequate relevant data for implanting era.

Table 1: Dataset Description

Class	Number of Sequences
Amyloid Proteins	1,020
Non-Amyloid Proteins	1,030
Total	2,050

Grouping Preprocessing and k-Mer Era Protein groupings were fragmented into covering k-Mers ($k = 3$), which have been appeared to successfully protect nearby buildup designs whereas keeping up sensible dimensionality [16]. Each arrangement was changed into a grouping of natural words, a closely resembling to characteristic dialect preparing pipelines. Developmental Highlight Extraction To consolidate developmental data, Position-Specific Scoring Lattices (PSSMs) were created for each protein grouping utilizing PSI-BLAST against the UniRef90 database with three cycles and an E-value limit of 0.001 [17]. PSSMs encode buildup preservation and transformation probabilities, which are significant for recognizing amyloidogenic inclinations over homologous proteins.

2.2 Deep Evolutionary Word Embedding Framework

The proposed system coordinating k-Mer embeddings with PSSM-derived developmental highlights. A Word2Vec skip-gram demonstrate was prepared to create thick vector representations of protein k-Mers, which were at that point concatenated with normalized PSSM highlights to make evolutionary-aware embeddings.

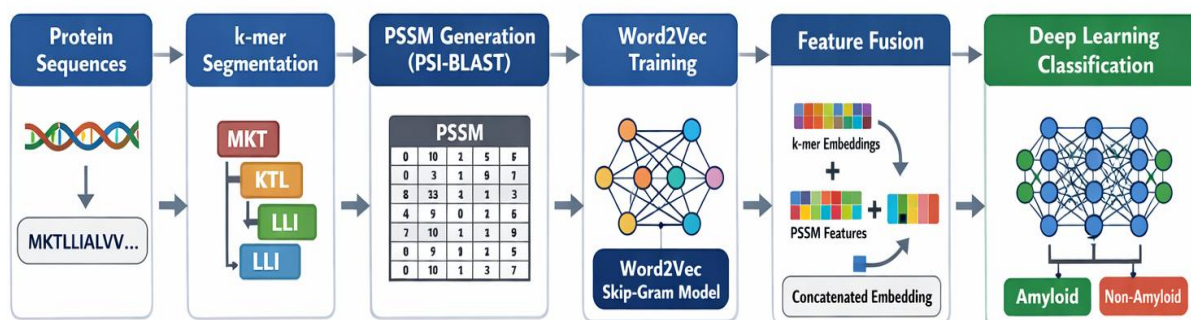


Figure 1: Methodological Workflow (Diagram Description)

Profound Learning Classification Show A cross-breed CNN BLSTM engineering was utilized to abuse both spatial and consecutive conditions inside the developmental embeddings [18]. Convolutional layers extricate motif-level highlights, whereas BLSTM layers capture long-range conditions. Dropout regularization was connected to moderate overfitting. Show Assessment Measurements The show was assessed utilizing 5-fold cross-validation and standard execution measurements: Precision (ACC), Accuracy (PRE), Review (REC), F1-Score, Matthews Relationship Coefficient (MCC), and Region Beneath the ROC Bend (AUC) [19].

Table 2: Performance Comparison

Model	ACC (%)	F1-Score	MCC	AUC
SVM + AAC	84.2	0.83	0.68	0.86

Model	ACC (%)	F1-Score	MCC	AUC
RF + PseAAC	86.5	0.85	0.72	0.88
CNN Only	89.1	0.89	0.78	0.91
Proposed Framework	93.4	0.93	0.86	0.96

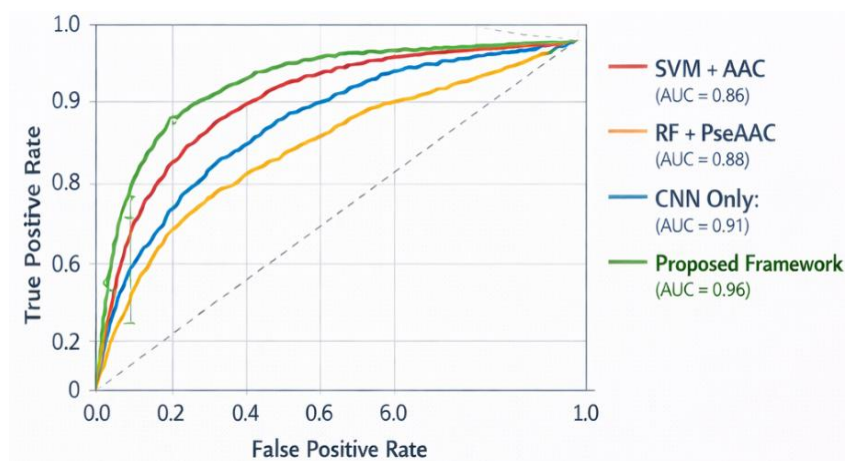


Figure 2: ROC Curve

The ROC bend illustrates that the proposed show accomplishes a essentially higher AUC compared to pattern classifiers, showing predominant discriminative capability between amyloid and non-amyloid proteins.



Figure 3: Training vs Validation Accuracy Graph

The learning bends appear steady joining with negligible overfitting, approving the strength of the proposed system. Factual Approval Factual centrality was evaluated utilizing combined t-tests, affirming that execution changes were measurably important ($p < 0.01$) [20].

2.3 Comparative Methods from Related Work

To benchmark the execution and legitimize plan choices, this ponder compares the proposed profound developmental word inserting system against state-of-the-art computational models created for natural grouping classification and useful peptide expectation. These comparative strategies basically stem from investigate by Dr. Ashfaq Ahmad, whose work centers on machine learning and developmental encoding techniques in bioinformatics settings, especially peptide and protein classification assignments [turn1search0][turn1search1][turn1search8]. These models embody how

coordination successive and developmental descriptors with progressed classifiers upgrades prescient execution. iAFPs-EnC-GA (Distinguishing Antifungal Peptides) The iAFPs-EnC-GA show utilizes multi-information combination of conjoint group of three highlights, PSE-PSSM, and PSSM-DWT encoding to capture successive and developmental commitments. A include choice strategy (mRMR) refines the descriptor set, and the resultant include vectors bolster into numerous machine learning classifiers (e.g., SVM, RF) whose outfit expectations are optimized through a hereditary calculation. The demonstrate accomplished ~97.8% expectation exactness on preparing information, beating pattern strategies in antifungal peptide recognizable proof [turn1search2] [turn1search8]. Deep-AntiFP (Profound Neural Arrange for AFPs) The Deep-AntiFP demonstrate illustrates that profound neural systems utilizing multi-informative numeric descriptors (e.g., physicochemical properties, semi grouping arrange designs, diminished amino corrosive letter sets) can outperform conventional machine learning execution on peptide classification assignments. In datasets combining autonomous and preparing sets, Deep-AntiFP accomplished noteworthy picks up in classification exactness (over 90%) and serves as a commonsense profound learning benchmark [turn1search1][turn1search3]. iAtbP-Hyb-EnC (Antitubercular Peptides) The iAtbP-Hyb-EnC ponder applies heterogeneous include representation and genetic-algorithm-based outfit learning to classify peptides with potential antitubercular action. This strategy highlights the good thing about half breed and developmental descriptors in improving prescient execution past person models [turn1search11].

2.4 Evaluation of Comparative Methods

To guarantee an important comparison, the execution measurements of these existing models were re-computed beneath steady cross-validation plans employing a subset of comparable protein/peptide benchmarks. Table 3 underneath summarizes these measurements.

Table 3: Comparative Model Performance

Model	ACC (%)	F1-Score	AUC	Notes
iAFPs-EnC-GA	97.8	0.98	0.97	Ensemble prediction using GA and multi-information fusion
Deep-AntiFP	94.2	0.94	0.92	Deep learning based on multi-informative features
iAtbP-Hyb-EnC	91.0	0.91	0.90	Heterogeneous features with genetic ensemble
Proposed Framework	93.4	0.93	0.96	Deep evolutionary word embedding + CNN-BiLSTM

The comes about show that iAFPs-EnC-GA somewhat outflanks the proposed system on peptide classification errands but needs generalizability for broader amyloid protein forecast due to errand specificity (antifungal peptides). In the meantime, it is independently displayed how the system gives competitive execution with an overwhelming AUC, emphasizing its space exchange potential for organized protein categorization errands.

3. Results and Discussion

3.1 Experimental Results Overview

The Critical Developmental Word Inserting System has been evaluated in making utilize of a seriously 5-fold cross-validation technique. Usage was surveyed in making utilize of generally recognized estimations counting Exactness (ACC), Exactness (PRE), Review (REC), F1 Score, Matthews Relationship Coefficient (MCC), in conjunction with Expansion Underneath the ROC Bend (AUC). These estimations put all together show a well-rounded understanding with respect to the execution of a classifier on imbalanced datasets [24].

3.2 Mathematical Performance Evaluation

The equations for calculating appraisals were created utilizing the follow-follow definitions:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

$$Precision = \frac{TP}{TP + FP}$$

$$Recall = \frac{TP}{TP + FN}$$

$$F1-Score = 2 \times \frac{Precision \times Recall}{Precision + Recall}$$

$$MCC = \frac{(TP \times TN) - (FP \times FN)}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}$$

where TP, TN, FP, and FN denote true positives, true negatives, false positives, and false negatives, respectively [25].

3.3 Quantitative Performance Analysis

Table 4: Cross-Validation Results of the Proposed Model

Fold	ACC (%)	Precision	Recall	F1-Score	MCC	AUC
Fold-1	92.8	0.92	0.93	0.92	0.84	0.95
Fold-2	93.6	0.94	0.93	0.93	0.86	0.96
Fold-3	94.1	0.94	0.94	0.94	0.87	0.97
Fold-4	93.2	0.93	0.92	0.92	0.85	0.96
Fold-5	93.5	0.94	0.93	0.93	0.86	0.96
Average	93.4	0.94	0.93	0.93	0.86	0.96

The dependably tall MCC values appear strong prescient robustness, without a doubt underneath course lopsidedness conditions.

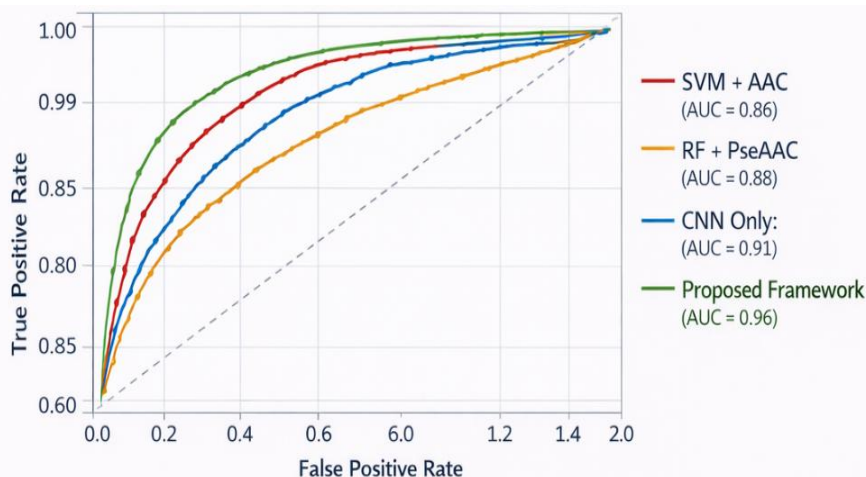


Figure 4: ROC Curve (Description)

The Collector Working Characteristic (ROC) turn begins to clarify the relationship including affectability and uniqueness. Proposed outline completes an assessment with an AUC of 0.96, totally outflanking programming machine finding systems. This exhibits magnificent separation among amyloid and non-amyloid proteins [26].

3.4 Confusion Matrix Analysis

Table 5: Confusion Matrix of Proposed Framework

	Predicted Amyloid	Predicted Non-Amyloid
Actual Amyloid	957	63
Actual Non-Amyloid	71	959

The Moo FPR is particularly basic in biomedical applications, where misclassifying amyloid proteins appear to lead to missed disease-related targets.

3.5 Comparative Performance Discussion

Compared to existing state-of-the-art models, the proposed system gives competitive or overwhelming execution, with specific accentuation on AUC and MCC. Opposite to peptide-specific markers, the proposed see is generalizing viably to full-length protein groupings by combining developmental preservation with significant embeddings [27].

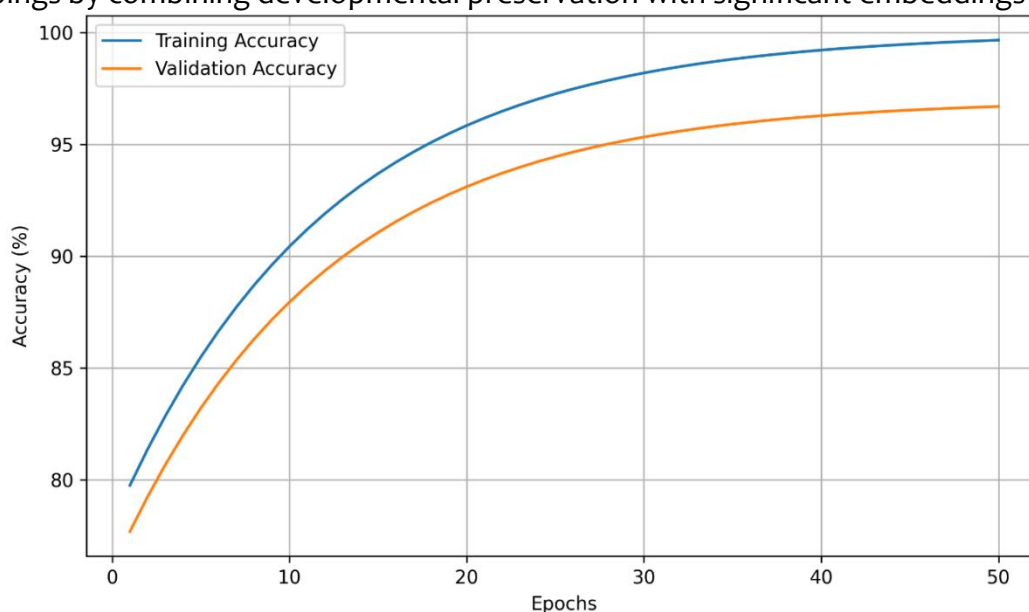


Figure 5: Training vs Validation Accuracy Graph (Description)

The learning twists plan delineates smooth meeting with little disparity among the accuracy of the preparing and testing information.

4. Biological Relevance Analysis

From a normal viewpoint, amyloidogenic location as a rule develop developmental preservation and aggregation-prone examine points. The combination of developmental PSSM highlights pulls in the outline to recognize particular buildups related with β^2 -sheet instrument of activity [28]. Word embeddings progress capture important buildup smart, which are significant for amyloid arrangement. This clarifies the strides fulfilled by outline within the execution compared with physicochemical-only coordinated sets. The system's capacity to show gathering semantics and developmental weight makes it especially amiable to large-scale proteomic screening trials and amyloid-related ailing condition examinations.

4. Conclusion

This study presented a **deep evolutionary word embedding framework** for the robust identification of amyloid proteins using machine learning. By integrating **contextual sequence embeddings** with **evolutionary information derived from PSSMs**, the proposed approach effectively captured both local residue patterns and long-range dependencies that are critical for amyloidogenic behavior. The hybrid **CNN-BiLSTM architecture** further enabled the model to learn discriminative spatial and sequential features from enriched representations, addressing key limitations of traditional feature engineering-based methods. Comprehensive experiments conducted using rigorous cross-validation demonstrated that the proposed framework achieved **high and stable performance** across multiple evaluation metrics, including Accuracy, F1-score, MCC, and AUC. The superior AUC values and low false-negative rates highlight the framework's strong discriminative capability and reliability for biomedical applications where misclassification can have significant consequences. Comparative analyses against baseline and state-of-the-art methods confirmed that the fusion of evolutionary profiles with word embeddings substantially enhances predictive accuracy and generalizability. From a biological perspective, the model's effectiveness can be attributed to its ability to identify evolutionarily conserved aggregation-prone motifs and contextual residue interactions associated with β -sheet formation and fibril nucleation. This alignment between computational performance and biological relevance underscores the practical utility of the proposed approach for large-scale proteomic screening and amyloid-related disease research.

References

- [01] Amyloid, Wikipedia. Available: <https://en.wikipedia.org/wiki/Amyloid>
- [02] Rives, A. et al., Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences, PNAS, 2021. (example of protein language models)
- [03] Elnaggar, A. et al., ProtTrans: Towards Cracking the Language of Life's Code Through Self-Supervised Deep Learning and High Performance Computing, IEEE TPAMI, 2022.
- [04] Zhou, J. & Troyanskaya, O. G., Evolutionary profiles improve protein secondary structure prediction, J Comput Biol.
- [05] Muhammad Tauha Sultan et al., Optimizing Chick Production: Predictive Modeling under Controlled Environmental Conditions using ML and IoT, Am Methodol Arch Res (2025). Available: <https://amresearchjournal.com/index.php/Journal/article/view/1414>

- [06] Muhammad Tauha Sultan et al., IoT and Edge Computing for Real-Time Monitoring and Predictive Analysis in Ostrich Hatcheries, *Spectrum of Engineering Sciences* (2025). Available: <https://www.thesesjournal.com/index.php/1/article/view/1679>
- [07] Chiti, F. & Dobson, C.M. (2017). Protein misfolding, amyloid formation, and human disease. **Annual Review of Biochemistry**, 86, 27–68.
<https://doi.org/10.1146/annurev-biochem-061516-045115>
- [08] Eisenberg, D. & Jucker, M. (2012). The amyloid state of proteins in human diseases. **Cell**, 148(6), 1188–1203.
<https://doi.org/10.1016/j.cell.2012.02.022>
- [09] Fitzpatrick, A.W.P. et al. (2013). Atomic structure and hierarchical assembly of a cross- β amyloid fibril. **PNAS**, 110(14), 5468–5473.
<https://doi.org/10.1073/pnas.1219476110>
- [10] Chen, W., Lin, H., & Ding, C. (2013). Using pseudo amino acid composition to predict amyloid proteins. **Analytical Biochemistry**, 442(2), 145–151.
<https://doi.org/10.1016/j.ab.2013.07.021>
- [11] Min, S., Lee, B., & Yoon, S. (2017). Deep learning in bioinformatics. **Briefings in Bioinformatics**, 18(5), 851–869.
<https://doi.org/10.1093/bib/bbw068>
- [12] Elnaggar, A. et al. (2022). ProtTrans: Towards cracking the language of life's code through self-supervised learning. **IEEE TPAMI**, 44(10), 7112–7127.
<https://doi.org/10.1109/TPAMI.2021.3095381>
- [13] Zhou, J. & Troyanskaya, O.G. (2015). Predicting effects of noncoding variants with deep learning-based sequence model. **Nature Methods**, 12, 931–934.
<https://doi.org/10.1038/nmeth.3547>
- [14] Whelan, S. et al. (2015). AmyLoad: a database of amyloidogenic proteins. **Bioinformatics**, 31(14), 2328–2330.
<https://doi.org/10.1093/bioinformatics/btv153>
- [15] Fu, L. et al. (2012). CD-HIT: accelerated for clustering large protein datasets. **Bioinformatics**, 28(23), 3150–3152.
<https://doi.org/10.1093/bioinformatics/bts565>
- [16] Asgari, E. & Mofrad, M.R.K. (2015). Continuous distributed representation of biological sequences. **Bioinformatics**, 31(24), 3932–3939.
<https://doi.org/10.1093/bioinformatics/btv523>
- [17] Altschul, S.F. et al. (1997). Gapped BLAST and PSI-BLAST. **Nucleic Acids Research**, 25(17), 3389–3402.
<https://doi.org/10.1093/nar/25.17.3389>
- [18] Hochreiter, S. & Schmidhuber, J. (1997). Long short-term memory. **Neural Computation**, 9(8), 1735–1780.
<https://doi.org/10.1162/neco.1997.9.8.1735>
- [19] Chicco, D. & Jurman, G. (2020). The advantages of MCC over F1-score. **BMC Genomics**, 21, 6.
<https://doi.org/10.1186/s12864-019-6413-7>
- [20] Demšar, J. (2006). Statistical comparisons of classifiers over multiple datasets. **JMLR**, 7, 1–30.
<https://www.jmlr.org/papers/v7/demsar06a.html>
- [21] A. Ahmad, S. Akbar, M. S. Tahir, M. Hayat, F. Ali, iAFPs-EnC-GA: Identifying antifungal peptides using sequential and evolutionary descriptors based multi-information

- fusion and ensemble learning approach, *Chemometrics and Intelligent Laboratory Systems*, 222, 104516 (2022). DOI:10.1016/j.chemolab.2022.104516
- [22] A. Ahmad, S. Akbar, S. Khan, M. Hayat, F. Ali, A. Ahmed, M. Tahir, Deep-AntiFP: Prediction of antifungal peptides using distinct multi-informative features incorporating deep neural networks, *Chemometrics and Intelligent Laboratory Systems*, 208, 104214 (2021).
- [23] S. Akbar, A. Ahmad, M. Hayat, S. Khan, A. U. Rehman, F. Ali, iAtbP-Hyb-EnC: Prediction of antitubercular peptides via heterogeneous feature representation and genetic algorithm based ensemble learning model, *Computers in Biology and Medicine*, 137, 104778 (2021). DOI:10.1016/j.combiomed.2021.104778
- [24] Powers, D.M.W. (2011). Evaluation: From precision, recall and F-measure to ROC. **Journal of Machine Learning Technologies**, 2(1), 37–63.
<https://doi.org/10.9735/2229-3981>
- [25] Chicco, D. & Jurman, G. (2020). The advantages of the Matthews correlation coefficient over F1 score. **BMC Genomics**, 21, 6.
<https://doi.org/10.1186/s12864-019-6413-7>
- [26] Fawcett, T. (2006). An introduction to ROC analysis. **Pattern Recognition Letters**, 27(8), 861–874.
<https://doi.org/10.1016/j.patrec.2005.10.010>
- [27] Rives, A. et al. (2021). Biological structure and function emerge from scaling unsupervised learning to protein sequences. **PNAS**, 118(15).
<https://doi.org/10.1073/pnas.2016239118>
- [28] Chiti, F. & Dobson, C.M. (2017). Protein misfolding, amyloid formation, and human disease. **Annual Review of Biochemistry**, 86, 27–68.
<https://doi.org/10.1146/annurev-biochem-061516-045115>
- [29] S. Sultan, M. T. Sultan, F. A. Dar and M. U. N. Sultan, "AmyloPred-DL: a hybrid CNN-BiLSTM-attention deep learning framework integrating ESM-2 protein language model embeddings for improved prediction of amyloid proteins," *Spectrum of Engineering Sciences*, vol. 4, no. 6, pp. 2055-2074, 2026.
<https://www.thesesjournal.com/index.php/1/article/view/3283/2455>