scientific reports

Check for updates

OPEN

A deep learning model for epidermal growth factor receptor prediction using ensemble residual convolutional neural network

Wajdi Alghamdi[1], Farman Ali[2]✉, Raed Alsini[3], Amal Babour[4], Naif Waheb Rajkhan[5] & Tamim Alkhalifah[6]✉

Epidermal growth factor receptor (EGFR) overexpression is a key oncogenic driver in breast cancer, making it an important therapeutic target. Conventional approaches for EGFR identification, including motif- and homology-based methods, often lack accuracy and sensitivity, while experimental assays such as immunohistochemistry are costly and variable. To address these limitations, we propose a novel deep learning–based predictor, ERCNN-EGFR, for the accurate identification of EGFR proteins directly from primary amino acid sequences. Protein features were extracted using composition distribution transition (CDT), amphiphilic pseudo amino acid composition (AmpPseAAC), k-spaced conjoint triad descriptor (KSCTD), and ProtBERT-BFD embeddings. To reduce redundancy and enhance discriminative power, features were refined using XGBoost-Feature Forward Selection (XGBoost-FFS) approach. Multiple deep learning frameworks, including Bidirectional Long Short-Term Memory (BiLSTM), Gated Recurrent Unit (GRU), Generative Adversarial Network (GAN), and Ensemble Residual Convolutional Neural Network (ERCNN), were evaluated. Among them, ERCNN demonstrated Superior performance, achieving 93.48% accuracy, 94.53% sensitivity, 92.58% specificity, and a Matthews correlation coefficient of 0.816 after feature selection, and maintained robust performance on an independent test set (82.85% accuracy). Ablation analysis confirmed that dual residual building blocks and ProtBERT-BFD features were critical to the model's predictive strength. ERCNN-EGFR offers a scalable, cost-effective, and accurate computational approach for EGFR identification, with potential applications in breast cancer diagnostics, therapeutic target discovery, and personalized treatment strategies.

Keywords Epidermal growth factor receptor, Deep learning, Machine learning

Breast cancer remains the most common malignancy among women worldwide and the second leading cause of cancer-related mortality[1]. The epidermal growth factor receptor (EGFR), a transmembrane glycoprotein involved in cell proliferation, survival, and differentiation, plays a pivotal role in breast cancer progression[2]. Figure 1 shows a structural depiction of EGFR.

Under physiological conditions, EGFR signaling is tightly regulated; however, in many breast cancer subtypes, EGFR is overexpressed or aberrantly activated, leading to uncontrolled cell growth, increased metastatic potential, and poor prognosis[3].

EGFR has emerged as an important therapeutic target, with treatments including small-molecule tyrosine kinase inhibitors (e.g., gefitinib, erlotinib) and monoclonal antibodies (e.g., cetuximab, panitumumab)[4]. While these therapies have demonstrated clinical benefit, challenges such as drug resistance, tumor heterogeneity, and lack of robust predictive biomarkers limit their long-term efficacy[5]. Accurate and early identification of EGFR

[1]Faculty of Computing and Information Technology, Department of Information Technology, King Abdulaziz University, Jeddah 21589, Saudi Arabia. [2]Department of Computer Science, Bahria University, Islamabad, Pakistan. [3]Department of Information Systems, Faculty of Computing and Information Technology, King Abdul Aziz University, Jeddah 21589, Saudi Arabia. [4]Faculty of Computing and Information Technology, Department of Information Systems, King Abdulaziz University, Jeddah 21589, Saudi Arabia. [5]Faculty of Computing and Information Technology, Department of Computer Science, King Abdul Aziz University, Jeddah 21589, Saudi Arabia. [6]Department of Computer Engineering, College of Computer, Qassim University, Buraydah, Saudi Arabia. ✉email: farman.buic@bahria.edu.pk; tkhliefh@qu.edu.sa

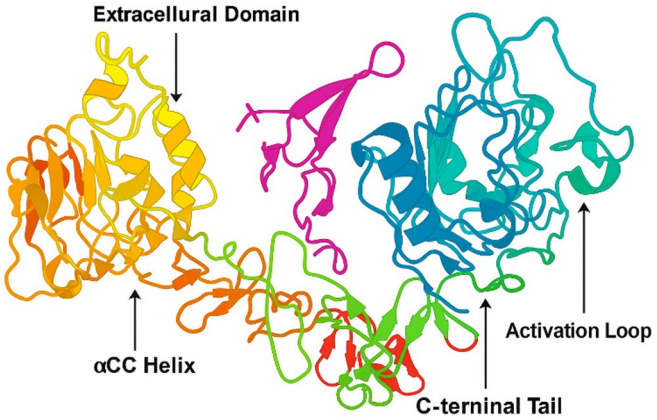

Fig. 1. Ribbon diagram of the EGFR structure, with distinct functional regions highlighted in different colors. The ligand-binding domain (yellow–orange) facilitates interaction with epidermal growth factor, while the kinase domain (blue–cyan) contains the catalytic loop responsible for phosphorylation activity. The juxtamembrane segment (red–green) plays a regulatory role in signal transduction. The magenta-colored loop represents a key motif involved in conformational changes upon activation. This color-coded representation enables clear visualization of EGFR's structural organization and functional domains. Structural model of the Epidermal Growth Factor Receptor (EGFR). Image adapted from Wikipedia (https://en.wikipedia.org/wiki/Epidermal_growth_factor_receptor), licensed under CC BY-SA 4.0

status in breast cancer patients is therefore essential to optimize treatment strategies. Traditional diagnostic techniques such as immunohistochemistry (IHC)[6] and fluorescence in situ hybridization (FISH)[7], though widely used, are labor-intensive, subjective, and dependent on tissue quality.

In recent years, several computational and machine learning approaches have been applied for EGFR-related research in cancer. For example, Studies have employed supervised screening approaches combined with structural biology techniques to identify potent EGFR inhibitors[8]. In contrast, others have integrated machine learning with pharmacogenomics to rank potential breast cancer drugs and associated biomarkers[9].

Comparative genomic and expression analyses of other prognostic markers, such as ANLN and KDR, have also provided insights that can complement EGFR-based investigations[10]. Moreover, stacking ensemble frameworks for predicting drug–drug synergy have demonstrated potential in designing combination therapies for EGFR-positive cancers[11].

The application of deep learning in breast cancer diagnostics has also seen rapid growth. Several studies have focused on using multi-modal imaging to improve diagnostic accuracy. Recent work, such as a brief survey on deep learning schemes for multi-image modalities, highlights this trend[12]. Other research has explored how combining radiomics features from mammography with deep learning models and using networks that fuse features from different modalities can enhance diagnostic precision[13,14]. These studies collectively highlight the potential of deep learning to analyze complex, multi-source data for improved clinical outcomes.

Sequence-based predictors eliminate the need for costly wet-lab procedures and allow rapid, large-scale screening. Sequence-based predictors using ML and DL have been developed for identification of many biological problems, such as druggable protein[15], angiogenic protein[16], and globular protein[17]. However, according to our best knowledge, no sequence-based computational model has been constructed for EGFR identification.

Considering these limitations and gaps, we propose an advanced DL model, namely Ensemble Residual Convolutional Neural Network (ERCNN), that integrates ensemble learning with the principles of residual and convolutional neural networks. The major contribution points are listed below.

- Developed a novel computational predictor for the fast and accurate identification of EGFR.
- Constructed a unique primary sequence-based dataset for training and testing, which serves as a valuable resource for advancing cancer research and drug discovery.
- Trained the model using ERCNN, a new architecture that integrates ensemble learning with residual and convolutional neural networks.
- The model was trained using a multi-perspective feature set: KSCTD, AmpPseAAC, CDT, and ProtBERT-BFD.

The graphical view of the applied methods is shown in Fig. 2.

Materials and methods

Datasets

To construct a robust dataset for our innovative predictor, we first sourced sequences of both EGFR and non-EGFR from the UniProt database[18]. After collection, we applied several filtering criteria to ensure high quality. We used the CD-hit tool[19] to remove sequences with more than 25% similarity, preventing redundancy.

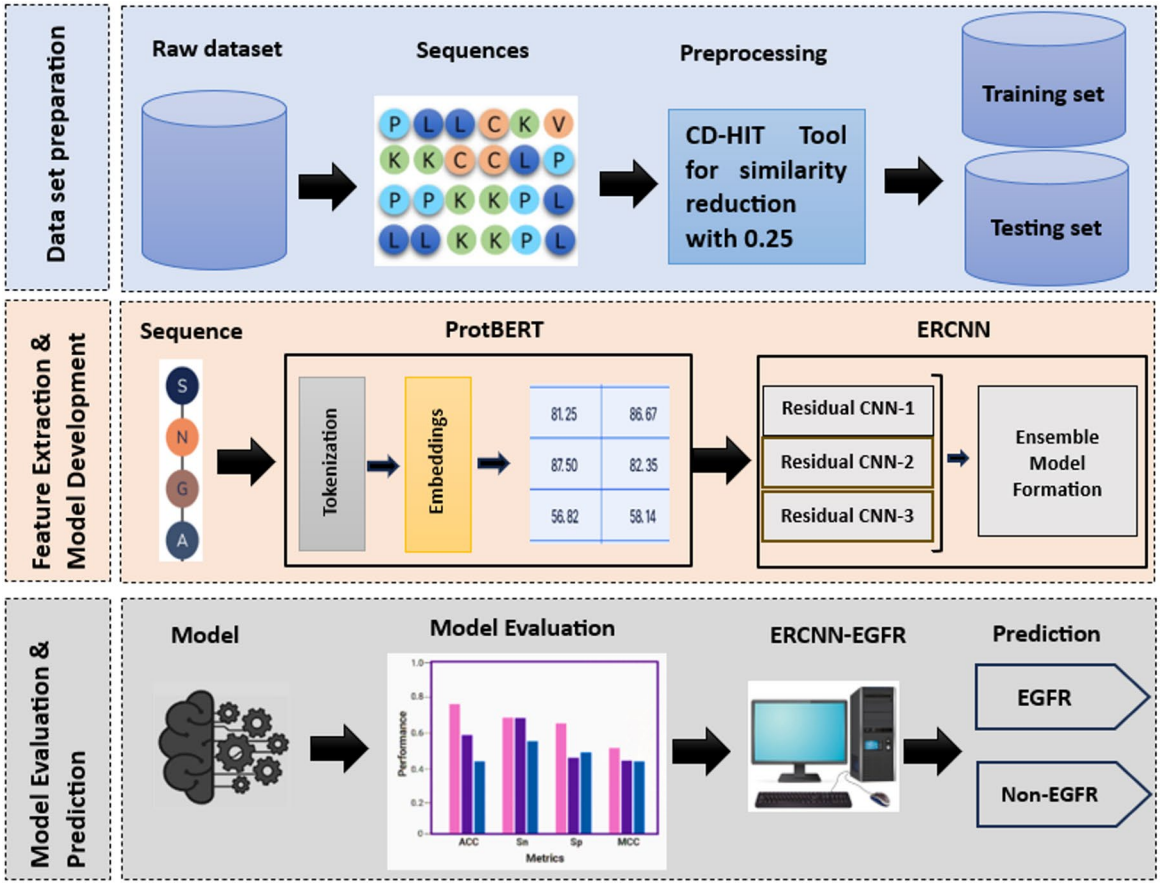

Fig. 2. Stepwise workflow of the ERCNN-EGFR. The process involves: (1) Data set preparation: raw protein sequences are collected, processed using the CD-HIT tool for similarity reduction, and split into training and testing sets; (2) Feature extraction and model development: sequences are tokenized and embedded using ProtBERT, then passed through multiple Residual CNN models to form an ensemble; (3) Model evaluation and prediction: ERCNN model is evaluated using performance metrics and final predictions classify sequences as EGFR or non-EGF.

Additionally, we excluded sequences shorter than 50 residues or those containing unknown characters. This process resulted in a final dataset of 1092 EGFR and 1143 non-EGFR sequences.

The finalized dataset was then randomly partitioned into training and testing subsets to evaluate the model's performance. The training set, used to train our model, consists of 845 EGFR and 858 non-EGFR sequences. The testing set, which contains 247 EGFR and 285 non-EGFR sequences, was reserved exclusively to assess the model's ability to generalize to new, unseen data.

Feature representative approaches

Feature extraction is the process of converting the primary sequences of proteins into numerical form that can be processed by a machine/deep learning model[20]. We use three feature extraction methods in our study: KSCTD, AmpPseAAC, CDT, and ProtBERT-BFD.

Amphiphilic Pseudo amino acid composition

AmpPseAAC was developed by Kuo-Chen Chou in 2005[21]. The updated pseudo amino acid composition method (AmpPseAAC) represents protein sequences using a set of numerical descriptors that capture the physicochemical characteristics and amino acid composition of the protein. By adding details about the amino acids' amphiphilicity, AmpPseAAC expands on PseAAC[22]. Amphiphilicity is a metric used to quantify how hydrophobic an amino acid is. Water tends to repel hydrophobic amino acids while drawing hydrophilic amino acids towards it. AmpPseAAC has been shown to be effective in several bioinformatics applications, including protein classification, such as identification of enzyme subfamily classes[23], prediction of apoptosis protein[24], and classification of human protein subcellular locations[25].

K-Spaced conjoint triad

KSCTD is a powerful tool for representing and analyzing biological sequences[26]. As a member of the sequence-based feature extraction family, it plays a crucial role in describing the functional and structural properties of biomolecules[27].

The KSCTD algorithm is rooted in the concept of triads, which are sequences of three consecutive amino acids. These triads represent local structural motifs that are fundamental to protein folding and function[28]. By considering the spatial arrangement of these triads through the concept of "k-spacing," KSCTD provides a more detailed and informative representation of sequences compared to traditional methods[29]. This approach allows for the capture of long-range interactions between amino acids, which are often crucial for protein function[30].

Composition distribution transition

The notion of Composition Distribution Transition (CDT) is utilized in the fields of computational biology and bioinformatics[31–33]. It serves as an effective tool for analyzing protein sequences. By analyzing the distribution of compositional changes within a sequence, the transition approach is a statistical technique that helps researchers understand sequence patterns, structural traits, and functional implications.

CDT examines how certain elements, like the amino acids in proteins or the nucleotides in DNA, vary in relative frequency throughout a sequence. This can provide important details on structural motifs, functional domains, or sequence evolution[34]. This technique is particularly helpful for pinpointing sequence segments of interest where compositional fluctuations can point to structural or functional alterations. A sequence is usually segmented into smaller areas for CDT analysis, and the compositional variations between these regions are then quantified. This may entail evaluating differences in amino acids. Such biological properties as conserved DNA sections, binding sites, or protein domains may be indicated by these compositional variations[35].

ProtBERT-BFD

ProtBERT-BFD is a large-scale protein language model built upon the Bidirectional Encoder Representations from Transformers (BERT) architecture, tailored specifically for protein sequence modeling. It is pretrained on the Big Fantastic Database (BFD), which contains over 2.1 billion protein sequences, making it one of the most extensive protein corpora ever used for model development. ProtBERT-BFD aims to produce rich, contextual embeddings that capture the evolutionary, structural, and functional characteristics of proteins directly from their amino acid sequences.

The model is trained using self-supervised learning with the masked language modeling strategy. In this approach, a fixed percentage (typically 15%) of amino acids in each sequence is randomly masked, and the model is tasked with predicting these masked residues from the surrounding context. By using a bidirectional attention mechanism, ProtBERT-BFD simultaneously incorporates information from both upstream and downstream positions, enabling it to learn long-range dependencies that are critical for understanding protein folding, structural domains, and functional sites. Unlike traditional profile-based approaches such as PSSMs, this model learns directly from raw sequences, avoiding the need for computationally intensive multiple sequence alignments.

ProtBERT-BFD offers several advantages over other protein language models, including UniRep and ProtTrans-T5. One major strength Lies in its extensive training dataset, the model is pretrained on 2.1 billion protein sequences, providing broad coverage of the protein sequence space and enabling better generalization to rare protein families and uncharacterized domains.

Another advantage is its bidirectional context capture, which allows the model to consider dependencies in both sequence directions simultaneously. This is in contrast to unidirectional approaches, which only model context in one direction, potentially missing important long-range interactions. Additionally, ProtBERT-BFD provides an alignment-free representation, eliminating the need for multiple sequence alignments. This significantly reduces preprocessing time and makes the model more scalable for large-scale bioinformatics studies.

Model training and prediction

During this stage, the tasks of model training will be carried out. In this connection, various deep learning frameworks will be employed, including BiLSTM, ERCNN, GAN, and GRU. The best performance is secured by ERCNN-based model, which is elaborated in the following subsection.

Ensemble residual convolutional neural network

ERCNN integrates ensemble learning with the principles of residual and convolutional neural networks[36,37]. This framework is designed to improve the performance of the model. Residual Neural Network (ResNet) is a pivotal component of the ensemble residual CNN. ResNet is a crucial part of the ensemble residual CNN. ResNet introduces shortcut connections to address the problem of vanishing gradients in CNN[38]. By enabling information to move through the network more directly, these shortcut links lower the possibility that performance may deteriorate as the network gets deeper. A machine learning method called ensemble learning makes use of several models to increase prediction accuracy[39]. ERCNN enhances model performance by combining multiple ResNet-CNN models into an ensemble framework. The concatenation of sub-models facilitates mitigating overfitting and capturing a broader range of features. This improved generalization is particularly advantageous when dealing with small datasets. Combining predictions from multiple sub-models leads to a more robust final decision, leveraging the consensus across several networks for increased reliability[40].

Keeping in view the advantages above, we implemented ensemble residual CNN. This work constructs the final model by two types of residual building block (RBB), i.e., RBB-1 and RBB-2. Both RBBs have three convolutional and batch normalization layers. Before convolutional layers, a padding layer is added to preserve the spatial dimensions of the input feature while performing the convolution operation. The process involves padding the input data with a layer of zeros around its edges. Each convolutional layer is tested with 32, 64, 128, and 256. The model achieved strong performance with 64 filters. Similarly, strider size of 3 is applied for

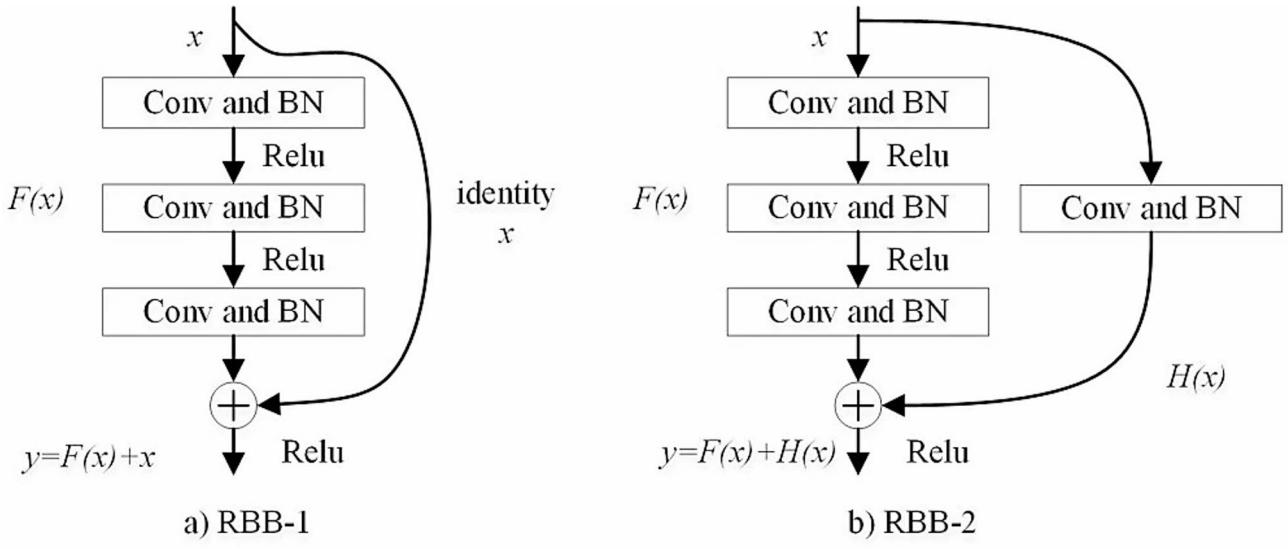

Fig. 3. Structural representation of RBB-1 and RBB-2.

Hyperparameter	Value
Number of epochs	80
Batch size	100
Learning rate	0.001
Convolution layers	3
Dropout	0.4
Activation function	Sigmoid
Optimizer	Adam

Table 1. ERCNN model hyperparameters.

all convolutional layers. The shortcut in RBB-1 is identified by x, while the shortcut in RBB-2 contains one convolutional and batch normalization layer. Both RBBs are represented in Fig. 3. RBB-1 is computed as:

$$y = F(x) + x \tag{1}$$

where, F is the nonlinear function for the convolutional path in RBB-1.

RBB-2 is expressed as:

$$y = F(x) + H(x) \tag{2}$$

where, H is the shortcut path.

Dropout layers are incorporated after the fully connected layers to enhance classification performance. The output layer contains a flatten layer and a sigmoid layer. The flatten layer transforms features into a vector shape. The sigmoid layer outputs the probability of each possible outcome. Other hyperparameters that we implemented in our model are listed in Table 1.

Feature selection approach

Feature selection is a crucial process in machine learning that involves choosing a subset of relevant features to use in model construction. The primary goal is to simplify the model, improve its performance, reduce training time, and make it more interpretable[41]. This is especially important in high-dimensional datasets where many features may be redundant or irrelevant. By removing noise and focusing on the most informative features, feature selection helps prevent overfitting and enhances the model's generalization ability. In this work, we used XGBoost-Feature Forward Selection (XGBoost-FFS) approach.

XGBoost-FFS is an advanced wrapper-based feature selection method that leverages the power of the XGBoost (Extreme Gradient Boosting) algorithm. Unlike filter methods that rank features independently of the model, wrapper methods evaluate subsets of features by training and testing a model on them. XGBoost-FFS is a greedy and iterative process that progressively builds an optimal feature set. It starts with an empty set and, in each step, adds the single feature that shows improvement to an XGBoost model's performance. This process continues until a predetermined number of features is selected.

The process begins with Initialization, where an empty set is created to hold the selected features. Next, during the Iteration phase, the algorithm loops through all features that are not yet in the set and temporarily adds each one. For each of these temporary sets, the Evaluation step occurs: an XGBoost model is trained and its performance is measured using a metric Like accuracy. After all options are evaluated, the Selection step identifies and permanently adds the single feature that resulted in the best performance improvement. This entire process is then Repeated until the desired number of features has been chosen. Although this method is computationally demanding because it trains a model in every loop, it is highly effective at identifying the most predictive features specifically for an XGBoost model. In this work, XGBoost-FFS selected 116, 18, 21, and 234 the best feature sets from KSCTD, AmpPseAAC, CDT, and ProtBERT-BFD.

Model assessment

To ensure the reliability of our novel approach, it is examined using assessment methods. The most extensively used method in bioinformatics is 10-fold cross-validation[42–45]. 10-fold is the division of our data into ten sets. The model is tested on the tenth set after being trained on nine folds of the sets. We repeat this ten times, testing with a new set each time. The final forecast is regarded as the mean of the 10-fold outcomes[46]. Additionally, we assess the model's performance using assessment metrics, such as MCC, sensitivity (Sn), specificity (Sp), and accuracy (Acc)[47,48]. The confusion matrix is used to formulate these parameters. These parameters are computed using the following equations.

$$Acc = 1 - \frac{EG_{-}^{+} + EG_{+}^{-}}{EG^{+} + EG^{-}} \tag{3}$$

$$Sn = 1 - \frac{EG_{-}^{+}}{EG^{+}} \tag{4}$$

$$Sp = 1 - \frac{EG_{+}^{-}}{EG^{-}} \tag{5}$$

$$MCC = \frac{1 - \left(\frac{EG_{-}^{+} + EG_{+}^{-}}{EG^{+} + EG^{-}}\right)}{\sqrt{\left(1 + \frac{EG_{-}^{+} + EG_{+}^{-}}{EG^{+}}\right)\left(1 + \frac{AG_{-}^{+} + EG_{+}^{-}}{EG^{-}}\right)}} \tag{6}$$

EG^{+} indicates accurately identified EGFR, while EG^{-} denotes non-EGFR instances that are accurately identified. Similarly, EG_{-}^{+} and EG_{+}^{-} refer to samples that are mistakenly identified.

Results and discussion

Results analysis of the DL algorithms using training set

The performance obtained by various DL models prior to feature selection is presented in Table 2. Using the CDT, BiLSTM model achieves 57.40% Acc, 65.48% Sn, and 49.47% Sp, with MCC of 0.276, indicating moderate predictive ability. GAN model with CDT performs slightly worse overall, with 53.13% Acc, 67.57% Sn, and 38.92% Sp, reflected by its MCC of 0.389. GRU model demonstrates stronger performance on CDT,

Algorithm	Method	Acc (%)	Sn (%)	Sp (%)	PR-AUC (%)	MCC
BiLSTM	CDT	57.40	65.48	49.47	64.88	0.276
GAN		53.13	67.57	38.92	60.36	0.389
GRU		69.40	77.48	61.47	75.98	0.396
ERCNN		72.69	88.62	57.03	78.78	0.485
BiLSTM	AmpPseAAC	52.38	77.86	27.42	55.25	0.103
GAN		54.61	68.76	40.85	57.11	0.465
GRU		72.46	83.40	61.70	75.66	0.467
ERCNN		78.32	74.31	82.26	83.47	0.596
BiLSTM	KSCTD	59.24	66.39	52.15	65.22	0.228
GAN		62.91	87.86	37.45	68.47	0.201
GRU		85.54	88.88	82.25	91.64	0.732
ERCNN		87.84	91.59	84.13	93.46	0.761
BiLSTM	ProtBERT-BFD	75.46	76.89	74.65	81.56	0.641
GAN		77.24	77.64	77.27	83.28	0.663
GRU		89.28	89.31	89.15	94.25	0.774
ERCNN		91.15	90.27	88.95	96.13	0.795

Table 2. Performance of DL frameworks before feature selection.

reaching 69.40% Acc, 77.48% Sn, and 61.47% Sp, with MCC of 0.396. ERCNN model secures the best CDT-based performance, with 72.69% Acc, 88.62% Sn, 57.03% Sp, and MCC of 0.485, showcasing its effectiveness in learning robust sequence features.

BiLSTM on AmpPseAAC attains only 52.38% Acc, with high sensitivity (77.86%) but very low specificity (27.42%), resulting in MCC of 0.103. GAN slightly improves on this with 54.61% Acc, 68.76% Sn, 40.85% Sp, and MCC of 0.465. GRU generated stronger outcomes, with 72.46% Acc, 83.40% Sn, 61.70% Sp, and MCC of 0.467. ERCNN achieves the highest results for AmpPseAAC, with 78.32% Acc, 74.31% Sn, 82.26% Sp, and MCC of 0.596, confirming its consistent superiority.

With KSCTD features, BiLSTM attains 59.24% Acc, 66.39% Sn, and 52.15% Sp, with MCC of 0.228. GAN performs slightly better, reaching 62.91% accuracy, but suffers from poor specificity (37.45%) despite high sensitivity (87.86%), resulting in a low MCC of 0.201. GRU with KSCTD attains 85.54% Acc, 88.88% Sn, 82.25% Sp, and MCC of 0.732. ERCNN further enhances performance, achieves 87.84% Acc, 91.59% Sn, 84.13% Sp, and MCC of 0.761, underscoring the power of ensemble and residual learning in extracting meaningful sequence-level features.

ProtBERT-BFD embeddings produce promising results overall. BiLSTM secures 75.46% Acc, 76.89% Sn, 74.65% Sp, and MCC of 0.641. GAN improves slightly to 77.24% accuracy and MCC of 0.663. GRU model further boosts performance with 89.28% Acc, 89.31% Sn, 89.15% Sp, and MCC of 0.774. ERCNN obtains the best ProtBERT-BFD results, with 91.15% Acc, 90.27% Sn, 88.95% Sp, and MCC of 0.795, making it the best performer across all descriptors.

Results analysis of the DL algorithms after feature selection approach

The performance of deep learning frameworks after feature selection is summarized in Table 3. Using the CDT, BiLSTM model achieves an Acc of 59.56%, Sn of 60.76%, and Sp of 57.64%, with MCC of 0.402, indicating a modest classification ability. GAN performs slightly worse with 56.35% Acc and MCC of 0.367. However, GRU shows a progressive improvement, achieving 71.57% Acc, 71.27% Sn, and 71.90% Sp, with MCC of 0.543. ERCNN secures the best CDT-based results, with 74.68% Acc, 75.76% Sn, 73.37% Sp, and MCC of 0.575, highlighting its superior learning of discriminative sequence patterns.

BiLSTM using AmpPseAAC shows lower performance with 54.57% accuracy and MCC of 0.346. GAN achieves slightly higher values (55.57% accuracy, 55.36% sensitivity, 55.85% specificity, 0.357 MCC). GRU model significantly outperforms both BiLSTM and GAN, secures 74.68% Acc, 73.35% Sn, 75.33% Sp, and MCC of 0.576. ERCNN consistently generated the highest results on AmpPseAAC, producing 79.89% Acc, 80.99% Sn, 78.28% Sp, and MCC of 0.622, confirming its robustness in feature learning.

When considering KSCTD, BiLSTM achieves 60.36% Acc with balanced Sn (60.42%) and Sp (60.24%), leading to MCC of 0.415. GAN model shows an improvement to 64.76% Acc and MCC of 0.451. GRU demonstrates remarkable performance with 87.33% Acc, 88.95% Sn, 87.02% Sp, and MCC of 0.764. ERCNN further enhances the results, obtaining 88.25% accuracy, 90.13% sensitivity, 87.93% specificity, and the highest MCC for KSCTD (0.772).

ProtBERT-BFD embeddings yield the best performance across all descriptors. BiLSTM achieves 77.12% Acc, 77.71% Sn, 77.34% Sp, and MCC of 0.664. GAN improves further with 79.44% accuracy and 0.688 MCC, while GRU generated 91.67% Acc, 90.87% Sn, 90.89% Sp, and MCC of 0.793. ERCNN obtains the best performance, with 93.48% accuracy, 94.53% sensitivity, 92.58% specificity, and MCC of 0.816. These findings confirm that ProtBERT-BFD embeddings, when combined with ERCNN, show the best generalizable representation for classification.

Algorithm	Method	Acc (%)	Sn (%)	Sp (%)	PR-AUC (%)	MCC
BiLSTM	CDT	59.56	60.76	57.64	66.27	0.402
GAN		56.35	57.89	54.36	63.85	0.367
GRU		71.57	71.27	71.90	80.34	0.543
ERCNN		74.68	75.76	73.37	83.63	0.575
BiLSTM	AmpPseAAC	54.57	57.47	55.78	62.16	0.346
GAN		55.57	55.36	55.85	63.86	0.357
GRU		74.68	73.35	75.33	83.97	0.576
ERCNN		79.89	80.99	78.28	86.09	0.622
BiLSTM	KSCTD	60.36	60.42	60.24	67.75	0.415
GAN		64.76	65.24	63.76	71.52	0.451
GRU		87.33	88.95	87.02	93.63	0.764
ERCNN		88.25	90.13	87.93	94.78	0.772
BiLSTM	ProtBERT-BFD	77.12	77.71	77.34	84.57	0.664
GAN		79.44	80.43	78.51	86.53	0.688
GRU		91.67	90.87	90.89	95.37	0.793
ERCNN		93.48	94.53	92.58	97.38	0.816

Table 3. Performance of DL frameworks after feature selection.

Classifier	Acc (%)	Sn (%)	Sp (%)	PR-AUC (%)	MCC
RF	80.68	81.35	80.01	86.44	0.684
ERT	83.65	82.65	84.27	89.13	0.709
Adaboost	85.72	84.25	86.16	91.36	0.735
ERCNN-EGFR	93.48	94.53	92.58	97.38	0.816

Table 4. Performance of ML models on the best feature selection approach.

Classifier	Acc (%)	Sn (%)	Sp (%)	PR-AUC (%)	MCC
BiLSTM	72.24	72.35	72.53	80.26	0.442
GAN	66.15	67.26	65.26	74.67	0.407
GRU	69.17	69.07	69.97	77.28	0.424
ERCNN-EGFR	82.85	81.49	83.04	90.24	0.652

Table 5. Results on the independent test Set.

Results analysis of the ML frameworks on the best feature selection approach

We have extended our experimentation to include several ML classifiers, with the results presented in Table 4. These ML classifiers were trained via XGBoost-FFS's best feature set. RF model, trained on these selected features, demonstrated a solid baseline performance, achieving an accuracy of 80.68% and MCC of 0.684. ERT slightly outperformed RF, yielding higher accuracy (83.65%), specificity (84.27%), and MCC (0.709). Adaboost, on the other hand, showed even better performance, Surpassing both RF and ERT with accuracy of 85.72% and MCC of 0.735. This indicates its superior ability to handle the classification task.

Our proposed model, ERCNN-EGFR, consistently exhibited remarkable performance. With accuracy of 88.56%, it outperformed all other machine learning classifiers. Notably, the model achieved a high sensitivity of 89.78%, demonstrating its strong capability to correctly identify EGFR-positive instances. Specificity of 87.29% also confirms its robust performance in correctly classifying EGFR-negative instances. High MCC of 0.773 further validates the model's best performance, confirming its efficacy in accurately distinguishing between EGFR and non-EGFR.

Performance of the classifiers on the independent test set

We further examined the performance of baseline classifiers on the Independent Test Set to assess their generalization ability. BiLSTM model demonstrated reasonable performance, achieving Acc of 72.24%, Sn of 72.35%, Sp of 72.53%, and MCC of 0.442. GAN showed slightly lower performance, with Acc of 66.15%, Sn of 67.26%, Sp of 65.26%, and MCC of 0.407. GRU performed better than GAN, and secured 69.17% Acc, 69.07% Sn, 69.97% Sp, and MCC of 0.424, however, it was lower than BiLSTM.

On the Independent Test Set, the proposed model demonstrated strong generalization, obtaining an accuracy of 82.85%, sensitivity of 81.49%, specificity of 83.04%, and MCC of 0.652. This performance highlights the great precision of EGFR prediction provided by ERCNN-EGFR. The findings validate our predictor's exceptional performance and promising generalization efficacy by showing it to outperform BiLSTM, GAN, and GRU models on the testing dataset across all assessment parameters. A thorough comparison of these classifiers is given by the results in Table 5, which also demonstrate how well ERCNN-EGFR performs in predicting EGFR accurately.

The proposed ERCNN-EGFR significantly outperformed all other classifiers, obtaining 82.85% Acc, 81.49% Sn, 83.04% Sp, and MCC of 0.654. These results highlight the superior predictive power of ERCNN-EGFR, demonstrating its ability to effectively capture discriminative features and achieve higher generalization capability than BiLSTM, GAN, and GRU models on the testing dataset.

As part of our model evaluation, we present the ROC curves for the ERCNN model on both the training and independent test datasets, as shown in Figs. 4 and 5, respectively. The ROC curve provides a comprehensive assessment of the model's discriminative ability by illustrating the trade-off between true positive rate and false positive rate across different classification thresholds.

In our results, the ERCNN achieves an AUC of 0.973 on the training set and 0.904 on the independent test set. These strong AUC values indicate that the probability scores generated by the ERCNN model reliably distinguish between positive and negative classes in both training and unseen data. The smooth and consistently elevated curves, well above the diagonal random classifier line, further confirm robust separation and good classification capacity.

In addition to the ROC analysis, we have computed Precision-Recall AUC values for the ERCNN model as presented in Tables 2 through 5 of the manuscript. Precision-Recall AUC is particularly meaningful for imbalanced datasets, as it focuses on the trade-off between precision and recall. The high Precision-Recall AUC values further corroborate the ERCNN model's strong performance and reliable probability outputs.

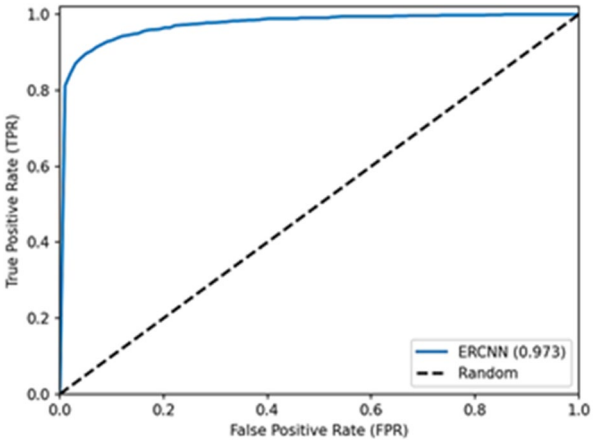

Fig. 4. ROC curve of the ERCNN on the training set.

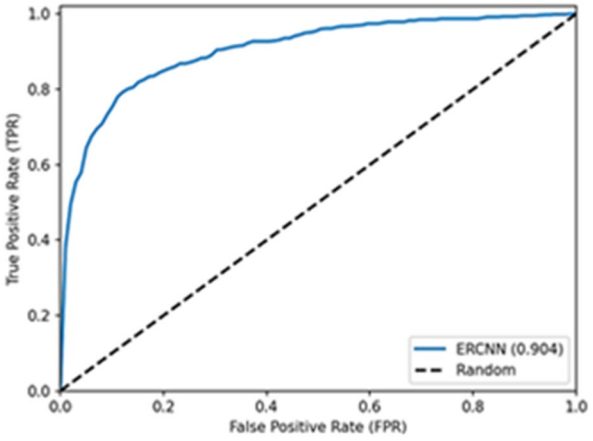

Fig. 5. ROC curve of the ERCNN on the Independent test set.

Variant	Acc (%)	Sn (%)	Sp (%)	PR-AUC (%)	MCC
Plain CNN	90.80	91.90	89.60	95.10	0.771
RCNN (Residual CNN without ensemble)	91.50	92.40	90.20	95.80	0.784
RCNN with RBB-1 only	92.20	93.50	90.80	96.40	0.801
RCNN with RBB-2 only	92.00	92.80	91.10	96.20	0.797
Full ERCNN (Proposed)	93.48	94.53	92.58	97.38	0.816

Table 6. Ablation study of the proposed model.

Ablation study analysis of the proposed model

To assess the contribution of different architectural components in ERCNN, we performed an ablation study using ProtBERT-BFD features (Table 6). The results demonstrate that each architectural enhancement meaningfully improves predictive performance.

The baseline Plain CNN achieved 90.80% Acc, 91.90% Sn, 89.60% Sp, a PR-AUC of 95.10%, and MCC of 0.771. While these results confirm the model's capacity to capture local sequence patterns, they also reveal its Limited generalization ability. Incorporating residual connections in the RCNN variant improved overall performance, achieving 91.50% Acc and MCC of 0.784. This improvement highlights the effectiveness of residual learning in stabilizing deeper networks and mitigating vanishing gradient issues.

Further analysis of individual residual building blocks showed that both RBB-1 and RBB-2 enhanced model capability beyond Plain CNN and RCNN. The RBB-1 variant attained 92.20% Acc, 93.50% Sn, and the highest MCC among the single-block models (0.801), demonstrating its stronger capacity to capture discriminative sequence features. Similarly, the RBB-2 variant achieved 92.00% Acc, 92.80% Sn, and MCC of 0.797, underscoring its complementary role in refining representation learning.

The Full ERCNN, which integrates ensemble learning with both RBB-1 and RBB-2, consistently outperformed all ablated variants. It achieved the best results with 93.48% Acc, 94.53% Sn, 92.58% Sp, PR-AUC of 97.38%, and MCC of 0.816. These findings confirm that the synergistic integration of multiple residual blocks within an ensemble framework allows ERCNN to capture both local and global sequence-level features more effectively, thereby ensuring robust and reliable EGFR prediction.

Computational cost and deployment challenges

ERCNN-EGFR was designed as a lightweight residual ensemble framework, making it computationally more efficient than deeper transformer-based models Such as ESM. The model was trained and tested on a workstation equipped with an NVIDIA RTX 3090 GPU (24 GB VRAM), 128 GB RAM, and an Intel Xeon Gold processor. Under this configuration, the average training time for ERCNN-EGFR was approximately 20–25 min, indicating feasibility for large-scale screening tasks.

In terms of expected hardware for deployment, the model can be efficiently run on a single modern GPU (e.g., NVIDIA A100/RTX series) or even optimized for CPU-based environments with slightly higher inference times (tens of milliseconds per sequence). This makes it scalable to high-throughput diagnostic pipelines.

However, we also acknowledge that in real-time diagnostic environments, challenges such as latency, throughput, and integration into clinical workflows must be carefully considered. High-throughput clinical pipelines often require efficient batch processing, seamless integration with hospital information systems, and potentially cloud-based deployment for distributed processing. To address this, we highlight in the Discussion that the model can be further optimized using model compression, pruning, or knowledge distillation to reduce memory footprint and further improve inference speed.

Conclusion and future directions

In this study, we developed ERCNN-EGFR, a novel deep learning–based predictor for the accurate identification of epidermal growth factor receptor (EGFR) proteins. By integrating ensemble learning with residual convolutional neural networks and leveraging multi-perspective feature representations, including CDT, AmpPseAAC, KSCTD, and ProtBERT-BFD, the proposed model demonstrated Superior performance compared to conventional deep learning architectures Such as BiLSTM, GRU, and GAN. Notably, ERCNN-EGFR achieved an accuracy of 87.84% on the training dataset and 82.85% on the independent testing dataset, underscoring its strong generalization ability. The ablation analysis further confirmed that the combination of ProtBERT-BFD embeddings and the dual residual building block design (RBB-1 and RBB-2) was central to the model's predictive strength. These architectural innovations enabled ERCNN-EGFR to capture both local and global sequence patterns more effectively than standard CNN ensembles or ResNet-based frameworks.

From a practical perspective, ERCNN-EGFR offers a scalable, cost-effective, and reliable computational tool that can complement existing laboratory-based EGFR detection methods such as IHC and FISH. Unlike conventional approaches or static variant databases, the model has the potential to identify previously uncharacterized EGFR variants, thereby accelerating the discovery of therapeutic targets and supporting precision oncology.

Our model was trained on a curated EGFR dataset, which may carry potential bias related to dataset composition and size. Furthermore, while ERCNN-EGFR achieved strong performance for EGFR identification in breast cancer, its generalization to other cancers or receptor families has not yet been validated. In future work, we plan to extend ERCNN-EGFR by integrating additional sources of biological information, such as structural features, omics-based datasets, and pathway-level annotations, to further enhance predictive performance. Moreover, we aim to evaluate the model on diverse cancer types and receptor families beyond EGFR to assess its generalizability across broader oncogenic contexts. Incorporating explainable AI techniques will also be a priority, enabling the identification of key sequence determinants that drive EGFR prediction and improving biological interpretability. Finally, we envision deploying ERCNN-EGFR as a user-friendly web server or software package to facilitate its accessibility for the wider research community in cancer biology and drug discovery.

Data availability

The datasets and code are available online freely at the public link: https://github.com/Farman335/ERCNN-EGFR.

Received: 12 June 2025; Accepted: 2 September 2025
Published online: 29 September 2025

References

1. DeSantis, C. E. et al. Breast cancer statistics, 2019, vol. 69, no. 6, pp. 438–451, (2019).
2. Levantini, E., Maroni, G., Del Re, M. & Tenen, D. G. EGFR signaling pathway as therapeutic target in human cancers, in *Seminars in Cancer Biology*, vol. 85, pp. 253–275: Elsevier. (2022).
3. Kallel, I. et al. EGFR overexpression relates to triple negative profile and poor prognosis in breast cancer patients in tunisia. *J Recept Signal Transduct Res.* **32**(3), 142–149 (2012).
4. Dassonville, O., Bozec, A., Fischel, J. L. & Milano, G. EGFR targeting therapies: monoclonal antibodies versus tyrosine kinase inhibitors: similarities and differences, vol. 62, no. 1, pp. 53–61, (2007).
5. Roskoski, R. Jr Small molecule inhibitors targeting the egfr/erbb family of protein-tyrosine kinases in human cancers. *Pharmacol Res.***139**, pp. 395–411, (2019).
6. Zaha, D. Significance of immunohistochemistry in breast cancer, vol. 5, no. 3, p. 382, (2014).
7. Persons, D. L. et al. Fluorescence in situ hybridization (FISH) for detection of HER-2/neu amplification in breast cancer: a multicenter portability study. *Ann Clin Lab Sci.* **30**(1), 41–48 (2000).

8. Mehmood, A., Li, D., Li, J., Kaushik, A. C. & Wei, D. Q. Supervised screening of EGFR inhibitors validated through computational structural biology approaches. *ACS Med. Chem. Lett.* **15** (12), 2190–2200 (2024).
9. Mehmood, A. et al. Ranking breast cancer drugs and biomarkers identification using machine learning and pharmacogenomics. *ACS Pharmacol. Translational Sci.* **6** (3), 399–409 (2023).
10. Mehmood, A., Li, R., Kaushik, A. C. & Wei, D. Q. Comparative analysis of the genomic and expression profiles of ANLN and KDR as prognostic markers in breast cancer. *Silico Pharmacol.* **13** (1), 15 (2025).
11. Mehmood, A., Kaushik, A. C. & Wei, D. Q. DDSBC: a stacking ensemble classifier-based approach for breast cancer drug-pair cell synergy prediction. *J. Chem. Inf. Model.* **64** (16), 6421–6431 (2024).
12. Mahmood, T. et al. A brief survey on breast cancer diagnostic with deep learning schemes using multi-image modalities. *IEEe Access.* **8**, 165779–165809 (2020).
13. Mahmood, T., Saba, T., Rehman, A. & Alamri, F. S. Harnessing the power of radiomics and deep learning for improved breast cancer diagnosis with multiparametric breast mammography. *Expert Syst. Appl.* **249**, 123747 (2024).
14. Mahmood, T., Saba, T. & Rehman, A. Breast cancer diagnosis with MFF-HistoNet: a multi-modal feature fusion network integrating CNNs and quantum tensor networks. *J. Big Data*. **12** (1), 60 (2025).
15. Alghushairy, O. et al. Machine learning-based model for accurate identification of druggable proteins using light extreme gradient boosting. *J. Biomol. Struct. Dynamics*. **42** (22), 12330–12341 (2024).
16. Almusallam, N. et al. An omics-driven computational model for angiogenic protein prediction: advancing therapeutic strategies with Ens-deep-AGP. *Int. J. Biol. Macromol.* **282**, 136475 (2024).
17. Zouari, S. et al. Deep-GB: A novel deep learning model for globular protein prediction using CNN-BiLSTM architecture and enhanced PSSM with trisection strategy. *IET Syst. Biol.* **18** (6), 208–217 (2024).
18. Ali, F. et al. IR-MBiTCN: computational prediction of insulin receptor using deep learning: A multi-information fusion approach with multiscale bidirectional Temporal convolutional network. *International J. Biol. Macromolecules*, p. 143844, (2025).
19. Barukab, O., Ali, F., Alghamdi, W., Bassam, Y. & Khan, S. A. DBP-CNN: Deep Learning-based prediction of DNA-binding proteins by coupling discrete cosine transform with Two-dimensional convolutional neural network. *Expert Syst. Applications.* **65**, 116729 (2022).
20. Zulfiqar, H. et al. Deep-STP: a deep learning-based approach to predict snake toxin proteins by using word embeddings. *Front. Med.* **10**, 1291352 (2024).
21. Ali, F. & Hayat, M. Classification of membrane protein types using voting feature interval in combination with chou' s Pseudo amino acid composition. *J. Theor. Biol.* **384**, 78–83 (2015).
22. Zhu, W. et al. A first computational frame for recognizing heparin-binding protein. *Diagnostics* **13** (14), 2465 (2023).
23. Chou, K. C. Using amphiphilic pseudo amino acid composition to predict enzyme subfamily classes, vol. 21, no. 1, pp. 10–19, (2005).
24. Su, W. et al. Prediction of apoptosis protein subcellular location based on amphiphilic Pseudo amino acid composition. *Bioinformatics*. **14**, 1157021 (2023).
25. Majid, A., Choi, T. S., Processing, I. & Recognition, P. A New Ensemble Scheme for Predicting Human Proteins Subcellular Locations, vol. 3, no. 1, pp. 1–8, (2010).
26. Chen, Z. et al. iFeature: a python package and web server for features extraction and selection from protein and peptide sequences. *Anal. Biochem*. **34**(14), 2499–2502 (2018).
27. Gu, Z. F. et al. Prediction of blood–brain barrier penetrating peptides based on data augmentation with augur. *BMC Biol.* **22** (1), 86 (2024).
28. Wang, H. and X. J. B. b. Hu, Accurate prediction of nuclear receptors with conjoint triad feature, vol. 16, no. 1, pp. 1–13, (2015).
29. Ali, F. et al. VEGF-ERCNN: A deep learning-based model for prediction of vascular endothelial growth factor using ensemble residual CNN. *J. Comput. Sci.* **83**, 102448 (2024).
30. Li, F. et al. Computational analysis and prediction of PE_PGRS proteins using machine learning, vol. 20, pp. 662–674, (2022).
31. Liu, T. et al. ApoPred: identification of apolipoproteins and their subfamilies with multifarious features. *Front. Cell. Dev. Biology*. **8**, 621144 (2021).
32. Ali, F. et al. Deep-GHBP: improving prediction of growth Hormone-binding proteins using deep learning model. *Biomed. Signal Process. Control*. **78**, 103856 (2022).
33. Khan, A., Uddin, J., Ali, F., Banjar, A. & Daud, A. Comparative analysis of the existing methods for prediction of antifreeze proteins. *Chemometr. Intell. Lab. Syst.* **232**, 104729 (2023).
34. Arif, I., Agrahari, G., Gautam, A. K. & Chatterjee, A. Inferring layer-by-layer composition in Au-Ag nanoparticles using a combination of X-ray photoelectron spectroscopy and Monte Carlo simulations. *Sci. Rep.* **691**, 121503 (2020).
35. Zhou, Y. N., Li, J. J. & Luo, Z. Synthesis of gradient copolymers with simultaneously tailor-made chain composition distribution and glass transition temperature by semibatch ATRP: from modeling to application. *iScience*. **50**(15), 3052–3066 (2012).
36. Fan, X. et al. Deep learning for intelligent traffic sensing and prediction: recent advances and future challenges, vol. 2, pp. 240–260, (2020).
37. Liu, T. et al. Cm-siRPred: predicting chemically modified SiRNA efficiency based on multi-view learning strategy. *Int. J. Biol. Macromol.* **264**, 130638 (2024).
38. Zhang, H. Q. et al. PMPred-AE: a computational model for the detection and interpretation of pathological myopia based on artificial intelligence. *Front. Med.* **12**, 1529335 (2025).
39. Alsini, R. et al. Deep-VEGF: deep stacked ensemble model for prediction of vascular endothelial growth factor by concatenating gated recurrent unit with two-dimensional convolutional neural network. *Journal Biomol. Struct. Dynamics*, **21**, 1–11 (2024).
40. Ali, F. et al. DEEP-EP: Identification of epigenetic protein by ensemble residual convolutional neural network for drug discovery, *Methods*, vol. 226, pp. 49–53, (2024).
41. Ali, F. et al. DBPPred-PDSD: machine learning approach for prediction of DNA-binding proteins using discrete wavelet transform and optimized integrated features space. *Chemometr. Intell. Lab. Syst.* **182**, 21–30 (2018).
42. Khan, A. et al. Prediction of antifreeze proteins using machine learning. *Sci. Rep.* **12** (1), 1–10 (2022).
43. Ali, F. et al. Comprehensive analysis of computational models for prediction of anticancer peptides using machine learning and deep learning. *Archives Comput. Methods Engineering.* **11**, 1–21 (2025).
44. Almusallam, N. et al. Multi-headed ensemble residual CNN: a powerful tool for fibroblast growth factor prediction. *Results Eng.* **24**, 103348 (2024).
45. Khan, Z. U., Ali, F., Ahmad, I., Hayat, M. & Pi, D. iPredCNC: computational prediction model for Cancerlectins and non-cancerlectins using novel cascade features subset selection. *Chemometr. Intell. Lab. Syst.* **195**, 103876 (2019).
46. Khan, A., Uddin, J., Ali, F., Banjar, A. & Daud, A. Comparative analysis of the existing methods for prediction of antifreeze proteins. *Chemometrics Intell. Lab. Systems.* **89**, 104729 (2022).
47. Ghulam, A. et al. Deep learning-based model for improving prediction of anticancer peptides using two-dimensional convolutional neural network. *Chemometr. Intell. Lab. Syst.* **226**, 104589 (2022).
48. Khan, Z. U., Ali, F., Khan, I. A., Hussain, Y. & Pi, D. iRSpot-SPI: deep learning-based recombination spots prediction by incorporating secondary sequence information coupled with physio-chemical properties via chou's 5-step rule and Pseudo components. *Chemometr. Intell. Lab. Syst.* **189**, 169–180 (2019).

Author contributions

Wajdi Alghamdi: Performed the experiments, contributed to data acquisition.Farman Ali: Drafted and composed the manuscript, contributed to the interpretation of results.Raed Alsini: Performed the experiments, assisted in data acquisition.Amal Babour: Conducted validation analyses, contributed to data verification.Naif Wahed Rajkhan: Substantively revised the manuscript, contributed to critical editing.Tamim Alkhalifah: Reviewed the manuscript, contributed to critical feedback, and interpretation.All authors have agreed to be personally accountable for their own contributions, and to ensure that questions related to the accuracy or integrity of any part of the work are appropriately investigated, resolved, and documented in the literature.

Funding

The Researchers would like to thank the Deanship of Graduate Studies and Scientific Research at Qassim University for financial support (QU-APC-2025).

Declarations

Competing interests

The authors declare no competing interests.

Ethical approval

No human or animal subjects are involved in this work.

Consent to publish

All authors agree on the publication of the paper.

Additional information

Correspondence and requests for materials should be addressed to F.A. or T.A.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.