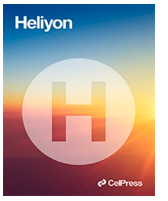
Heliyon 10 (2024) e30373

Contents lists available at ScienceDirect

Heliyon

journal homepage: www.cell.com/heliyon



Research article

Efficient screening of pharmacological broad-spectrum anti-cancer peptides utilizing advanced bidirectional Encoder representation from Transformers strategy

Yupeng Niu [a,c,1], Zhenghao Li [a,c,1], Ziao Chen [b,c,1], Wenyuan Huang [a,c], Jingxuan Tan [a,c], Fa Tian [a], Tao Yang [a,c], Yamin Fan [a,c], Jiangshu Wei [a], Jiong Mu [a,c,*]

[a] College of Information Engineering, Sichuan Agricultural University, Ya'an 625000, China
[b] College of Law, Sichuan Agricultural University, Ya'an 625000, China
[c] Artificial intelligence laboratory, Sichuan Agricultural University, Ya'an 625000, China

ARTICLE INFO

Keywords:
Anti-cancer peptides
Deep learning
Natural language processing
Transformers

ABSTRACT

In the vanguard of oncological advancement, this investigation delineates the integration of deep learning paradigms to refine the screening process for Anticancer Peptides (ACPs), epitomizing a new frontier in broad-spectrum oncolytic therapeutics renowned for their targeted antitumor efficacy and specificity. Conventional methodologies for ACP identification are marred by prohibitive time and financial exigencies, representing a formidable impediment to the evolution of precision oncology. In response, our research heralds the development of a groundbreaking screening apparatus that marries Natural Language Processing (NLP) with the Pseudo Amino Acid Composition (PseAAC) technique, thereby inaugurating a comprehensive ACP compendium for the extraction of quintessential primary and secondary structural attributes. This innovative methodological approach is augmented by an optimized BERT model, meticulously calibrated for ACP detection, which conspicuously surpasses existing BERT variants and traditional machine learning algorithms in both accuracy and selectivity. Subjected to rigorous validation via five-fold cross-validation and external assessment, our model exhibited exemplary performance, boasting an average Area Under the Curve (AUC) of 0.9726 and an F1 score of 0.9385, with external validation further affirming its prowess (AUC of 0.9848 and F1 of 0.9371). These findings vividly underscore the method's unparalleled efficacy and prospective utility in the precise identification and prognostication of ACPs, significantly ameliorating the financial and temporal burdens traditionally associated with ACP research and development. Ergo, this pioneering screening paradigm promises to catalyze the discovery and clinical application of ACPs, constituting a seminal stride towards the realization of more efficacious and economically viable precision oncology interventions.

1. Introduction

According to the World Health Organization (WHO), there are approximately 20 million incidents of neoplasm and 10 million

* Corresponding author.
E-mail address: jmu@sicau.edu.cn (J. Mu).
[1] These authors contributed equally to this work.

https://doi.org/10.1016/j.heliyon.2024.e30373
Received 12 December 2023; Received in revised form 24 April 2024; Accepted 24 April 2024
Available online 1 May 2024

tumor-induced fatalities every year. It is envisaged that the global pharmaceutical marketpertaining to oncological agents will amount to 436.02 billion US dollars by 2023. Amongst the diverse therapeutic strategies, precision oncology has demonstrated promising outcomes in recent clinical studies. However, the exorbitant cost and the substantial capital investment entailed for research and development persist as major impediments. The process of identifying suitable molecular targets, whether genomic, proteomic, or specific tissue microenvironments, is often arduous and time-consuming. In recent years, broad-spectrum oncolytic peptides have garnered significant attention due to their attributes of robust efficacy, high potency, and expeditious action against a wide repertoire of malignancies. Anti-cancer peptides (ACPs), with their distinctive mechanisms of pharmacological action, are increasingly being acknowledged as a promising modality in oncological therapeutics, paving the way for novel avenues in pharmacological research and drug discovery. They offer many advantages over conventional treatments, including lower molecular weights, simpler structures, higher tumor selectivity, reduced side effects, easy absorption, diverse routes of administration, and lower risk of inducing multidrug resistance [1]. Targeting specificity, ACP-cell interactions, peptide permeability, stability, and efficacy are all impacted by factors such as the chemical moiety of the ACP sequence [2]. Therapeutic peptides have been praised for their target specificity and low toxicity [3]. In addition, peptide-functionalized liposomes have been shown to provide increased cancer cell specificity, enhanced tumor penetration capacity, and significant tumor growth inhibition [4]. Anti-cancer peptides can be categorized into three main groups based on their mechanism of action: inhibitory peptides, necrosis-inducing peptides, and pro-apoptotic peptides [5]. They act through a variety of mechanisms, including induction of apoptosis, cell cycle arrest, cell membrane rupture, inhibition of intracellular signaling, topoisomerases and proteases, and antiangiogenic activity [6]. The objective is to harness advanced deep-learning techniques in conjunction with Pharmacological Broad-Spectrum Anti-Neoplastic Peptides, known for their pan-cancer cellular efficacy. Antineoplastic peptides emblemize a burgeoning domain within the oncology sphere, presenting the potential for innovative therapeutic strategies that could significantly bolster therapeutic efficacy whilst mitigating toxic side effects. The model streamlines drug discovery by efficiently identifying peptides with anticancer potential, cutting down on preliminary screening time and focusing on the most promising candidates. It reduces the number of required experiments and associated costs, minimizing the resources needed for biological activity verification. Furthermore, the model accelerates the transition from lab research to clinical trials, speeding up the development of new anticancer peptide drugs.

Currently, cellular research and clinical trials—which primarily rely on conventional biochemical and animal investigations—are the primary screening methods for anticancer peptides (ACPs). These procedures have certain efficiency limits, are expensive, and take a long time to complete [7]. To get over these restrictions, researchers are coming up with new tactics. For example, a virtual peptide library containing 677 peptides based on database and literature searches was generated by a cheminformatics approach. The candidate peptides were screened to five by screening for anticancer potential, non-toxicity, non-allergenicity, and non-hemolysis. According to molecular docking, PSYLNTPLL was the best potential peptide to stably bind to critical p47phox residues, whereas LYSPH was the most promising for targeting myeloperoxidase, xanthine oxidase, and Keap 1 [8]. More extensive study is being done for clinical uses. Researchers are deeply exploring the target selection of peptide-based vaccines, the design and screening of epitope peptides, clinical efficacy and adverse events, and the combination of peptide-based vaccines with other therapeutic strategies [9]. For example, the antitumor peptide CIGB-552 is a new targeted anticancer therapy whose molecular mechanism is related to the stabilization of the COMMD1 protein, thereby inhibiting the transcription factor NF-kB [10].

Faced with the high cost of experimental design and synthesis of anticancer peptides (ACPs), as well as the exponential growth of protein sequence data generated through high-throughput sequencing, experimental methods often take months or even years of speculation and experimentation, making it difficult to identify ACPs through experimental methods alone. However, these limitations can be appropriately optimized by applying machine learning (ML) methods.ML is a branch of artificial intelligence that automates analytical model building for fast and accurate result prediction [11]. For instance, Wan et al. developed a model using machine learning techniques such as support vector machine (SVM) and sequential minimum optimization (SMO) to discriminate between ACPs and hold peptides. (The accuracy of the model was 95.2 %) [12]. On the other hand, Charoenkwan et al. proposed a new, flexible scorecard method (FSCM) to efficiently predict and characterize peptides with anticancer activity using only sequence information. However, this method is not as intuitive as the latest integrated methods [13]. In the study of Zhao et al. they developed a new method called “DRACP” [14]. Although this method improves the recognition accuracy to some extent, it has some challenges in parameter selection and tuning. These studies show that although machine learning plays an important role in the screening and discovery of peptide drugs, these methods mainly stay in the traditional learning stage, and the accuracy and screening precision still need to be improved.

In this study,the purpose of this study is to solve the problem of using deep learning to predict the sequence of anticancer peptides and distinguish whether peptides have anticancer properties, this study employs a groundbreaking methodology utilizing a comprehensive BERT (Bidirectional Encoder Representations from Transformers) natural language processing (NLP) model [15]. To enhance model performance, it integrates the robust feature extraction capabilities of Convolutional Neural Networks (CNN) [16]. Recognizing the impact of optimizer choice on training outcomes, AdamW was selected for its superior performance among various optimizers [17]. Extensive comparisons regarding loss function selection led to the adoption of the binary cross-entropy loss function (BCELoss), identified as the most effective [18]. The model’s excellence is validated against traditional machine learning methods, such as support vector machine (SVM) [19], GaussianNB model [20], k-nearest neighbor (KNN) model [21], and decision tree model [22]. Utilizing five-fold cross-validation and external independent testing, the model demonstrates superior performance, achieving an accuracy of 0.9382, a recall of 0.9385, and an F1 score of 0.9371 in the external test. These results affirm the hypothesis that a fusion of BERT and CNNs, optimized with AdamW and employing BCELoss, effectively addresses complex NLP challenges. Fig. 1 (workflow diagram) meticulously outlines the research methodology of this paper.

2. Methods

2.1. Data collection

The study has meticulously curated and assembled two comprehensive data compendiums, denoted as ACP1 and ACP2. In this study, two datasets, ACP1 and ACP2, were used to construct and evaluate the model. The ACP1 data set is divided into training set, verification set and test set according to the ratio of 7:1.5:1.5. The dataset contains 574 examples that are resistant and 594 examples that are not resistant. This partitioning strategy aims to represent positive and negative outcomes equally in each subset.At the same time, the ACP2 dataset, consisting of 256 examples that behaved as resistant and 256 examples that behaved as non-resistant, was used directly as an external test set. Despite their shared attribute of large-scale magnitudes, each dataset manifests distinct characteristics. ACP1 is predominantly tailored towards encapsulating positive and negative instances with anticancer properties, providing an enriched source of data for the study's model's training. Conversely, ACP2 is architectured to showcase the model's robust generalization capability during external validation. Leveraging the samples from ACP2, the study can ascertain the study's model's proficiency in discerning anticancer peptides from non-anticancer peptides [23], even in a more rigorous and exigent training milieu.Fig. 2 illuminates a detailed distribution of instances within both ACP1 and ACP2, offering a precise portrayal of the distributional dynamics of positive and negative instances. This graphical exemplification furnishes readers an intuitive conduit to comprehend the data, particularly spotlighting the dispersion of peptide sequence lengths within the "activity" column. This visual elucidation empowers

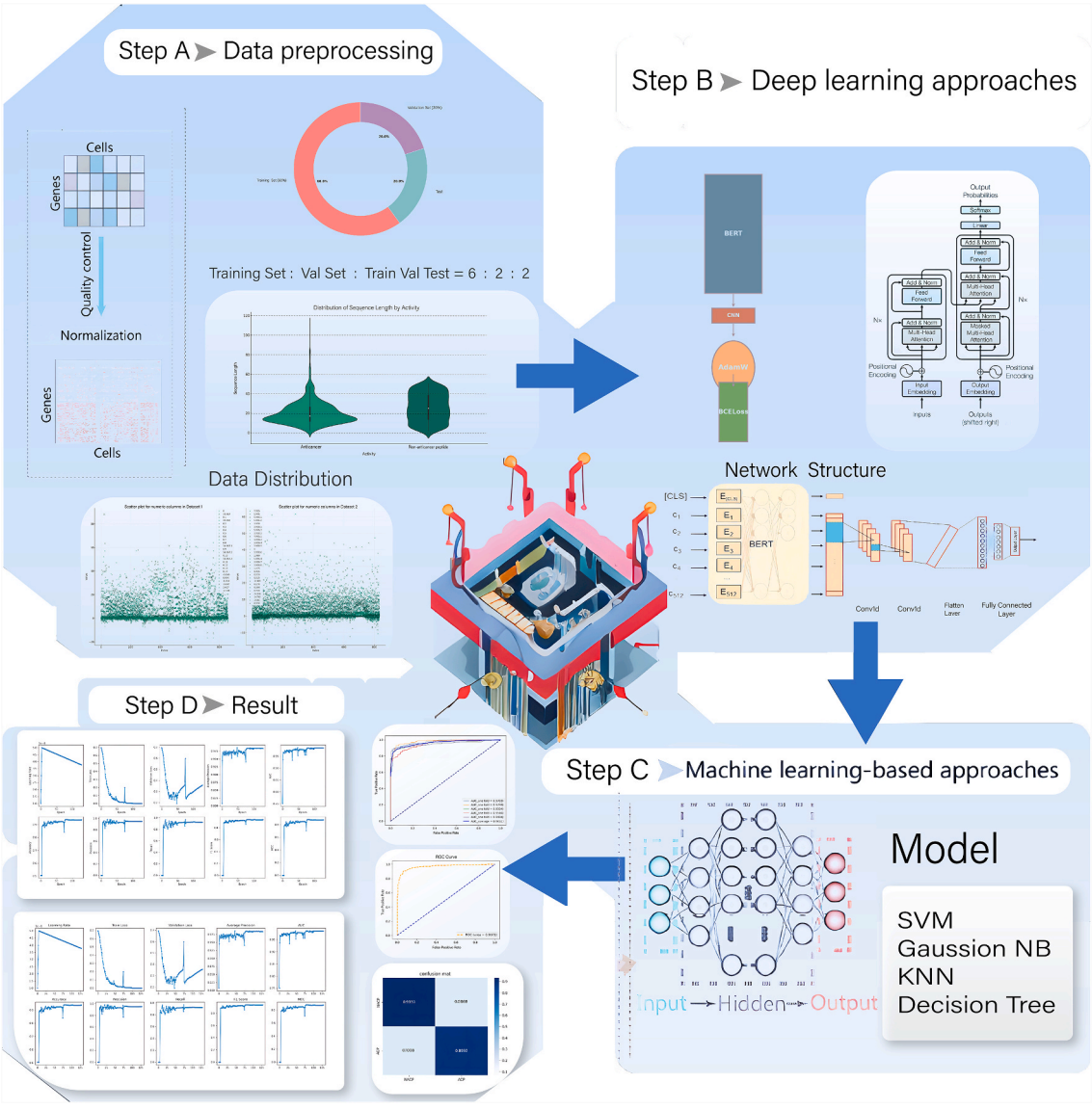


Fig. 1. Workflow diagram. It shows the research idea of this paper in detail.

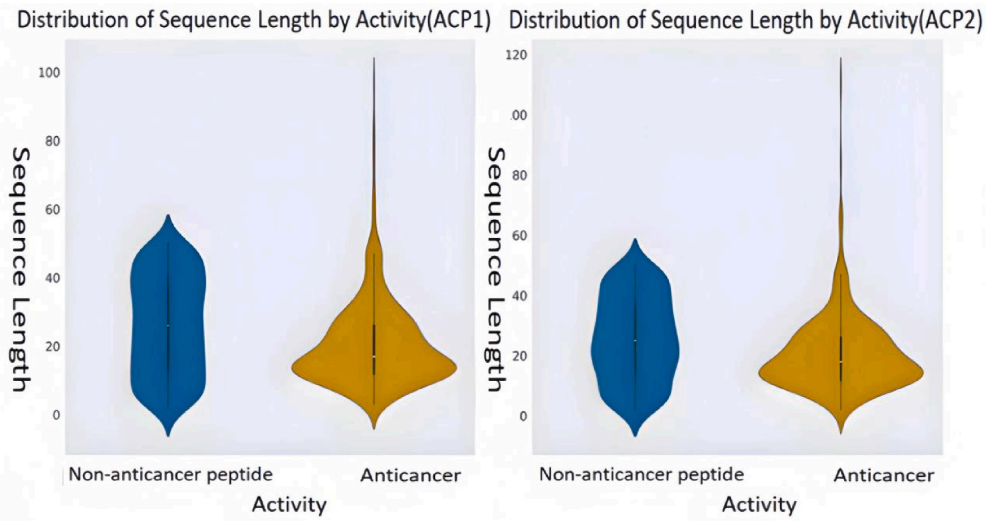


Fig. 2. Detailed distribution of data.
This graph shows the detailed distribution of anticancer peptides and non-anticancer peptides of ACP1 and ACP2, respectively, and the distribution of their peptide sequence lengths is revealed in the "Activity" column, which provides an intuitive way of interpreting the data.

readers to grasp lucidly the length distribution dynamics between anticancer peptides and non-anticancer peptides.Table 1, a comprehensive tabulation, provides granular insights into the count and provenance of positive instances within ACP. The study manually harvested data from APD3 (252 entries) [24], DPL (50+ entries) [25], and PlantPepDB (36 entries) [26], and coalesced data from BioPepDB (635 entries) [27] through an amalgamation of web scraping and manual collection techniques. Additionally, the study exported data from FeptideDB (45 entries) [28]. Consequently, the study aggregated an impressive total of 1639 positive instances from these heterogeneous databases.For the study's negative instances, the study leveraged active peptides from the prolific Uniprot database. Adhering to a defined ratio, the study stochastically bifurcated the overall repository into two sub-libraries. One sub-library was purposed for training and cross-validation, while the other was designated for independent validation. It bears underscoring that there is no overlap between positive and negative instances. Moreover, ACP1 and ACP2 are independent entities, devoid of intersections or overlaps.the study's In the design of ACP2 as an external test set, the intention was to validate the model's ability to generalize and predict anticancer peptides in future databases. This step not only demonstrated the accuracy of the model in identifying anticancer peptides in unknown datasets but also highlighted the model's applicability and flexibility when encountering new, untrained data. By conducting tests in this independent, more challenging environment, the potential of the model for the broad identification of anticancer peptides was further demonstrated, providing a powerful computational tool for the discovery of anticancer peptides and future clinical applications.

2.2. Data Preprocessing

Initially, raw amino acid sequence data were de-duplicated to ensure dataset uniqueness and representativeness, resulting in a clean dataset devoid of duplicate entries, thereby enhancing model training efficiency. Subsequently, a statistical approach was employed to identify and exclude outliers, specifically sequences with significant deviations in length or base composition, thus improving dataset quality by removing extreme values and noise. Further, a series of denoising operations were conducted to eliminate irrational amino acid sequences containing spaces, special symbols, and similar anomalies. This purification step allowed the model to focus more effectively on identifying and learning significant patterns and associations. Lastly, amino acid codes occurring at very low frequencies, deemed rare within the dataset, were filtered out. This approach enabled the model to concentrate on analyzing common and representative amino acid sequences.

2.3. Amino acid sequence feature extraction

The enhanced protein sequence representation known as pseudo amino acid composition (PseAAC) goes considerably beyond the conventional method based on the composition of 20 amino acids. PseAAC further incorporates sequence order information, thereby capturing amino acid interactions. PseAAC operates on the assumption that amino acid residues located at position i and position (i + k) in a protein sequence correlation exist. This comprehensive formulation elucidates the properties of proteins, including, but not limited to, their structure, function, and interactions. By adjusting the parameters, PseAAC can generate multiple pseudo-amino acid composition vectors, thus enhancing its versatility and adaptability [29].

This study extensively utilizes PseAAC for protein sequence analysis. In 2015, Hong-Bin and Kuo-Chen Chou introduced a web server named PseAAC, facilitating the generation of various PseAA combinations to enhance analysis capabilities [30]. PseAAC of type

Table 1
data sources.

No.	Database Name	Number of Entries	Link
1	APD3	252	https://aps.unmc.edu/database/anti
2	BioPepDB	635	http://bis.zju.edu.cn/biopepdbr/index.php?p=search&field=category&query=anticancer
3	DPL	50+	http://www.peptide-ligand.cn/search/?csrfmiddlewaretoken=PSqYxvTcUbmCHIAOCjLDJa0tzZky9MoQ6YR9NrAVsHHqeUB6uBkdMNyrmIJ1o2Zf&q0=&q1=&q2=&q4=Anticancer+&q3=&q5=&submit=Search
4	PlantPepDB	36	http://14.139.61.8/PlantPepDB/pages/browse_result.php
5	FeptideDB	45	http://www4g.biotec.or.th/FeptideDB/peptide_search.php

2 was selected, focusing on the correlation of consecutive amino acids to more efficiently capture protein sequence order information. A weight factor of 0.05 was applied, dictating the significance of sequence correlation terms within the PseAAC vector. The λ value was set to 1, emphasizing the consideration of neighboring amino acids' correlation in protein sequences. For representing amino acids' physicochemical properties, parameters such as hydrophobicity, hydrophilicity, mass, pK1 (pKa value of the α-carboxylic group), pK2 (pKa value of the amino group), and pI (isoelectric point at 25 °C) were chosen. These elements collectively constitute the PseAAC model employed in this research, offering a comprehensive and adaptable approach for probing into protein sequence intricacies [31].

2.4. Technological route

2.4.1. Applications of BERT and its variants in the field of biopeptides

BERT models have played an important role in the advancement of biomedical research, especially in the field of drug-target interactions (DTIs) [32]. By applying and fine-tuning the BERT pre-training model ChemBERTa, Kang et al. significantly improved DTI prediction [33]. BioBERT is an optimized BERT model that emphasizes domain-specific training for its effectiveness [34]. PharmBERT is another iteration of the BERT model for addressing unique language in drug labels and further exploits the potential of domain-specific BERT models by pre-training on drug labels [35]. The CT-BERT model is pre-trained on a large corpus of COVID-19-related Twitter messages to provide valuable insights for COVID-19 content analysis [36].2022 Mingyu et al. developed a model that utilizes BERT for text feature extraction and BiLSTM to obtain the internal information of audio, which achieves an effective fusion of multimodal features [37]. In conclusion, BERT modeling has made significant progress in the biomedical field, providing a powerful tool for data analysis, prediction, and understanding.

This research explores various BERT variants [38]. Initially, AdamW is integrated with BERT to enhance convergence speed and model stability, albeit at the cost of increased training complexity and time due to hyperparameter fine-tuning. Attention then shifts to employing the Binary Cross Entropy Loss (BCELoss) function for model optimization. However, this approach might intensify issues in datasets with significant category imbalance, necessitating further mitigation techniques. Furthermore, an exploration into merging Convolutional Neural Networks (CNNs) with BERT is conducted to leverage CNNs' performance benefits in computer vision tasks, particularly for multimodal inputs [39]. Yet, this amalgamation challenges with potential increases in computational demands and extended training durations due to BERT and CNN's structural variances [40].

2.4.2. 2 Model construction

Building on the theoretical and practical foundations laid out, a BERT enhancement model was developed. This model seamlessly combines BERT with the AdamW optimizer, BCELoss, and CNN, achieving significant improvements.The detailed architecture of this model is illustrated in Fig. 3.

In conclusion, BERT variations, particularly the model integrating the AdamW optimizer, BCELoss, and CNN, exhibit strong performance. However, challenges such as fine-tuning hyperparameters, addressing class imbalance, and optimizing computational resources and training time present areas for further exploration and enhancement. Addressing these issues is crucial for refining model training processes. This endeavor represents an ongoing academic journey, pushing the boundaries of research in this field.

2.4.3. Traditional machine learning models

To deeply investigate the prediction of anticancer peptides, four classical machine learning models are used in this study for analysis and comparison. The first, the Support Vector Machine (SVM), which primarily distinguishes between different classes by locating the ideal hyperplane, was first developed by Vapnik and Lerner in 1963 [41]. In this study, the parameters of SVM are configured as follows: the kernel function is linear (Tinear), the *C*-value is 1.0, the Degree is 3, and the Gamma is set to scale. Second, the GaussianNB model is explored, a probability-based classifier that operates on the foundational principle of applying Bayes' theorem with a distinct probability distribution [42].The parameters of this model are set as None for the prior probability (priors) and 1e-09 for the variance smoothing parameter (var_smoothing). Next is the Nearest Neighbor (KNN) model, which is an instance-based classification algorithm, the main idea of this algorithm is to find the "nearest neighbor" data points of the test data in feature space, and based on that, it is possible to find the "nearest neighbor" data points of the test data in the feature space. The main idea of this algorithm is to find the "nearest neighbor" data points of the test data in the feature space and make predictions based on the categories of these "neighbor" data points [43]. The parameters are configured as follows: the number of neighbors (n_neighbors) is 1; the weight selection method is UNIFORM so that all neighbors have equal weights, i.e., each neighbor has the same influence on the prediction results; the search algorithm is AUTO; and the p-value is 2. Finally, the decision tree model is a tree-structured model that can be used for classification and regression tasks. It achieves the best classification results by constantly judging and dividing the features [44]. Its parameters are configured as follows: the decision criterion is the Gini coefficient (gini), the segmentation method is selected as best, the maximum depth (max_depth) is None, the minimum number of samples split (min_samples_split) is 2, the minimum number of samples leaf (min_samples_leaf) is 1, and the maximum number of features (max_features) is None. features) as None. Through in-depth comparison and parameter tuning of these four models, the study aim to find the most suitable model for anticancer peptide prediction, thus ensuring the accuracy and generalization ability of the prediction.

2.5. Model evaluation

In evaluating deep learning models, several key metrics are utilized to assess the performance, including Accuracy (ACC), F1 score, Area Under Curve (AUC), and Matthews Correlation Coefficient (MCC) [45].Together, these evaluation metrics depict the performance

of the model in various aspects, including accuracy of prediction, comprehensiveness, and consistency of the predicted results with the true results.

Firstly, the Accuracy Criterion (ACC) is the most intuitive evaluation metric and is the ratio of the sample data that were correctly categorized to the total number of samples. It is expressed in a mathematical formula as shown in Equation (1):

$$ACC = \frac{(TP + TN)}{(TP + TN + FP + FN)} \tag{1}$$

Where TP denotes the number of true positive samples, TN denotes the number of true negative samples, FP denotes the number of false positive samples, and FN denotes the number of false negative samples. The higher accuracy rate represents the more effective the classifier is and the higher the precision of the prediction results.

Second, the F_1 score is the reconciled average of Precision as shown in Equation (2) and Recall as shown in Equation (3). Precision represents the number of samples that were judged to be positive examples; Recall represents the proportion of correct predictions in all actual positive samples. The F score is calculated as shown in Equation (4):

$$PRE = \frac{TP}{(TP + FP)} \tag{2}$$

$$REC = \frac{TP}{(TP + FN)} \tag{3}$$

$$F = \frac{(\alpha^2 + 1)P * R}{\alpha^2(P + R)} \tag{4}$$

In the above equation, when $\alpha = 1$, it is the common F_1 score as shown in Equation (5).

$$F_1 = \frac{2P * R}{P + R} \tag{5}$$

After that, the area under the curve (AUC) is an important metric used to evaluate the predictive performance of the model. The larger the area under the ROC curve, the better the predictive performance of the model [46].

Finally, Matthews correlation coefficient (MCC) is used to measure the model binary classification performance, which usually takes values ranging from −1 to 1 [47]. The MCC is calculated as shown in Equation (6):

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \tag{6}$$

The closer the value is to 1, the higher the consistency of the model's predictions with the true results. On the contrary, it means lower consistency.

Using these evaluation metrics, it is possible to comprehensively assess the model's performance in anticancer peptide (ACP) screening, facilitating a deeper understanding and optimization of its capabilities.

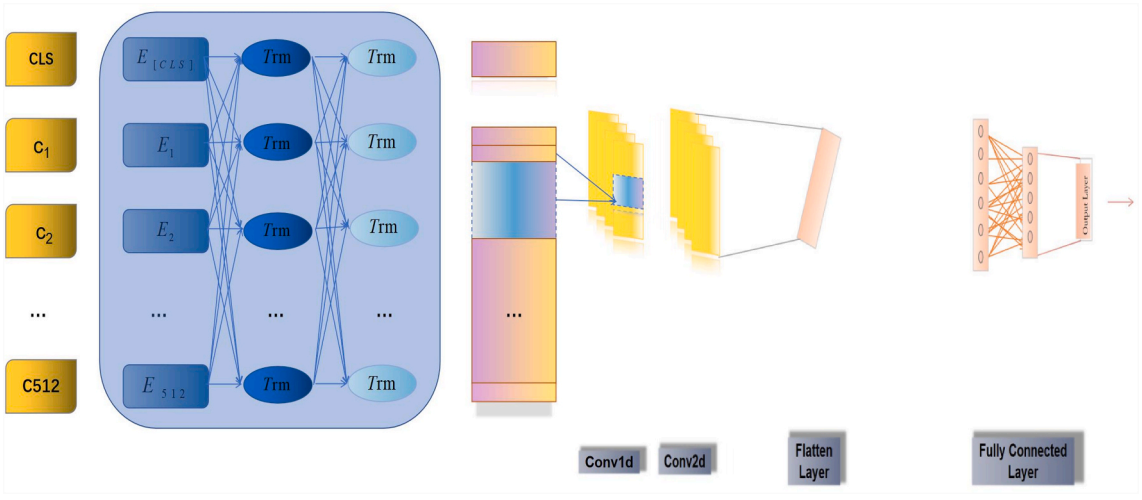


Fig. 3. Model architecture diagram.

3. Result

3.1 Experimental set up.

In this research, deep learning models were utilized to explore the anticancer peptide (ACP) screening process. Rigorous experimental setup and optimization of multiple models were conducted to ensure high recognition efficiency of ACPs. Models underwent extensive iterations with parameter tuning for optimal convergence and to effectively mitigate overfitting. Specifically, the AdamW-based model required 61 iterations for optimization, while the BERT model integrated with Convolutional Neural Networks (CNNs) was optimized after 54 iterations. Optimization was achieved after 44 iterations for the model employing Binary Cross Entropy Loss (BCELoss), 45 iterations for the baseline model, and the best performing module reached its peak after 77 iterations. To augment the models' robustness and predictive accuracy, GridSearch CV was implemented for hyperparameter fine-tuning, in conjunction with five-fold cross-validation and external dataset testing. These experiments ran on Python 3.8 (Ubuntu 20.04 system), utilizing PyTorch 1.10.0 as the primary deep learning framework and Cuda 11.3 for computational acceleration. Experiments were conducted on an Intel (R) Xeon(R) Platinum 8358P CPU @ 2.60 GHz server, equipped with RTX A5000 GPUs and 15 vCPUs. The introduced model is based on the 12-layer Transformer architecture [48], incorporating 12 self-attention mechanisms in each layer [49], with a hidden layer size of 768 and an intermediate layer size of 3072 in the feedforward neural network. A fully connected layer was added atop BERT for classification purposes. Furthermore, a novel CNN-BERT model was developed, adding a convolutional layer on top of BERT and integrating a maximum pooling layer before the fully connected classification layer [50]. Open-source libraries such as Sklearn, Transformers, Numpy, Pandas, and Matplotlib were extensively leveraged, providing powerful and versatile tools for feature extraction from the ACP database.

3.1. Result for -fold cross validation

3.1.1. Five deep learning models

Five deep learning models were carefully designed and run for each prediction task in the study's main dataset: the BERT baseline model, the fusion of BERT with the AdamW optimizer, the union of BERT with BCELoss, the combination of BERT with Convolutional Neural Networks (CNNs), and the top-performing model (BERT + AdamW + BCELoss + CNN hybrid model). To ensure the fairness and impartiality of the comparison, all the models have undergone rigorous and fine parameter optimization.

Table 2 below will show the cross-validation results of the five models in detail, and Fig. 4 shows the results for the anticancer peptide (ACP) grade prediction task and the receiver operating characteristic (ROC) curve of the study's proposed model [51]. Among all the models examined, the BERT + AdamW + BCELoss + CNN model exhibits the best performance with an average area under the ROC curve (AUC) of 0.9726, a result that is significantly better than the other four models: the BERT baseline model (0.9593), BERT + AdamW (0.9642), BERT + BCELoss (0.9622) and BERT + CNN (0.9611). This finding clearly reveals that the model incorporating the optimizer, loss function, and convolutional neural network has a significant advantage in this prediction task.

This study demonstrates the remarkable progress made by deep learning models in predicting ACP mutation status and grade. These models provide a viable new option for non-invasively identifying ACP grade and genetic characteristics in patients, and will likely accelerate the development of novel peptide-based anti-cancer therapies. In evaluating the performance of the study's models, the study employed a confusion matrix that provides a detailed view of each model's performance on the test dataset (see Fig. 5), including the number of true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN) [52]. Of particular note, the BERT + AdamW + BCELoss + CNN model shows excellent performance in all categories, generating low numbers of false positives and false negatives, which suggests that the model performs superbly in terms of both prediction precision and recall.

The training process is clearly illustrated through the training iteration graph, highlighting the variation of training and validation losses over the number of iterations. For the BERT + AdamW + BCELoss + CNN model, a decreasing trend in both training and validation losses was observed, signaling the model's convergence throughout the learning phase. Notably, the training loss diminishes more rapidly in the initial stages, with the rate of reduction slowing as iterations progress, indicating the model's approach towards an optimized solution. Conversely, the pattern of validation loss reduction is crucial; it diminishes in the early phases but begins to plateau after reaching a certain number of iterations, suggesting the onset of overfitting and serving as a cue to cease training [53]. Figs. 6 and 7 provide a detailed view of the BERT training iteration dynamics.

3.1.2. Machine learning models

In addition to the main focus of this research, four traditional machine learning models were developed: Support Vector Machines (SVMs), Gaussian Naïve Bayes (GaussianNB), k-nearest Neighbors (KNN), and Decision Tree Models. To ensure a fair evaluation, each model underwent thorough parameter optimization, with performances directly benchmarked against those of deep learning models. Among the traditional approaches, SVM stood out, achieving the highest average area under the receiver operating characteristic (ROC) curve (AUC) of 0.9441.It is closely followed by the Gaussian Plain Bayesian model, with an AUC of 0.9321. k-Nearest Neighbors and Decision Tree models perform relatively weakly, with AUCs of 0.8675 and 0.7844, respectively. These results are presented in the shown in detail in the subsequent figures (Table 3, Fig. 8, Fig. 9), including the performance metrics and ROC curves for each model.

3.2. External validation

This research not only conducted a rigorous five-fold cross-validation on an internal dataset [54] but also carried out comprehensive validation on an external independent dataset. This approach was taken to confirm the robustness and applicability of the

model for real-world clinical scenarios [55].

3.2.1. Deep learning models

The study ran five deep learning models on external datasets: the BERT baseline model, the combination of BERT and the AdamW optimizer, the union of BERT and BCELoss, the fusion of BERT and a convolutional neural network [56] (CNN), and the study's best model, the BERT + hybrid model of AdamW + BCELoss + CNN. All these models are finely parameterized to ensure a fair performance comparison.

The results show that the BERT + AdamW + BCELoss + CNN model performs best on the tasks of anticancer peptide (ACP) grade prediction and ACP mutation status prediction. On the external validation dataset, the area under the receiver operating characteristic (ROC) curve (AUC) of the model reaches 0.9848, a result that starkly reveals the superior performance of the study's model when dealing with real-world data.

The AUCs of the remaining models on the external validation dataset are 0.9641 for the BERT baseline model, 0.9699 for BERT + CNN, 0.9750 for BERT + BCELoss, and 0.9671 for BERT + AdamW.These results provide further evidence of the superiority of the deep learning models in processing complex bioinformatics data.

Further analysis of the confusion matrix (Table 4) reveals that the BERT + AdamW + BCELoss + CNN model demonstrates excellent prediction accuracy and recall, underscoring its significant potential for practical deployment. Fig. 10 showcases the ROC curves of five deep learning models, with the model developed in this paper outperforming the rest.ACP2 consists of peptide sequences that have never participated in the model training process. These sequences were collected from different databases through independent methods to test the model's performance when faced with entirely new data. The model demonstrated exceptional performance on this external test set, accurately identifying anticancer peptides and non-anticancer peptides.This result not only confirms the model's high generalization ability but also showcases its potential in recognizing unknown anticancer peptides in practical applications.

3.2.2. Traditional machine learning models

For a comprehensive performance comparison, this paper ran four traditional machine learning models on an external dataset: support vector machine (SVM), Gaussian plain Bayes (GaussianNB), k-nearest neighbor (KNN), and decision tree models. Despite rigorous parameter optimization, the performance of these traditional models is still far below that of the study's deep learning model when processing real data. Fig. 11 illustrates the ROC curves for the four models.

As shown in Table 5: Specifically, the area under the receiver operating characteristic (ROC) curve (AUC) is 0.9494 for the SVM model, 0.9385 for the Gaussian Plain Bayesian model, 0.8999 for the k-nearest neighbor model, and 0.8519 for the decision tree model. These results further corroborate the fact that the deep learning models are handling complex biomedical prediction tasks leading position on the complex biomedical prediction task [57].

While traditional machine learning models are still valuable in some aspects, the study's model performs significantly better than traditional models in the tasks of ACP grade prediction and mutation status prediction. This finding emphasizes the leading position of deep learning in handling complex biomedical prediction tasks [58] and reveals its great potential in the development of future peptide-based cancer therapies.

In conclusion, in this study, this paper comprehensively evaluated the performance of deep learning models versus traditional machine learning models in predicting ACP grade and mutation status through internal five-fold cross-validation and external independent validation. The study's results reveal the superior performance of deep learning models and their potential applications in cancer therapy research [59], providing strong support for the development of future peptide-based cancer therapies.

3.3. Discussion

In the clinic, determining whether a biological peptide has anticancer properties usually requires extensive experiments, which are not only time-consuming but also costly [60]. All these factors have limited the rapid development of the field of anticancer peptide research. In this study, this paper developed the study's model based on the dataset processed by two feature encoding methods, PseAAC expression bioinformatics and natural language processing, which can come to predict whether a biopeptide is an anticancer peptide or not. This strategy realizes efficient and highly accurate screening of anticancer peptides. Future work may benefit from

Table 2
Cross-validation results.

ver	Acc (95%CI)	P (95%CI)	R (95%CI)	F1 (95%CI)	MCC (95%CI)
baseline	0.92544 ± 0.0073	0.94314 ± 0.0171	0.9065 ± 0.0252	0.92408 ± 0.0141	0.85192 ± 0.0146
bceloss	0.92368 ± 0.0069	0.9372 ± 0.0083	0.90848 ± 0.0174	0.92252 ± 0.0068	0.84752 ± 0.0066
AdamW	0.93262 ± 0.011	0.96632 ± 0.0131	0.89738 ± 0.0212	0.93022 ± 0.011	0.86806 ± 0.011
Conv	0.9195 ± 0.0103	0.95668 ± 0.023	0.8804 ± 0.019	0.91634 ± 0.0158	0.84258 ± 0.0201
Best	0.9326 ± 0.0103	0.95014 ± 0.0122	0.91332 ± 0.0124	0.93126 ± 0.009	0.86583 ± 0.0132

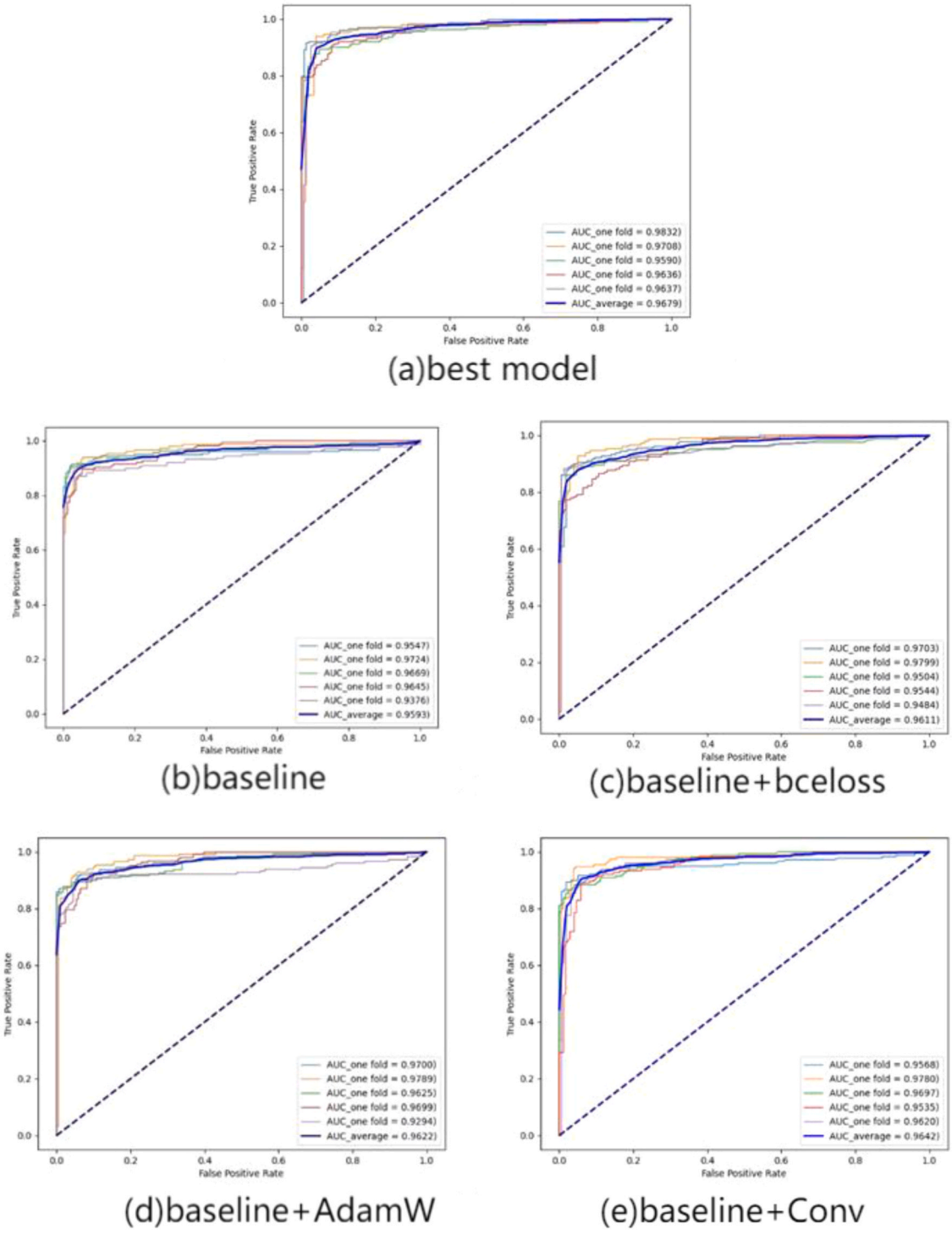


Fig. 4. ROC curves for each model.

expanding the dataset size to further validate the study's findings and enhance the generalizability of the model developed.

While there are many precedents of combining traditional machine learning with peptide screening, traditional machine learning models usually miss many nuances in the information and the accuracy is not very high [61]. In recent years, the advent of pre-training models has propelled uni-modal domains like computer vision (CV) and natural language processing (NLP) into a new era, with large-scale models integrating vision and language, such as M-FLAG [62], CLIP [63], and Med-UniC [64], demonstrating commendable

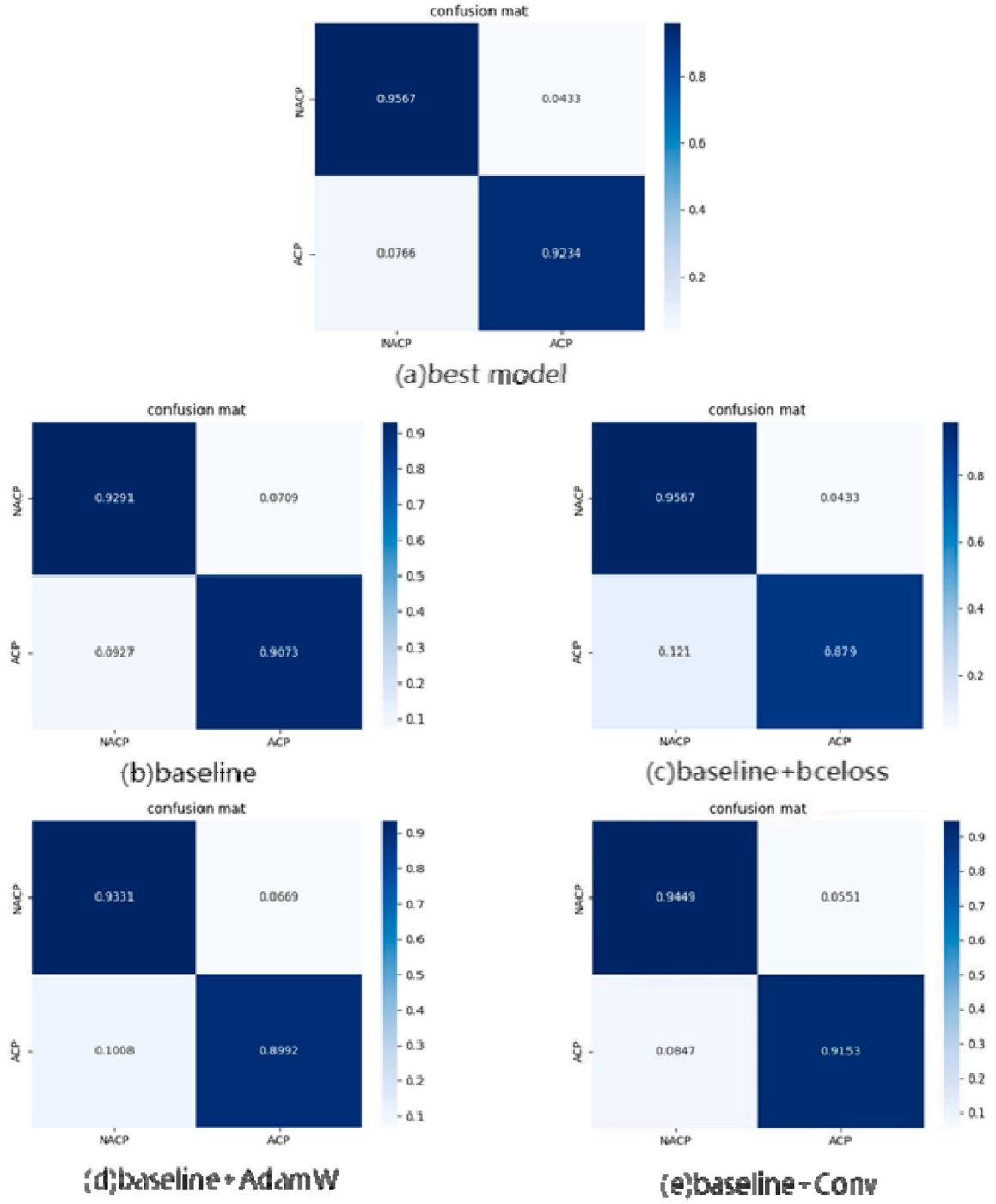


Fig. 5. Confusion matrix for the 4 models. The study's model performs well across samples.

effectiveness in their respective fields. Although this study focuses solely on natural language large-scale models and does not involve image processing, it paves the way for new research directions and ideas for our future studies. BERT's main innovation is based on the bi-directional structure of the Transformer (a self-attentive mechanism model) [65], which allows the model to take into account all the information in the context when processing the language. Although BERT has been able to learn some important features from the original amino acid sequence [66], PseAAC contains information that is difficult for BERT to capture. It can enhance the model's understanding of the global and local properties of proteins by independently calculating the two outputs and then combining them to form a larger feature vector. This vector is fed into a convolutional neural network optimized for AdamW and BCELoss to minimize

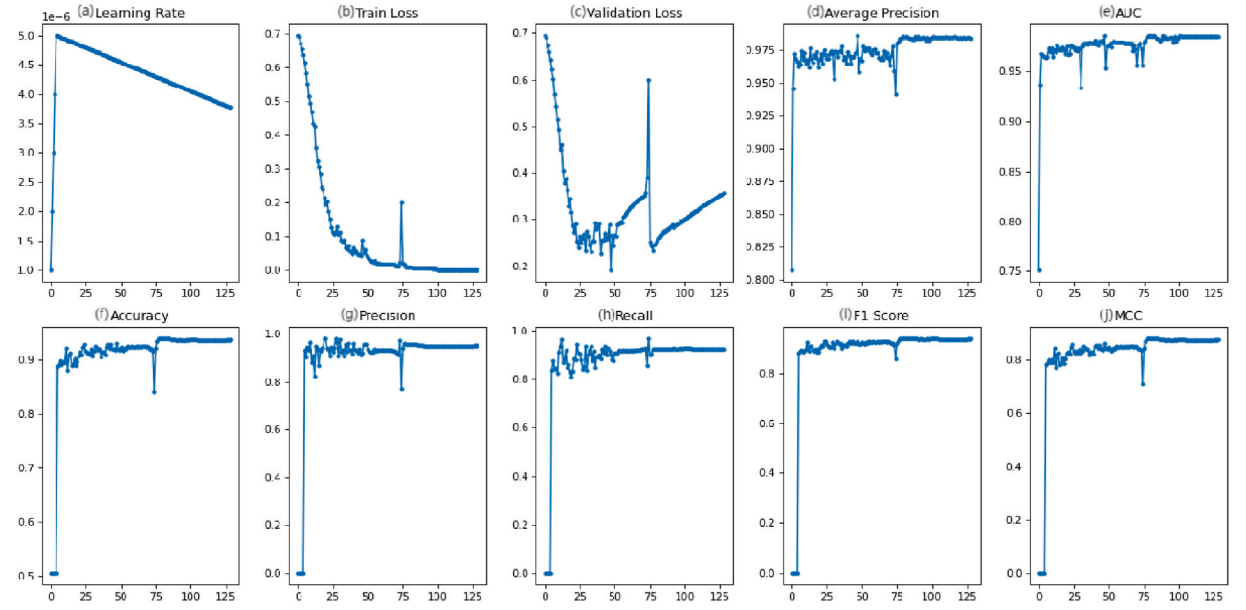


Fig. 6. Bert training iterative graph.
Top left to right: learning rate Training loss. Validation loss. Average Accuracy. Area Under Curve; Bottom left to right: Accuracy. Precision Rate. Recall rate. F1 value. Area Under Curve.

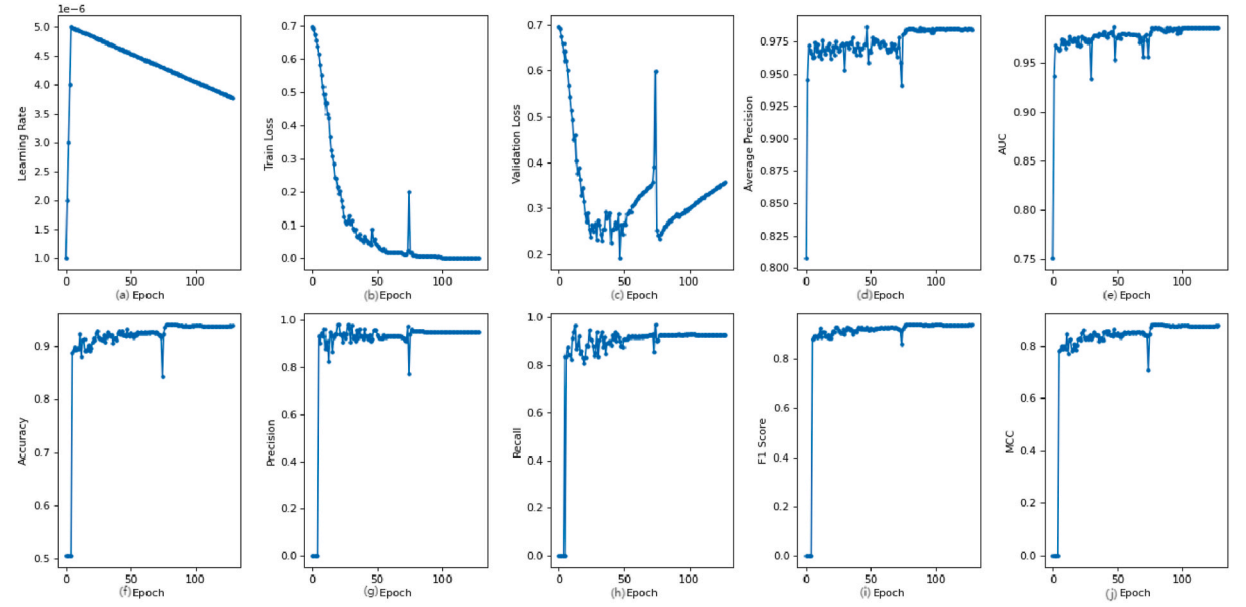


Fig. 7. Bert training iterative graph.

Table 3
Fifty-fold cross-validation of four machine learning models.

Models	Accuracy (95%CI)	Precision (95%CI)	Recall (95%CI)	F1 Score (95%CI)	MCC (95%CI)
KNN	0.7806 ± 0.0344	0.7876 ± 0.0641	0.7806 ± 0.0344	0.7797 ± 0.0344	0.5676 ± 0.0679
NB	0.8484 ± 0.0194	0.8521 ± 0.0382	0.8484 ± 0.0194	0.8480 ± 0.0194	0.6994 ± 0.0389
DT	0.7961 ± 0.0229	0.7971 ± 0.0447	0.7961 ± 0.0229	0.7961 ± 0.0229	0.5923 ± 0.0464
SVM	0.8783 ± 0.0142	0.8910 ± 0.0277	0.8783 ± 0.0142	0.8772 ± 0.0142	0.7683 ± 0.0289

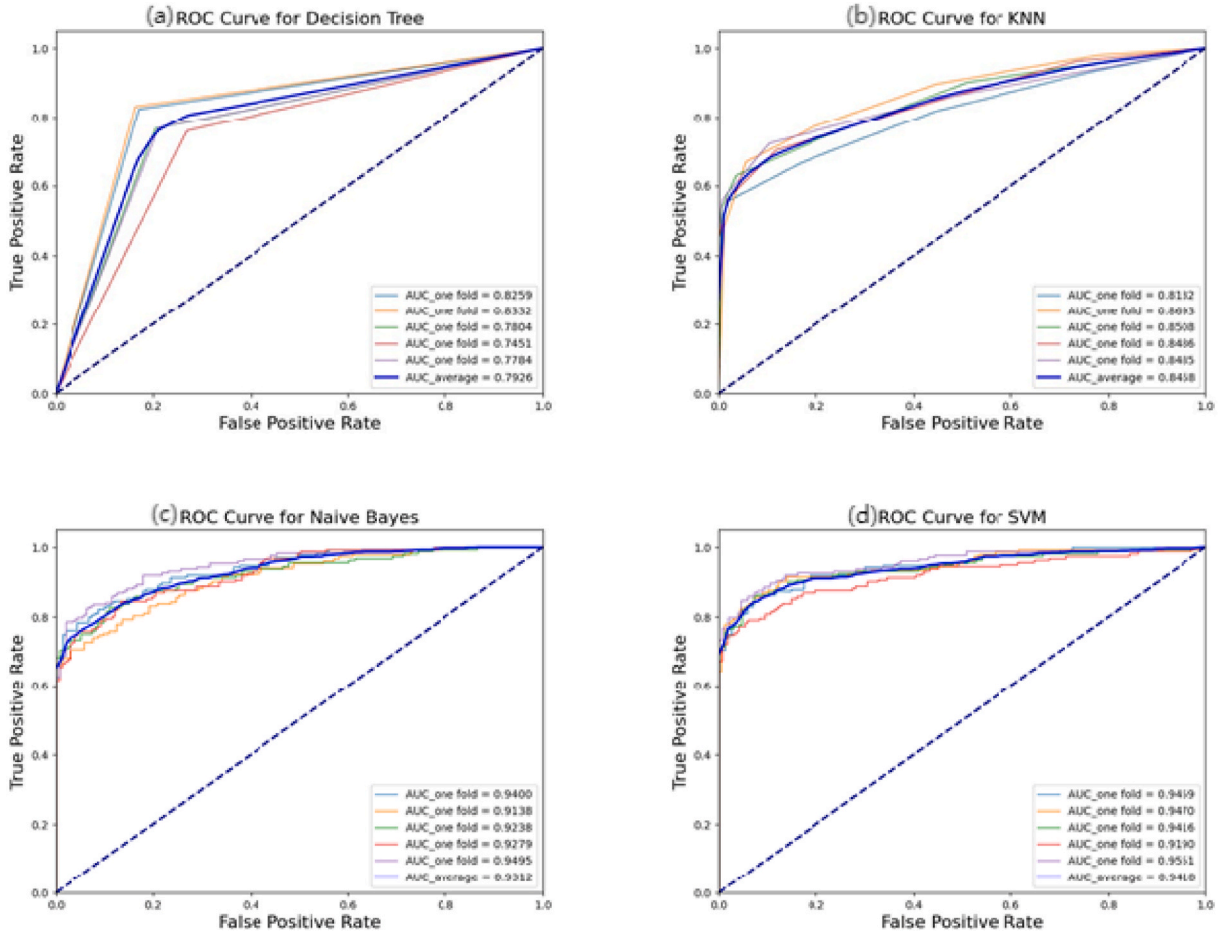


Fig. 8. Four model ROC curves: decision tree AUC:0.7844; KNN AUC:0.8675; plain Bayes: 0.9321; support vector machine: 0.9441.

prediction errors and maximize prediction accuracy.

The application of deep learning techniques to the discovery and development of anticancer peptides has a lot of value: the application of the model will likely facilitate research advances in the field of tumor therapy because it can be used for initial screening to quickly and efficiently identify peptides with anticancer activity from a large number of candidate peptides [67], This method significantly accelerates the initial screening phase, allowing researchers to promptly concentrate on the substances most likely to achieve success. Second, the ability of the model does not stop at screening but can guide the design of anticancer peptides by extracting useful patterns and features from amino acid sequences, the optimization of which is expected to create more potent anticancer peptides [68]. By identifying potential peptides in advance, the time and resources required to validate biological activity in the laboratory can be reduced. In addition, the analytical power of the model can also reveal the mechanism of action of anticancer peptides and deepen the study's understanding of the active features of anticancer peptides, thus promoting the development of anticancer therapies. Finally, the application of the model in the field of precision medicine is also expected [69], using the model to analyze protein/peptide sequences in tumor samples from specific patients to predict peptides with possible antitumor activity, which can help develop personalized anticancer therapies.

ESM-2 was evaluated in preliminary experiments and was found to exhibit suboptimal performance. Additionally, the model's considerable size and extensive parameter count make it particularly challenging to train, which undermines its suitability for practical clinical application. Given these limitations, a decision was made to develop a proprietary model tailored to overcome these specific challenges, aiming for both high performance and feasibility in clinical translation. This hybrid model combining BERT, CNN, AdamW, BCELoss, and PseAAC exhibits high performance on the task of anticancer peptide prediction, especially when compared with traditional machine learning methods such as Support Vector Machines (SVMs), Gaussian Spartan Bayes (GaussianNB), k Nearest Neighbors (knn), and Decision Trees (DTs) in terms of accuracy, precision, recall, F1 score, and Matthews correlation coefficient (MCC) were significantly improved. This hybrid model shows excellent performance on the anticancer peptide prediction task, and this performance is stable and reliable as the model performs well in both the five-fold cross-validation and ablation experiments. In external tests, this hybrid model achieved an accuracy of 94.02 %, a precision of 95.42 %, a recall of 92.34 %, an F1 score of 93.85 %, and a Matthews correlation coefficient (MCC) of 88.09 %. These results are significantly better than the baseline model as well as

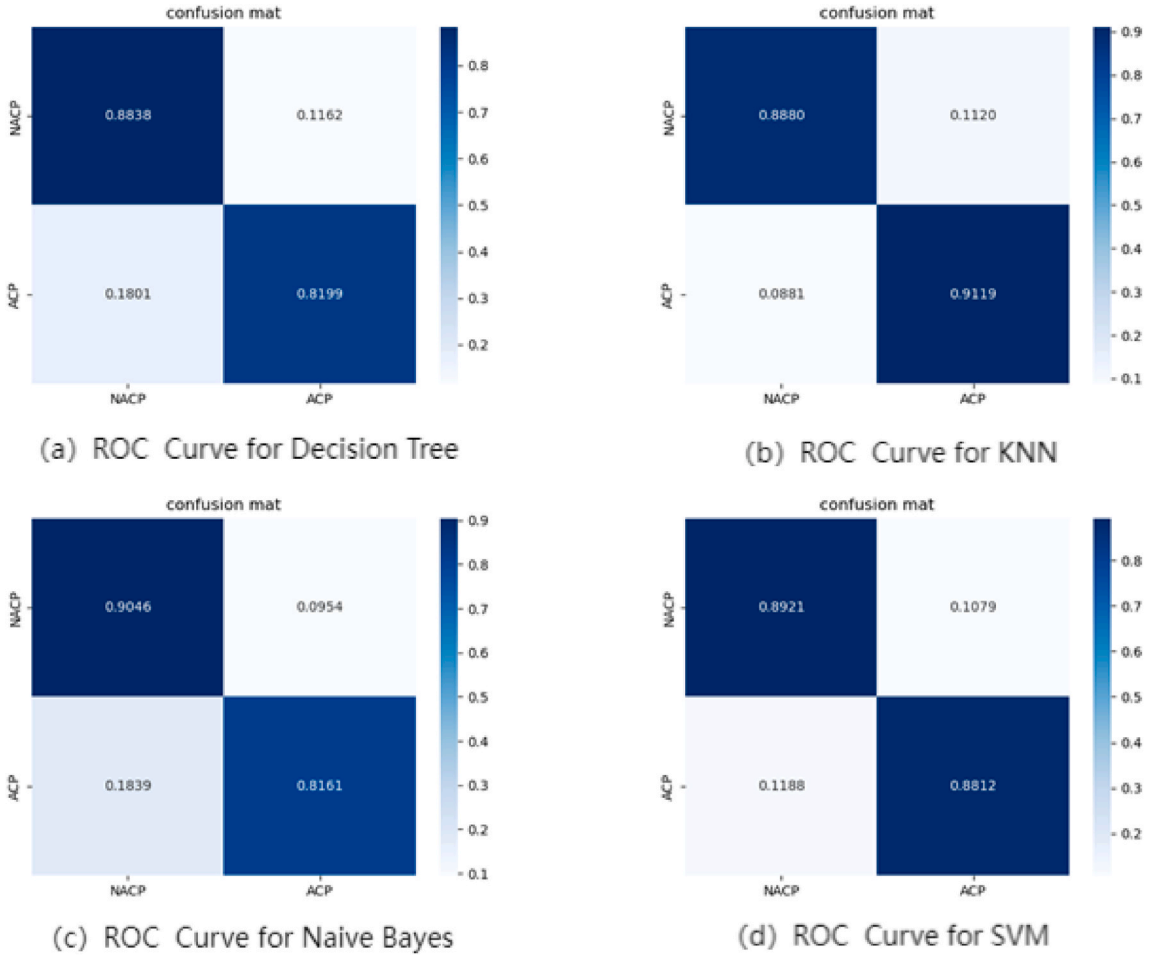


Fig. 9. Confusion Matrix for 4 Models of Machine Learning. Confusion matrices for four machine learning models, namely, decision tree, KNN, plain Bayes, and support vector machine, are shown in turn for comparison with the performance of the study's model in each sample below.

Table 4
Comparison of outcome indicators.

	BERT	BERT	CNN	BCELOSS	ADAMW	Acc	P	R	F1	MCC
baseline	✓	✓				0.9183	0.952	0.879	0.914	0.8389
bceloss			✓			0.9343	0.964	0.9032	0.9314	0.87
AdamW				✓		0.9183	0.9259	0.9073	0.9165	0.8368
Conv					✓	0.9303	0.9419	0.9153	0.9284	0.8608
best	✓	✓	✓	✓	✓	0.9402	0.9542	0.9234	0.9385	0.8809

traditional machine learning methods, as shown in Table 6, demonstrating the superiority of the hybrid model for the task of anticancer peptide prediction. The study thoroughly examined the model's generalization capabilities, with a particular focus on its predictive performance when encountering peptide sequences derived from databases of diverse origins. The employment of an external test set, ACP2, underscores the model's robust performance in handling entirely novel data.

The model developed in this study has demonstrated significant strengths in predicting anticancer peptides, yet it is essential to confront several inherent limitations. Initially, the research relies predominantly on theoretical calculations, marking merely the initial step. Future work will necessitate validating these predictions through biological experiments in a laboratory setting to merge theory with practice and ensure the predictions translate into practical outcomes. Moreover, despite promising results on an external test set, a deeper understanding of the model's generalizability requires extending the validation to more external test sets. This expansion is critical for a comprehensive evaluation of the model's stability and reliability across varying data environments. Additionally, the model requires further optimization to handle a broader range of tumor cells [70], potentially necessitating the introduction of new feature representations or adjustments to the model's structure and parameters for enhanced adaptability to specific tumor cell types. Notably, the current model does not account for the prediction of transmembrane ACEase inhibitory peptides [71], a

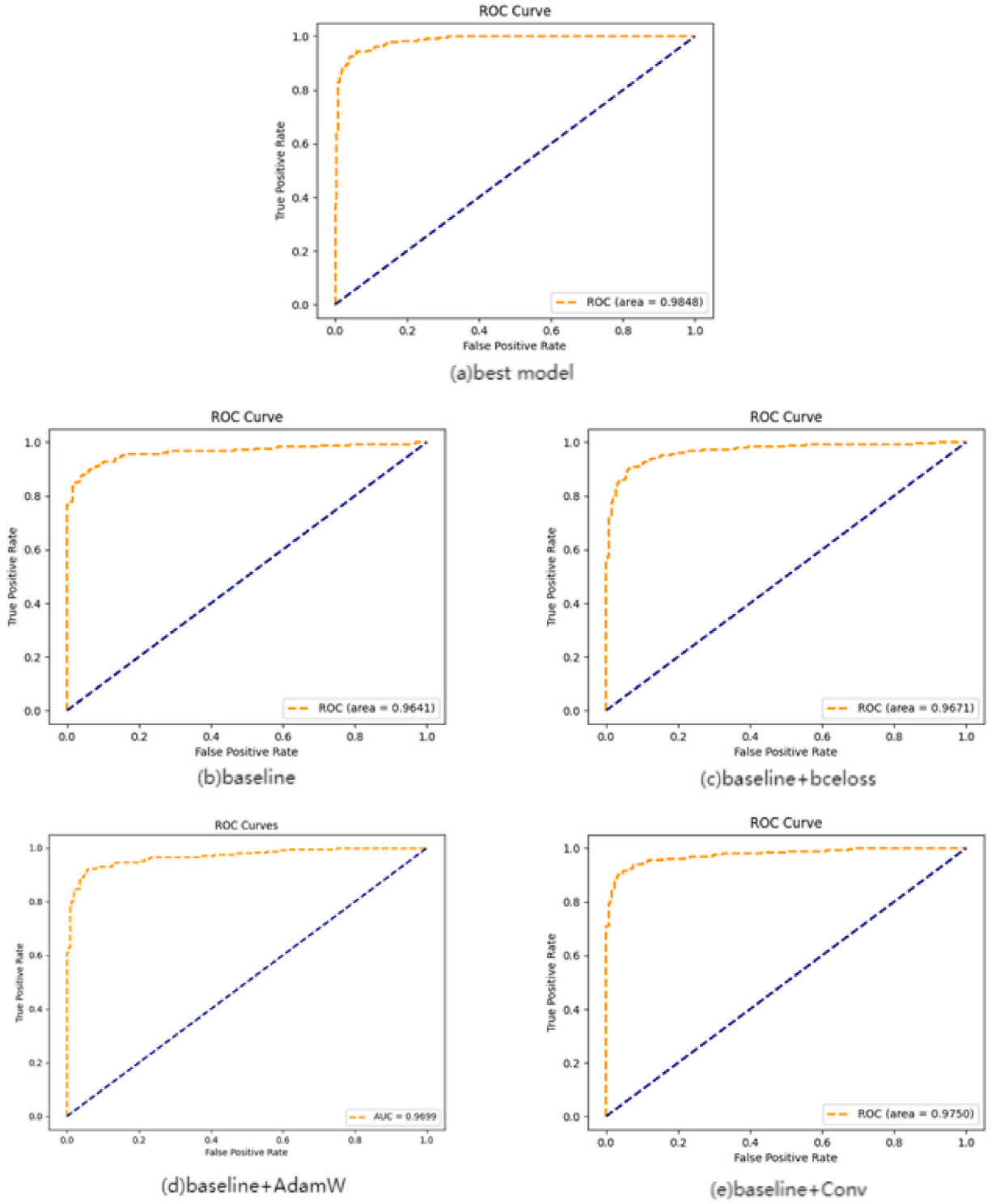


Fig. 10. ROC curves of 5 deep learning models: The ROC curves of 5 deep learning models are listed, and the results show that the study's model developed in this paper has the best performance among them.

vital category of anticancer peptides. Future updates will involve collecting relevant data and adjusting the model's input and output structures to encompass these peptides. In previous studies, the focus on anticancer peptide prediction was predominantly centered around short peptides, and the application of machine learning and deep learning techniques in related experiments yielded promising

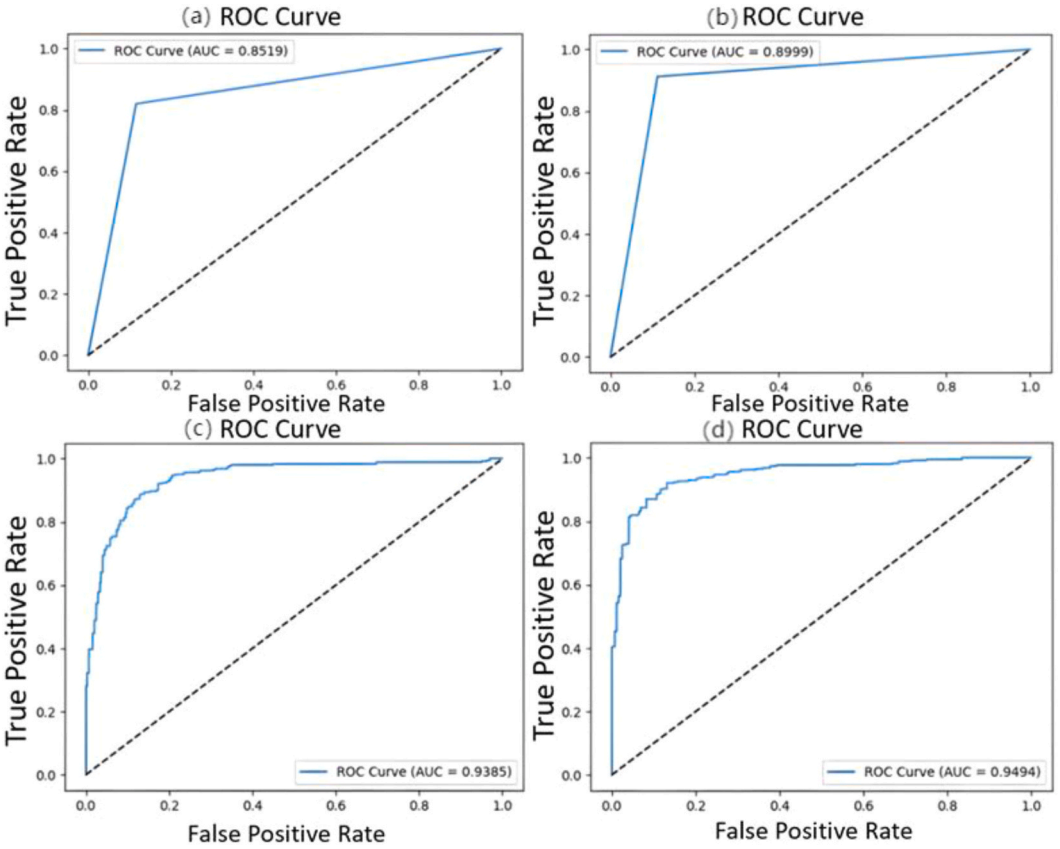


Fig. 11. Four model ROC curves.

Table 5
Machine learning results for each model.

	Acc	P	R	F1	MCC
svm	0.8865	0.8984	0.8812	0.8897	0.7729
GaussianNB	0.8586	0.9025	0.8161	0.8571	0.7214
knn	0.9004	0.8981	0.9119	0.9049	0.8005
decision tree	0.8506	0.8843	0.8199	0.8509	0.7036

outcomes [11–14]. Moving forward, the scope of research will be expanded to encompass other types of peptides, taking into consideration factors such as tertiary structures.Moving forward, addressing these challenges will be a priority to improve the model's performance and utility, thus offering a more robust tool for the discovery and development of anticancer peptides [72]. In future endeavors, the model will undergo comprehensive enhancements in several key areas. Primarily, tools such as Gradient-weighted Class Activation Mapping (Grad-CAM) will be employed to augment the model's interpretability by vividly demonstrating the focal points during the identification of anticancer peptides. Furthermore, a stronger collaborative framework with biologists, pharmacologists, and clinical physicians will be established, ensuring the model's efficacy in both biological and clinical contexts. Moreover, to maintain the accuracy and relevance of the model, continual assessments and updates will be conducted in response to emerging data and technological advancements. These initiatives are aimed at enriching the theoretical depth and practical applicability of the deep learning model, thereby advancing the research and development of anticancer peptides.Simultaneously, we plan to harness the power of ensemble learning methods. This strategy will involve employing suitable ensemble strategies to integrate the strengths of various deep learning architectures.

4. Conclusion

This research has developed a pioneering model that seamlessly integrates deep learning, preference-ordered amino acid coding (PseAAC), and natural language processing (NLP), establishing a novel approach in anticancer peptide screening. The model's unique configuration allows for exceptional performance in identifying anticancer peptides, offering a more cost-effective and economically

Table 6
Fifty-fold cross-validation results for 5 deep learning models.

	Acc	P	R	F1	MCC
svm	0.8865	0.8984	0.8812	0.8897	0.7729
GaussianNB	0.8586	0.9025	0.8161	0.8571	0.7214
knn	0.9004	0.8981	0.9119	0.9049	0.8005
decision tree	0.8506	0.8843	0.8199	0.8509	0.7036
best	**0.9402**	**0.9542**	**0.9234**	**0.9385**	**0.8809**

viable alternative to conventional screening methods [73]. Demonstrated through rigorous five-fold cross-validation and external testing, this model has consistently outperformed existing models, showcasing its proficiency in predicting the anticancer properties of peptides across various databases. Validation using an independent external test set further verifies the model's accuracy in detecting known anticancer peptides and its capability to predict the anticancer potential of previously uncharacterized peptide sequences.

Data availability statement

The study's dataset has been deposited in the public repository Figshare with the accession number 10.6084/m9.figshare.24,746,712. It is now accessible to the global research community, promoting transparency, reproducibility, and further advancements in related studies.

Funding

This project is supported by the Provincial Education Reform Project in Sichuan, China (Project No.: JG2021-453).

CRediT authorship contribution statement

Yupeng Niu: Writing – review & editing, Writing – original draft, Visualization, Validation, Supervision, Software, Resources, Project administration, Methodology, Investigation, Formal analysis, Data curation. **Zhenghao Li:** Visualization, Validation, Software, Project administration, Methodology, Investigation. **Ziao Chen:** Software. **Wenyuan Huang:** Writing – review & editing. **Jingxuan Tan:** Writing – original draft. **Fa Tian:** Writing – review & editing, Supervision, Resources, Project administration. **Tao Yang:** Visualization, Validation, Software. **Yamin Fan:** Data curation. **Jiangshu Wei:** Writing – review & editing, Supervision, Project administration. **Jiong Mu:** Funding acquisition.

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: Jiong Mu reports financial support was provided by Sichuan Provincial Department of Education. If there are other authors, they declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

Thanks to all partners in AI Studio for their support.

Abbreviations

ACP	Anti-Cancer Peptides
NLP	Natural Language Processing
PseAAC	Pseudo Amino Acid Composition
BERT	Bidirectional Encoder Representation from Transformers
CNN	Convolutional Neural Networks
ML	machine learning
BCELoss	binary cross-entropy loss function
SVM	support vector machine
TP	true positives
TN	true negatives
FP	false positives
FN	false negatives
ACC	Accuracy
AUC	Area Under Curve

MCC Matthews Correlation Coefficient

References

[1] M. Xie, D. Liu, Y. Yang, Anti-cancer peptides: classification, mechanism of action, reconstruction and modification, Open biology 10 (7) (2020 Jul 22) 200004.
[2] W. Chiangjong, S. Chutipongtanate, S. Hongeng, Anticancer peptide: physicochemical property, functional aspect and trend in clinical application (Review), Int. J. Oncol. 57 (3) (2020 Sep) 678–696, https://doi.org/10.3892/ijo.2020.5099. Epub 2020 Jul 10. PMID: 32705178; PMCID: PMC7384845.
[3] J.L. Lau, M.K. Dunn, Therapeutic peptides: historical perspectives, current development trends, and future directions, Bioorg. Med. Chem. 26 (10) (2018 Jun 1) 2700–2707.
[4] N. d'Avanzo, ri G. Torrie, P. Figueiredo, C. Celia, D. Paolino, A. Correia, K. Moslova, T. Teesalu, M. Fresta, H.A. Santos, LinTT1 peptide-functionalized liposomes for targeted breast cancer therapy, Int. J. Pharm. 597 (2021 Mar 15) 120346.
[5] E. Fisher, K. Pavlenko, A. Vlasov, G. Ramenskaya, Peptide-based therapeutics for oncology, Pharmaceut. Med. 33 (1) (2019 Feb) 9–20, https://doi.org/10.1007/s40290-018-0261-7. PMID: 31933267.
[6] S.R. Rajendran, C.E. Ejike, M. Gong, W. Hannah, C.C. Udenigwe, Preclinical evidence on the anticancer properties of food peptides, Protein Pept. Lett. 24 (2) (2017) 126–136, https://doi.org/10.2174/0929866523666160816152755. . PMID: 27538700.
[7] W. Liu, H. Tang, L. Li, X. Wang, Z. Yu, J. Li, Peptide-based therapeutic cancer vaccine: current trends in clinical application, Cell Prolif. 54 (5) (2021 May) e13025, https://doi.org/10.1111/cpr.13025. Epub 2021 Mar 22. PMID: 33754407; PMCID: PMC8088465.
[8] T.T. Chai, J.A. Koh, C.C. Wong, M.Z. Sabri, F.C. Wong, Computational screening for the anticancer potential of seed-derived antioxidant peptides: a cheminformatic approach, Molecules 26 (23) (2021 Dec 6) 7396, https://doi.org/10.3390/molecules26237396. PMID: 34885982; PMCID: PMC8659047.
[9] W. Liu, H. Tang, L. Li, X. Wang, Z. Yu, J. Li, Peptide-based therapeutic cancer vaccine: current trends in clinical application, Cell Prolif. 54 (5) (2021 May) e13025, https://doi.org/10.1111/cpr.13025. Epub 2021 Mar 22. PMID: 33754407; PMCID: PMC8088465.
[10] Rodriguez Y. Gomez, B. Oliva Arguelles, M. Riera-Romo, J. Fernandez-De-Cossio, H.E. Garay, J. Fernandez Masso, M. Guerra Vallespi, Synergic effect of anticancer peptide CIGB-552 and Cisplatin in lung cancer models, Mol. Biol. Rep. 49 (4) (2022 Apr) 3197–3212, https://doi.org/10.1007/s11033-022-07152-3. Epub 2022 Jan 30. PMID: 35094208.
[11] S. Basith, B. Manavalan, H.T. Shin, D.Y. Lee, G. Lee, Evolution of machine learning algorithms in the prediction and design of anticancer peptides, Curr. Protein Pept. Sci. 21 (12) (2020) 1242–1250, https://doi.org/10.2174/1389203721666200117171403.
[12] Y. Wan, Z. Wang, T.Y. Lee, Incorporating support vector machine with sequential minimal optimization to identify anticancer peptides, BMC Bioinf. 22 (1) (2021) 286, https://doi.org/10.1186/s12859-021-03965-4.
[13] P. Charoenkwan, W. Chiangjong, V.S. Lee, C. Nantasenamat, M.M. Hasan, W. Shoombuatong, Improved prediction and characterization of anticancer activities of peptides using a novel flexible scoring card method, Sci. Rep. 11 (1) (2021) 3017, https://doi.org/10.1038/s41598-021-82513-9.
[14] T. Zhao, Y. Hu, T. Zang, DRACP: a novel method for identification of anticancer peptides, BMC Bioinf. 21 (16) (2020 Dec) 1, 1.
[15] A. Abbood, A. Ullrich, R. Busche, S. Ghozzi, EventEpi—a natural language processing framework for event-based surveillance, PLoS Comput. Biol. 16 (11) (2020 Nov 20) e1008277.
[16] L. Alzubaidi, J. Zhang, A.J. Humaidi, A. Al-Dujaili, Y. Duan, O. Al-Shamma, J. Santamaría, M.A. Fadhel, M. Al-Amidie, L. Farhan, Review of deep learning: concepts, CNN architectures, challenges, applications, future directions, Journal of big Data 8 (2021 Dec) 1–74.
[17] Reddi SJ, Kale S, Kumar S. On the convergence of Adam and beyond. In: Proceedings of the 6th International Conference on Learning Representations; 2018 Apr 30-May 3; Vancouver, BC, Canada.
[18] J. Su, Z. Liu, J. Zhang, V.S. Sheng, Y. Song, Y. Zhu, Y. Liu, DV-Net: accurate liver vessel segmentation via dense connection model with D-BCE loss function, Knowl. Base Syst. 232 (2021 Nov 28) 107471.
[19] H. Graf, E. Cosatto, L. Bottou, I. Dourdanovic, V. Vapnik, Parallel support vector machines: the cascade SVM, Adv. Neural Inf. Process. Syst. 17 (2004).
[20] J. Telo, Supervised machine learning for detecting malicious URLs: an evaluation of different models, Sage Science Review of Applied Machine Learning 5 (2) (2022 Nov 15) 30–46.
[21] L. Xiong, Y. Yao, Study on an adaptive thermal comfort model with K-nearest-neighbors (KNN) algorithm, Build. Environ. 202 (2021 Sep 1) 108026.
[22] Y.Y. Song, L.U. Ying, Decision tree methods: applications for classification and prediction, Shanghai archives of psychiatry. 27 (2) (2015 Apr 4) 130.
[23] S. Gharpure, R. Yadwade, B. Ankamwar, Non-antimicrobial and non-anticancer properties of ZnO nanoparticles biosynthesized using different plant parts of Bixa orellana, ACS Omega 7 (2) (2022 Jan 5) 1914–1933.
[24] G. Wang, X. Li, Z. Wang, APD3: the antimicrobial peptide database as a tool for research and education, Nucleic Acids Res. 44 (D1) (2016 Jan 4) D1087–D1093.
[25] D. Kumar, P. Kumar, K.N. Rai, A study on DPL model of heat transfer in bi-layer tissues during MFH treatment, Comput. Biol. Med. 75 (2016 Aug 1) 160–172.
[26] D. Das, M. Jaiswal, F.N. Khan, S. Ahamad, S. Kumar, PlantPepDB: a manually curated plant peptide database, Sci. Rep. 10 (1) (2020 Feb 10) 2194.
[27] Q. Li, C. Zhang, H. Chen, J. Xue, X. Guo, M. Liang, M. Chen, BioPepDB: an integrated data platform for food-derived bioactive peptides, Int. J. Food Sci. Nutr. 69 (8) (2018 Nov 17) 963–968.
[28] T. Panyayai, C. Ngamphiw, S. Tongsima, W. Mhuantong, W. Limsripraphan, K. Choowongkomon, O. Sawatdichaikul, FeptideDB: a web application for new bioactive peptides from food protein, Heliyon 5 (7) (2019 Jul 1) e02076.
[29] M. Awais, W. Hussain, N. Rasool, Y.D. Khan, iTSP-PseAAC: identifying tumor suppressor proteins by using fully connected neural network and PseAAC, Curr. Bioinf. 16 (5) (2021 Jun 1) 700–709.
[30] H.B. Shen, K.C. Chou, PseAAC: a flexible web server for generating various kinds of protein pseudo amino acid composition, Anal. Biochem. 373 (2) (2008 Feb 15) 386–388.
[31] J. Jia, X. Li, W. Qiu, X. Xiao, K.C. Chou, iPPI-PseAAC (CGR): identify protein-protein interactions by incorporating chaos game representation into PseAAC, J. Theor. Biol. 460 (2019 Jan 7) 195–203.
[32] I. Lee, J. Keum, H. Nam, DeepConv-DTI: prediction of drug-target interactions via deep learning with convolution on protein sequences, PLoS Comput. Biol. 15 (6) (2019 Jun 14) e1007129.
[33] H. Kang, S. Goo, H. Lee, J.W. Chae, H.Y. Yun, S. Jung, Fine-tuning of bert model to accurately predict drug–target interactions, Pharmaceutics 14 (8) (2022 Aug 16) 1710.
[34] J. Lee, W. Yoon, S. Kim, D. Kim, S. Kim, C.H. So, J. Kang, BioBERT: a pre-trained biomedical language representation model for biomedical text mining, Bioinformatics 36 (4) (2020 Feb 15) 1234–1240.
[35] T. ValizadehAslani, Y. Shi, P. Ren, J. Wang, Y. Zhang, M. Hu, L. Zhao, H. Liang, PharmBERT: a domain-specific BERT model for drug labels, Briefings Bioinf. 24 (4) (2023 Jul) bbad226.
[36] M. Müller, M. Salathé, P.E. Kummervold, Covid-twitter-bert: a natural language processing model to analyse covid-19 content on twitter, Frontiers in artificial intelligence 6 (2023 Mar 14) 1023281.
[37] J. Mingyu, Z. Jiawei, W. Ning, AFR-BERT: attention-based mechanism feature relevance fusion multimodal sentiment analysis model, PLoS One 17 (9) (2022 Sep 9) e0273936.
[38] Devlin J, Chang M-W, Lee K, Toutanova K. BERT: Pre-training of deep bidirectional transformers for language understanding. In: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers); 2019 Jun 2-7; Minneapolis, Minnesota. Association for Computational Linguistics; 2019. p. 4171-86.

[39] T. ValizadehAslani, Y. Shi, P. Ren, J. Wang, Y. Zhang, M. Hu, L. Zhao, H. Liang, PharmBERT: a domain-specific BERT model for drug labels, Briefings Bioinf. 24 (4) (2023 Jul 20).

[40] D. Jimenez-Carretero, V. Abrishami, L. Fernandez-de-Manuel, I. Palacios, A. Quilez-Alvarez, A. Diez-Sanchez, M.A. Del Pozo, M.C. Montoya, Tox_ (R) CNN: deep learning-based nuclei profiling tool for drug toxicity screening, PLoS Comput. Biol. 14 (11) (2018 Nov 30) e1006238.

[41] V.N. Vapnik, A.Y. Lerner, Recognition of patterns with help of generalized portraits, Avtomat. i Telemekh. 24 (6) (1963 Jun) 774–780.

[42] K. Bong, J. Kim, Analysis of intrusion detection performance by smoothing factor of Gaussian NB model using modified NSL-KDD dataset, in: 2022 13th International Conference on Information and Communication Technology Convergence (ICTC), IEEE, 2022, pp. 1471–1476.

[43] G. Guo, H. Wang, D. Bell, Y. Bi, K. Greer, KNN model-based approach in classification, in: On the Move to Meaningful Internet Systems 2003: CoopIS, DOA, and ODBASE: OTM Confederated International Conferences, CoopIS, DOA, and ODBASE 2003, Catania, Sicily, Italy, November 3-7, 2003. Proceedings, Springer Berlin Heidelberg, 2003, pp. 986–996.

[44] A.J. Myles, R.N. Feudale, Y. Liu, N.A. Woody, S.D. Brown, An introduction to decision tree modeling, J. Chemometr.: A Journal of the Chemometrics Society 18 (6) (2004 Jun) 275–285.

[45] C. Halimu, A. Kasem, S.S. Newaz, Empirical comparison of area under ROC curve (AUC) and Mathew correlation coefficient (MCC) for evaluating machine learning algorithms on imbalanced datasets for binary classification, in: Proceedings of the 3rd International Conference on Machine Learning and Soft Computing, 2019 Jan 25, pp. 1–6.

[46] J.M. Lobo, A. Jiménez-Valverde, R. Real, AUC: a misleading measure of the performance of predictive distribution models, Global Ecol. Biogeogr. 17 (2) (2008 Mar) 145–151.

[47] Jurman G, Riccadonna S, Furlanello C. A Comparison of MCC and CEN Error Measures in Multi-Class Prediction.

[48] T Wolf, L Debut, V Sanh, J Chaumond, C Delangue, A Moi, P Cistac, T Rault, R Louf, M Funtowicz, J Davison, Transformers: State-of-the-art natural language processing, InProceedings of the 2020 conference on empirical methods in natural language processing: system demonstrations, 2020 Oct, pp. 38–45.

[49] A. Vaswani, P. Ramachandran, A. Srinivas, N. Parmar, B. Hechtman, J. Shlens, Scaling local self-attention for parameter efficient visual backbones, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 12894–12904.

[50] A Safaya, M Abdullatif, D Yuret, BERT-CNN for offensive speech identification in social media. InProceedings of the Fourteenth Workshop on Semantic Evaluation, KUISAIL at SemEval-2020, Task 12 (2020) 2054–2059.

[51] D. Chicco, N. Tötsch, G. Jurman, The Matthews correlation coefficient (MCC) is more reliable than balanced accuracy, bookmaker informedness, and markedness in two-class confusion matrix evaluation, BioData Min. 14 (2021) 13.

[52] K. Han, et al., A survey on vision transformer, IEEE Trans. Pattern Anal. Mach. Intell. 45 (1) (2023) 87–110, https://doi.org/10.1109/TPAMI.2022.3152247.

[53] B. Barz, J. Denzler, Deep learning on small datasets without pre-training using cosine loss, in: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, 2020, pp. 1371–1380.

[54] M.W. Browne, Cross-validation methods, J. Math. Psychol. 44 (1) (2000 Mar 1) 108–132.

[55] H.R. Ali, O.M. Rueda, S.F. Chin, C. Curtis, M.J. Dunning, S.A. Aparicio, C. Caldas, Genome-driven integrated classification of breast cancer validated in over 7,500 samples, Genome Biol. 15 (2014 Aug) 1–4.

[56] D. Jimenez-Carretero, V. Abrishami, L. Fernandez-de-Manuel, I. Palacios, A. Quilez-Alvarez, A. Diez-Sanchez, M.A. Del Pozo, M.C. Montoya, Tox_ (R) CNN: deep learning-based nuclei profiling tool for drug toxicity screening, PLoS Comput. Biol. 14 (11) (2018 Nov 30) e1006238.

[57] S. Alaparthi, M. Mishra, BERT: a sentiment analysis odyssey, J Market Anal 9 (2021) 118–126.

[58] I. Tobore, J. Li, L. Yuhang, Y. Al-Handarish, A. Kandwal, Z. Nie, L. Wang, Deep learning intervention for health care challenges: some biomedical domain considerations, JMIR mHealth and uHealth 7 (8) (2019 Aug 2) e11966.

[59] B.J. Marafino, M. Park, J.M. Davies, R. Thombley, H.S. Luft, D.C. Sing, D.S. Kazi, C. DeJong, W.J. Boscardin, M.L. Dean, R.A. Dudley, Validation of prediction models for critical care outcomes using natural language processing of electronic health record data, JAMA Netw. Open 1 (8) (2018 Dec 7) e185097.

[60] K. Kardani, A. Bolhassani, Antimicrobial/anticancer peptides: bioactive molecules and therapeutic agents, Immunotherapy 13 (8) (2021 Jun) 669–684.

[61] H.P. Chan, R.K. Samala, L.M. Hadjiiski, C. Zhou, Deep learning in medical image analysis, Adv. Exp. Med. Biol. 1213 (2020) 3–21, https://doi.org/10.1007/978-3-030-33128-3_1. PMID: 32030660; PMCID: PMC7442218.

[62] C. Liu, S. Cheng, C. Chen, et al., M-FLAG: medical vision-language pre-training with frozen language models and latent space geometry optimization[C], in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer Nature Switzerland, Cham, 2023, pp. 637–647.

[63] Radford A, Kim JW, Hallacy C, Ramesh A, Goh G, Agarwal S, Sastry G, Askell A, Mishkin P, Clark J, Krueger G. Learning transferable visual models from natural language supervision. InInternational conference on machine learning 2021 Jul 1 (pp. 8748-8763). PMLR.Radford A., Kim J.W., Hallacy C., et al., Clip: learning transferable visual models from natural language supervision[J], arXiv preprint arXiv:2103.00020 (2021).

[64] Z. Wan, C. Liu, M. Zhang, et al., Med-unic: unifying cross-lingual medical vision-language pre-training by diminishing bias, Adv. Neural Inf. Process. Syst. (2024) 36.

[65] Y. Jiang, M. Yang, S. Wang, X. Li, Y. Sun, Emerging role of deep learning-based artificial intelligence in tumor pathology, Cancer Commun. 40 (4) (2020 Apr) 154–166, https://doi.org/10.1002/cac2.12012. Epub 2020 Apr 11. PMID: 32277744; PMCID: PMC7170661.

[66] C.M. Li, P. Haratipour, R.G. Lingeman, J.J.P. Perry, L. Gu, R.J. Hickey, L.H. Malkas, Novel peptide therapeutic approaches for cancer treatment, Cells 10 (11) (2021 Oct 27) 2908, https://doi.org/10.3390/cells10112908. PMID: 34831131; PMCID: PMC8616177.

[67] A.L. Tornesello, A. Borrelli, L. Buonaguro, F.M. Buonaguro, M.L. Tornesello, Antimicrobial peptides as anticancer agents: functional properties and biological activities, Molecules 25 (12) (2020 Jun 19) 2850.

[68] S. Basith, B. Manavalan, T. Hwan Shin, G. Lee, Machine intelligence in peptide therapeutics: a next-generation tool for rapid disease screening, Med. Res. Rev. 40 (4) (2020 Jul) 1276–1314.

[69] M.R. Kosorok, E.B. Laber, Precision medicine, Annual review of statistics and its application 6 (2019 Mar 7) 263–286.

[70] C. Alix-Panabières, H. Schwarzenbach, K. Pantel, Circulating tumor cells and circulating tumor DNA, Annu. Rev. Med. 63 (2012 Feb 18) 199–215.

[71] J. Yang, Z. Xu, W.K. Wu, Q. Chu, Q. Zhang, GraphSynergy: a network-inspired deep learning model for anticancer drug combination prediction, J. Am. Med. Inf. Assoc. 28 (11) (2021 Nov 1) 2336–2345.

[72] A. Ghulam, F. Ali, R. Sikander, A. Ahmad, A. Ahmed, S. Patil, ACP-2DCNN: deep learning-based model for improving prediction of anticancer peptides using two-dimensional convolutional neural network, Chemometr. Intell. Lab. Syst. 226 (2022 Jul 15) 104589.

[73] Y. Zhang, Z. Dai, X. Zhao, C. Chen, S. Li, Y. Meng, Z. Suonan, Y. Sun, Q. Shen, L. Wang, Y. Xue, Deep learning drives efficient discovery of novel antihypertensive peptides from soybean protein isolate, Food Chem. 404 (2023 Mar 15) 134690.