

Article

A Stacking Machine Learning Method for IL-10-Induced Peptide Sequence Recognition Based on Unified Deep Representation Learning

Jiayu Li ^{1,2}, Jici Jiang ¹, Hongdi Pei ¹ and Zhibin Lv ^{1,*} 

¹ Student Innovation Competition Team, College of Biomedical Engineering, Sichuan University, Chengdu 610065, China; lijayu927@foxmail.com (J.L.); jiangjici@foxmail.com (J.J.); peihongdi@foxmail.com (H.P.)

² College of Life Science, Sichuan University, Chengdu 610065, China

* Correspondence: lvzhibin@pku.edu.cn

Abstract: Interleukin-10 (IL-10) has anti-inflammatory properties and is a crucial cytokine in regulating immunity. The identification of IL-10 through wet laboratory experiments is costly and time-intensive. Therefore, a new IL-10-induced peptide recognition method, IL10-Stack, was introduced in this research, which was based on unified deep representation learning and a stacking algorithm. Two approaches were employed to extract features from peptide sequences: Amino Acid Index (AAindex) and sequence-based unified representation (UniRep). After feature fusion and optimized feature selection, we selected a 1900-dimensional UniRep feature vector and constructed the IL10-Stack model using stacking. IL10-Stack exhibited excellent performance in IL-10-induced peptide recognition (accuracy (ACC) = 0.910, Matthews correlation coefficient (MCC) = 0.820). Relative to the existing methods, IL-10Pred and ILeukin10Pred, the approach increased in ACC by 12.1% and 2.4%, respectively. The IL10-Stack method can identify IL-10-induced peptides, which aids in the development of immunosuppressive drugs.

Keywords: unified representation learning; interleukin-10; machine learning; bioinformatics



Citation: Li, J.; Jiang, J.; Pei, H.; Lv, Z.

A Stacking Machine Learning Method for IL-10-Induced Peptide Sequence Recognition Based on Unified Deep Representation Learning. *Appl. Sci.* **2023**, *13*, 9346. <https://doi.org/10.3390/app13169346>

Academic Editor: Chilukuri K. Mohan

Received: 21 July 2023

Revised: 14 August 2023

Accepted: 15 August 2023

Published: 17 August 2023



Copyright: © 2023 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

Interleukin-10 (IL-10), a pleiotropic cell-signaling cytokine, contributes to immune modulation and inflammation [1]. Its production occurs in various lymphocyte subtypes, including macrophages, B cells, granulocytes, dendritic cells, and multiple T cell subsets [2,3]. IL-10 was initially discovered by Mossman and Coffman through cloning of T helper (Th) 2 cells, and was shown to inhibit cytokines produced by Th1 cells [4]. As an immunosuppressive molecule, IL-10 can restrict immune responses against pathogens and microbial communities, which is a key mechanism underlying its anti-inflammatory properties. Evidence suggests that IL-10 is expressed not only in bone marrow and lymphoid lineage cells but also in tumor-associated macrophages [5], epithelial cells [6], and innate immune cells of the central nervous system [7]. This broad anti-inflammatory profile has significant effects in preventing autoimmune diseases [8], balancing neuroimmune responses [7], and cancer therapy [9].

Identifying IL-10 target proteins through wet laboratory experiments is time-consuming and costly. Additionally, the complexity of the immune system and the workload involved in prediction are high. With the rapid development of immunoinformatics tools and immune epitope databases, combining machine learning (ML) with massive epitope data to construct direct or indirect peptide prediction models has become an area of increased focus [10–16]. Some methods to predict IL-10-induced epitopes have been developed based on ML and the largest immune epitope database, IEDB (Immune Epitope Database) [17,18]. Nagpal et al. first developed a computational model, IL-10Pred, in 2017 to predict peptides

that can induce IL-10 production. IL-10Pred is a cytokine-specific prediction method [19] that extracts features of IL-10-induced peptides using singular amino acid sequences to construct models using support vector machines (SVM). A model based on random forest developed using DPC achieved the best performance, with accuracy (ACC) = 0.812. Subsequently, in 2021, Singh et al. proposed ILeukin10Pred, another prediction model for IL-10-induced peptides developed based on amino acid sequence features [20]. Independent testing and five-fold cross-validation both indicated that ILeukin10Pred showed improved predictive performance compared to IL-10Pred. Although significant progress has been made in the prediction of IL-10-induced peptides, the performance of ML-based models utilizing sequence information still requires further improvement. Many research methods have made it feasible. In other fields, such as image processing, certain algorithmic models can offer assistance for biomedical issues. For instance, generative adversarial networks [21], deformation models [22], gated recurrent units [23], and dual-level representation enhancement networks [24] exhibit strengths in data handling, feature representation, and model optimization. As considered by ILeukin10Pred, it is essential to extract and encode features based on a wider range of amino acid characteristics for accurately predicting amino acids. However, the key to enhancing model performance lies in effectively fusing different types of data [25,26]. Additionally, Singh et al. mentioned in their study that using ensemble models may improve predictive performance [20]. To address these issues, this research proposes a novel IL-10-induced peptide recognition method, IL10-Stack, based on unified deep representation learning and the stacking algorithm. Unified deep representation learning can harmonize diverse data types, facilitating comprehensive analysis after data fusion [27]. Meanwhile, stacking, through integrating different models, demonstrates outstanding performance in improving prediction accuracy and generalization [28,29].

The Amino Acid Index (AAindex) is a numerical index database that provides information on the physicochemical and biological properties of the 20 amino acids [30]. It provides three categories of lists for delineating amino acid properties. These lists delve into the biological and chemical attributes of single or paired amino acids, encompassing aspects like charge, polarity, mutability, and contact potential. Indexing amino acid features based on the amino acid index list has become a common method in bioinformatics [30]. Unified representation (UniRep) is a method that transforms any protein sequence into a fixed-length vector representation, addressing the scarcity of protein informatics data by leveraging full utilization of the original sequence [31,32]. A key feature of UniRep is the numerical encoding of oligonucleotides, allowing for comparison and analysis of all oligonucleotide pairs occurring in the sequence (including overlaps). UniRep can learn the buried features of amino acids and incorporate the physicochemical properties of amino acid residue clusters in the protein. Using UniRep to predict amino acid features has significantly improved the performance of models and the method has been widely applied in protein engineering informatics.

In this study, we developed the IL10-Stack model based on stacking using UniRep feature encoding. A non-stacking IL10-Fuse model was also constructed for comparison. Both models were used for IL-10-induced peptide identification. We also compared the performance of AAindex and UniRep single and fusion feature extraction methods using different ML algorithms, and found that the UniRep single-feature stacking model had the highest accuracy. Notably, IL10-Stack outperformed existing prediction techniques in terms of five-fold cross-validation (MCC = 0.796 and ACC = 0.897) and independent testing (MCC = 0.820 and ACC = 0.910). Relative to IL-10Pred and ILeukin10Pred, the independent testing accuracy of IL10-Stack was improved by 12.1% and 4.0%, respectively. Overall, the IL10-Stack model developed in this study demonstrated higher accuracy and exhibited good robustness and generalization performance.

Our main contributions are summarized as follows:

- To address IL-10-induced peptide recognition from sequences, we transformed arbitrary protein sequences into fixed-length vector representations using sequence-based unified representation (UniRep).
- We employed the powerful ensemble learning algorithm, stacking, to construct an IL-10-induced peptide prediction model, effectively enhancing the predictive accuracy.
- After modeling single or fused sequence features using various machine learning algorithms, we observed that the stacking model based on UniRep encoding yielded the best results. Therefore, we proposed a novel IL-10-induced peptide recognition method, IL10-Stack, with significantly superior performance compared to existing methods.

2. Materials and Methods

2.1. Computational Framework

Figure 1 depicts the computational workflow framework used to build the IL-10-induced peptide prediction model. The analysis workflow involved multiple phases, such as data retrieval from the IL10Pred server [33], feature extraction, handling of imbalanced data, feature fusion, feature selection, utilization of ML algorithms, model assessment, and IL10-Stack server construction.

Initially, an IL-10 dataset was obtained from the IL10Pred method. Next, protein sequence features were extracted using AAIndex and UniRep. Additionally, we adopted the SMOTE method to rectify the data imbalance in the dataset. After feature fusion, the process of feature selection was performed utilizing LGBM.

We proceeded to split the data into training and testing sets, maintaining a ratio of 4:1. The methodology employed for training and validating the model incorporated both 5-fold cross-validation and independent testing. Two new methods for IL-10-induced peptide prediction were proposed, namely the non-stacking model, IL10-Fuse, and the stacking model, IL10-Stack. In the construction of IL10-Stack, we first trained individual models using SVM and LGBM separately. Subsequently, we made predictions on the test and validation sets using these two models, obtaining the test set predictions and validation set predictions. Next, we used the predictions from both models to create new test and validation sets, and then trained a new SVM model. Through this process, we developed a stacking algorithm based on amino acid sequence features for predicting IL-10-induced peptides. To facilitate better understanding and usage of our new algorithm by other scientists, we also established a web server in which users simply need to input a peptide sequence for prediction, and they will receive results indicating whether the peptide is an IL-10-induced peptide, along with corresponding confidence levels.

2.2. Dataset Acquisition and Preprocessing

IEDB [17] serves as a repository of immunological epitope-related details, presenting extensive antibody and T-cell epitope data for disease investigations. In our study, we utilized the foundational database mentioned in the IL-10Pred publication to formulate the IL-10-induced peptide prediction model [19]. In building the dataset, we designated MHC class II binders as IL-10-induced peptides based on the results of whether they can induce the release of the cytokine IL-10, and vice versa. After removing duplicate peptides, the dataset ultimately employed comprised 848 non-IL-10-induced peptides and 394 IL-10-induced peptides. We utilized the Synthetic Minority Over-sampling Technique, commonly known as SMOTE, to balance the positive and negative instances [34].

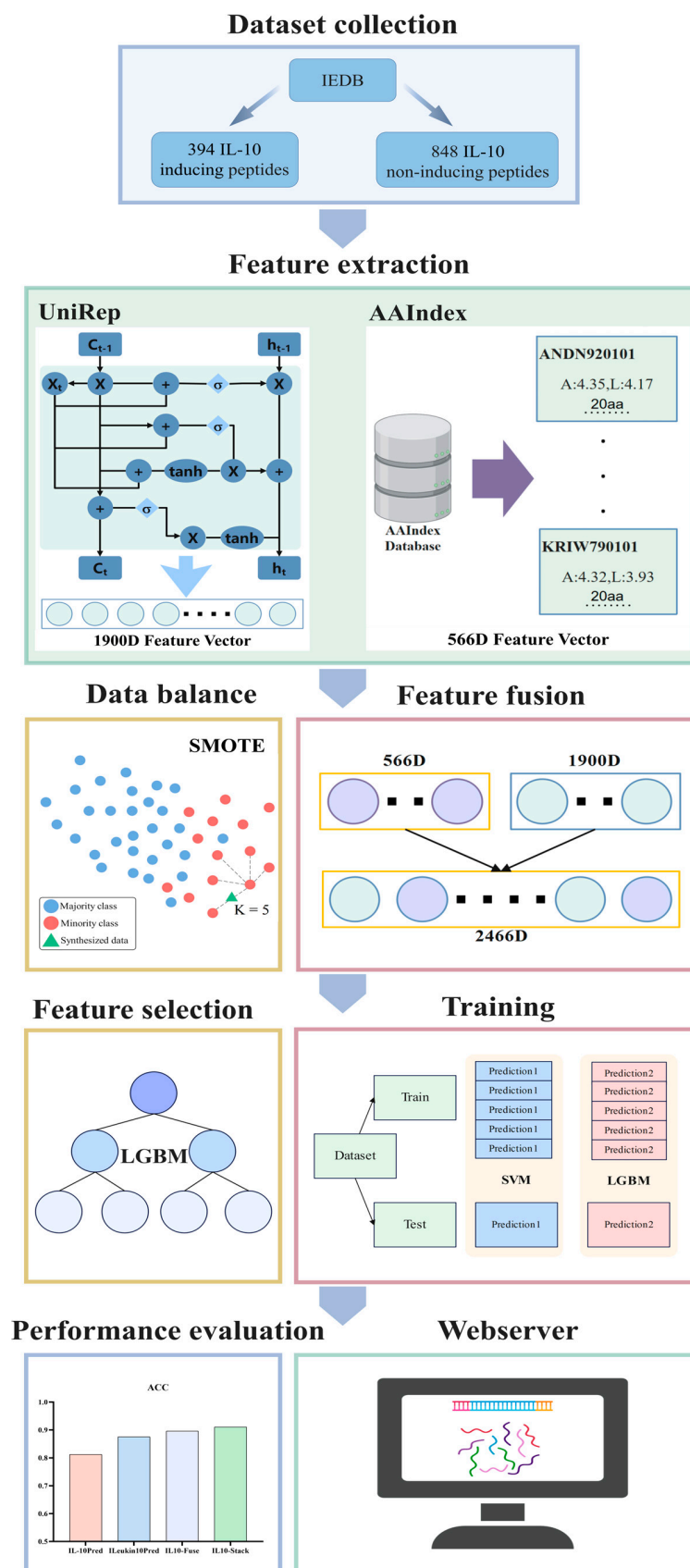


Figure 1. Overview of model development in this study, including dataset collection, feature extraction, data balancing, feature fusion, feature selection, model training, evaluation process, and web server.

2.3. Feature Encoding

Amino acid sequence encoding is the process of converting an amino acid sequence into a numerical or discrete encoded representation for ML analysis and processing [35]; it is the first step in ML prediction. To examine how different features affect the recognition of IL-10-induced peptides, two feature representation methods, AAIndex and UniRep, were employed; single and fused feature encoding using these two methods were compared, to construct a broader and more complete predictive model.

2.3.1. AAIndex Embedding Model

The AAIndex database [36] consists of three sections: AAIndex1, AAIndex2, and AAIndex3. Since the latter two lists involve the relationship between two proteins and did not apply to this study, 566 amino acid indices were selected from AAIndex1, where each index contained 20 amino acid values.

2.3.2. Pre-Trained UniRep Embedding Model

UniRef50 contains 240,000 amino acid sequences, which UniRep utilizes for training. UniRep learns how to represent proteins by minimizing cross-entropy loss in predicting the next amino acid [37]. After training, the model can encode input sequences into a single fixed-length vector using an mLSTM encoder [38]. The best-performing machine learning model is obtained by training to predict output vectors. Supervised learning in various bioinformatics assignments can be achieved by utilizing the input sequences as features.

First, encode sequences featuring L amino acid residues using one-hot encoding and then embed the outcomes into an $R^{L \times 10}$ matrix. Then, the matrix was passed through the mLSTM encoder to generate the hidden state outputs of size $R^{1900 \times L}$, serving as the embedding matrix. Ultimately, the 1900-dimensional (1900D) UniRep feature vector was derived through the process of average pooling.

The mLSTM encoder computation is expressed by the following Formulas (1)–(7):

$$m_t = (X_t W_{xm}) \otimes (h_{t-1} W_{hm}) \quad (1)$$

$$\hat{h}_t = \tan h(X_t W_{xh} + m_t W_{mh}) \quad (2)$$

$$f_t = \sigma(X_t W_{xf} + m_t W_{mf}) \quad (3)$$

$$i_t = \sigma(X_t W_{xi} + m_t W_{mi}) \quad (4)$$

$$o_t = \sigma(X_t W_{xo} + m_t W_{mo}) \quad (5)$$

$$C_t = i_t \otimes \hat{h}_t + C_{t-1} \otimes f_t \quad (6)$$

$$h_t = o_t \otimes \tan h(C_t) \quad (7)$$

where \otimes represents element-wise multiplication; m_t , the current modulation state; X_t , the current input; h_{t-1} , the previous hidden state; i_t , the input gate; \hat{h}_t , the input before the hidden state; f_t , the forgotten gate; C_t , the current cell state; C_{t-1} , the previous cell state; h_t , the output hidden state; and o_t , the output gate. Both σ and $\tan h$ are functions representing the sigmoid and tangent functions, respectively.

2.3.3. Feature Fusion

In pursuit of enhancing the model's predictive capabilities and robustness, the fusion of the 566D AAIndex feature vector with the 1900D UniRep feature vector results in the

2466D AAIndex + UniRep fused feature vector, referred to as AAIndex + UniRep fusion eigenvector.

2.4. Feature Selection Method

The process of feature selection entails picking the most relevant subset from the pool of original features [39], reducing dimensionality, eliminating redundant features, and improving the performance and generalization ability of a model [40]. In this study, light gradient-boosting machine (LGBM) was employed to optimize the model. LGBM was used to identify the best features and rank them based on their importance values, ultimately selecting the features with importance values greater than the threshold, the ‘average feature importance value’ [41].

2.5. Balancing Strategy

To address the issue of the imbalanced dataset, which could lead to bias towards the majority class (non-IL-10-induced peptides) [33], we employed various methods to balance the data. Among various methods of over-sampling and under-sampling, the SMOTE [42] demonstrated the most effective results (please see Supplementary Table S1). SMOTE is a method for constructing classifiers from imbalanced datasets and is superior to random oversampling algorithms, as it can effectively improve classification performance; it generates synthetic minority samples by randomly interpolating between minority sample points and their neighboring points, which are identified using the k-nearest neighbors algorithm within minority class (IL-10-induced peptides) samples. The distance between sample points is calculated using the Euclidean distance. This data balancing strategy not only addresses the issue of class imbalance and improves model performance but also avoids information loss or sample duplication, thereby enhancing the model’s generalization ability.

2.6. ML Algorithms

This study utilized four widely used high-performance ML models, including logistic regression (LR), SVM, LGBM, and stacking.

LR [43] is a ML algorithm suitable for classification problems. It constructs a linear model, estimates parameters using maximum likelihood estimation, and utilizes a logistic function for classification. LR is widely used due to its simple data preparation, ease of implementation, efficiency, and strong interpretability. The parameters of the best model that we used were as follows: ‘C’: 0.1097, ‘penalty’: ‘l2’.

SVM [44–46] finds the optimal hyperplane by mapping samples into a high-dimensional feature space; it is a powerful classification algorithm. The parameters of the best model that we used were as follows: ‘C’: 9.2367, ‘gamma’: 0.0007, ‘kernel’: ‘rbf’.

LGBM [47] is a gradient-boosting framework based on decision tree algorithms; it offers lower memory usage and faster training speed. The parameters of the best model that we used were as follows: ‘max depth’: 9, ‘n_estimators’: 750, ‘learning rate’: 0.05.

Stacking [28] is an ensemble learning algorithm that constructs a secondary learner (also known as a meta-learner) by using the predictions of multiple base learners (also known as primary learners) as inputs for the final prediction. Through multiple iterations, stacking achieves good performance and generalization ability. The parameters of the best model that we used were as follows: ‘learning rate’: 0.05, ‘n estimators’: 800, ‘n jobs’: 60, ‘meta classifier C’: 0.1, ‘meta classifier gamma’: 0.1, ‘svc C’: 10, ‘svc gamma’: 0.001, ‘max depth’: 6.

2.7. Performance Evaluation

Several statistical metrics were employed to appraise the model’s effectiveness [48], including area under the receiver operating characteristic curve (AUC), ACC, MCC, Sn, specificity (Sp), true positive (TP), false positive (FP), true negative (TN), false negative (FN), and

precision (P). The computation methods for these metrics are as follows (Equations (8)–(12)):

$$ACC = \frac{TP + TN}{FP + FN + TP + TN} \quad (8)$$

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \quad (9)$$

$$Sn = \frac{TP}{TP + FN} \quad (10)$$

$$Sp = \frac{TN}{TN + FP} \quad (11)$$

$$P = \frac{TP}{TP + FP} \quad (12)$$

where *TP* and *TN* correspond to the count of accurate IL-10-induced peptide identifications and non-IL-10-induced peptide identifications, respectively; *FP* and *FN* correspond to the count of erroneous IL-10-induced peptide identifications and non-IL-10-induced peptide identifications, respectively. Models were compared using AUC values.

The practice of employing K-fold cross-validation and carrying out independent testing is a customary way to measure the performance of ML models. Split the raw data into K equally sized subsets, selecting one of them as the validation set and using the others as the training set. Execute this operation K times in succession, each time using different training and validation sets, and the results are evaluated and averaged. In this study, we employed 5-fold ($K = 5$) cross-validation. In independent testing, a distinct dataset is utilized, which is entirely separate from the training set, to assess how well the model performs and generalizes in real-world scenarios.

3. Results and Discussion

3.1. Analysis of Three Different Feature Models Based on Non-Stacking Algorithms

To explore the biological features that can be used to identify IL-10-induced peptides, we first applied two deep representation learning methods for feature extraction, namely AAindex and UniRep. For each type of feature, we employed three different ML methods (LR, SVM, and LGBM) to develop models and perform initial optimization. Next, we combined these two features to generate fused features, namely AAindex + UniRep, and the fused feature combination was used as input for LR, SVM, and LGBM to construct prediction models and refine their performance. The model evaluation results of AAindex and UniRep in the non-stacking LR, SVM, and LGBM models are visualized in the form of Figure 2. Overall, the LR model performed worst, while the LGBM model had moderate performance, and the SVM model performed well. A comparison of the single features of AAindex (red) and UniRep (blue) with the fused feature (green) demonstrated that the SVM model using the fused feature exhibited the best performance in the non-stacking models; it had 2466 feature dimensions, with an ACC of 0.896 (Figure 2(B₁)) and an MCC of 0.792 (Figure 2(B₂)) in independent testing. Although independent testing showed that AAindex + UniRep had the highest ACC in the SVM algorithm, the UniRep feature demonstrated the highest ACC in LR and LGBM (Figure 2(B₁)). Further, five-fold cross-validation results indicated that the UniRep feature achieved the highest ACC in LR and SVM (Figure 2(A₁)). Hence, direct fusing of the two features can improve the performance of certain models, but it is not always effective.

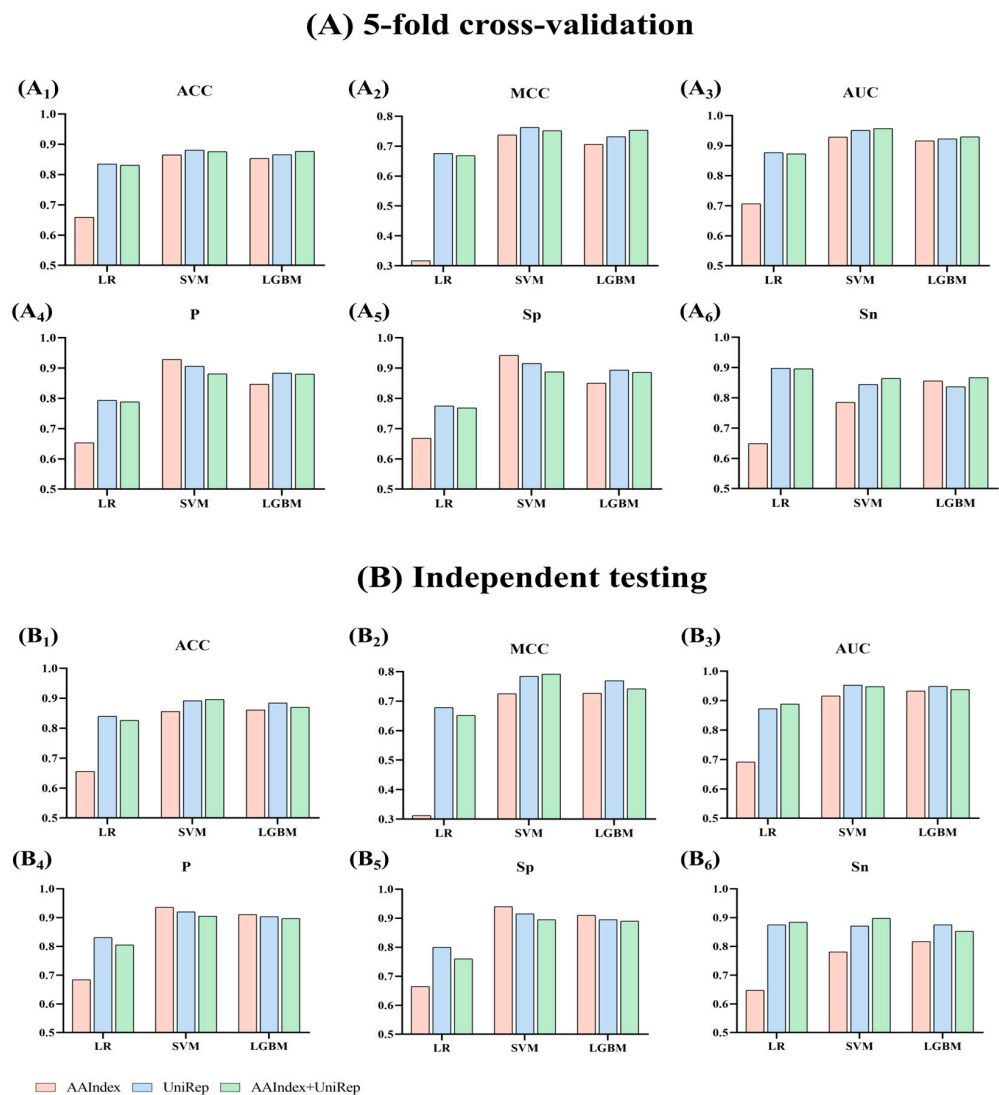


Figure 2. Performance metrics using different features and a range of algorithms. (A) The results of 5-fold cross-validation; (B) the results of independent testing. Different colors represent different features used by each algorithm, with red indicating AAIndex single feature, blue indicating the UniRep single feature, and green indicating AAIndex + UniRep fused feature. (A₁,B₁) ACC, accuracy; (A₂,B₂) MCC, Matthews correlation coefficient; (A₃,B₃) AUC, the area under the receiver operating characteristic curve; (A₄,B₄) P, precision; (A₅,B₅) Sp, specificity; (A₆,B₆) Sn, recall.

3.2. Analysis of Three Different Feature Models Based on Stacking

During the exploratory phase of model development work, we used paired feature combinations to generate fused features and estimated models using three ML methods [49]. The results showed that fused feature encoding outperformed non-fused feature encoding in certain models. The single features of AAIndex and UniRep had dimensions of 566 and 1900, respectively, while the dimension of the fused feature reached 2466. The risks of model overfitting and redundancy in information became more prominent as the dimensionality of features increased. To address this issue, we employed the stacking algorithm to build models using the three feature input methods separately. Figure 3 presents a comparison of the performance of models built on stacking using different dimensions of single and fused features.

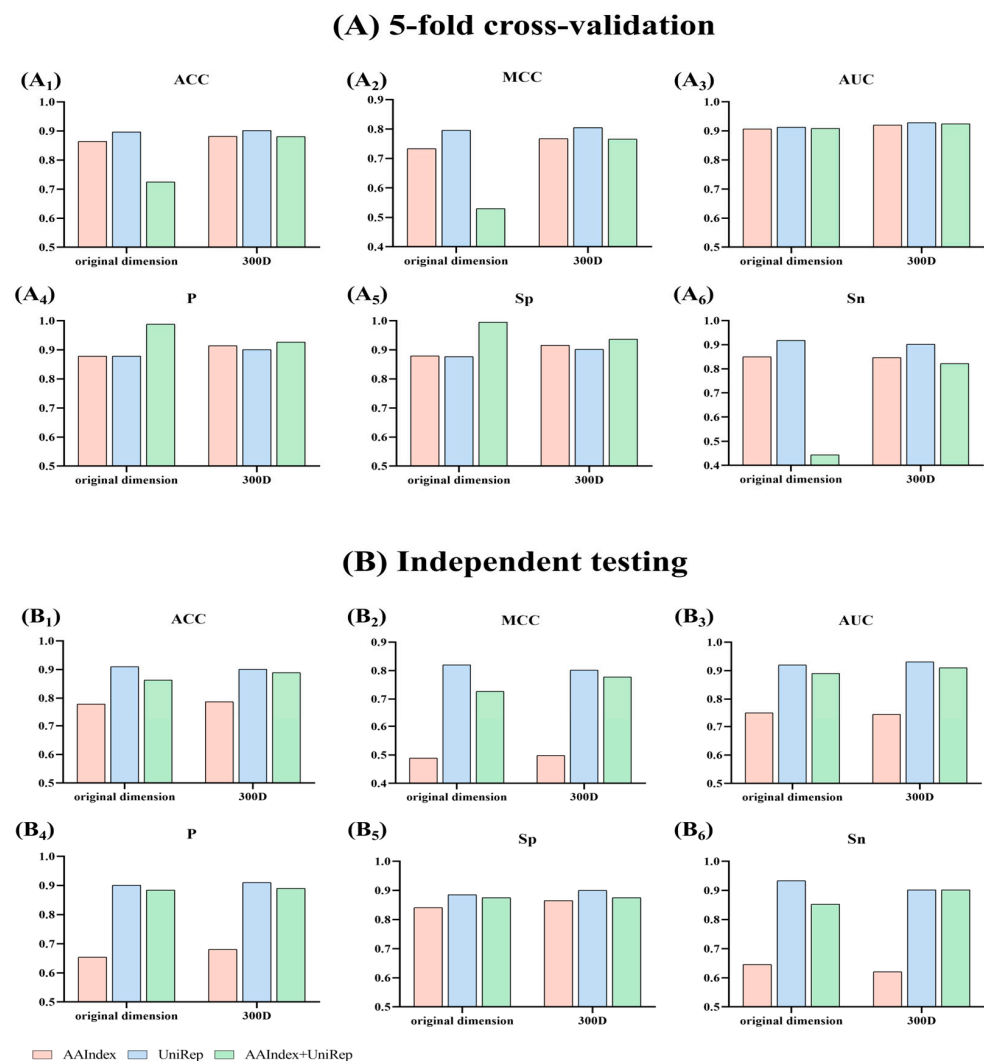


Figure 3. Performance metrics of single or fused feature models using the stacking algorithm. (A) The results of 5-fold cross-validation; (B) the results of independent testing. Different colors represent different features used by each algorithm, with red indicating AAindex single feature, blue indicating the UniRep single feature, and green indicating AAindex + UniRep fused feature. (A₁,B₁) ACC, accuracy; (A₂,B₂) MCC, Matthews correlation coefficient; (A₃,B₃) AUC, the area under the receiver operating characteristic curve; (A₄,B₄) P, precision; (A₅,B₅) Sp, specificity; (A₆,B₆) Sn, recall.

From the comparison presented in Figure 3, it can be observed that, for models using the fused feature, AAindex + UniRep (green), stacking models with feature selection exhibited better performance (Figure 3(A₁)), 21.5% improvement in ACC in five-fold cross-validation (Figure 3(A₂)), 44.6% improvement in MCC in five-fold cross-validation; however, for models using single features, there was no significant effect, and the performance may even have been worse (Figure 3(A₆,B₆)). Notably, in contrast to the higher performance of the fused feature models mentioned in the previous section, stacking models using UniRep single feature exhibited the best overall performance (Figure 3(A₁–A₆,B₁–B₆)). In the end, we chose the single UniRep feature to construct a stacking-based IL-10-induced peptide recognition model. Although the stacking model with a feature-selected 300D UniRep single feature showed better AUC (1.1% higher), P (1.0% higher), and Sp (1.7% higher) values in independent testing (Figure 3(B₃–B₅)), we still considered the stacking model based on the original dimension of UniRep single feature as the one that provided the best overall performance.

3.3. Comparison with Existing Methods

To evaluate the effectiveness of our technique, we compared IL10-Stack with the non-stacking method, IL10-Fuse, and the existing methods, IL-10Pred and ILeukin10Pred (Table 1). As shown in Table 1, the IL10-Stack model based on stacking achieved the best performance in independent testing, with ACC = 0.910 and MCC = 0.820, outperforming the IL10-Fuse model (ACC = 0.896 and MCC = 0.792).

Table 1. Comparison of independent testing results between IL10-Stack and other existing methods.

Classifier	ACC	MCC	AUC	P	Sp	Sn
IL-10Pred	0.812	0.590	0.880	0.674	0.819	0.797
ILeukin10Pred	0.875	0.755	0.931	0.927^a	0.947	0.804
IL10-Fuse	0.896	0.792	0.948	0.905	0.895	0.897
IL10-Stack	0.910	0.820	0.920	0.901	0.885	0.933

^a The best performance values are indicated in bold and underlined.

As shown in Figure 4, we employed multiple metrics to examine model performance (Figure 4A–F). Relative to the existing method, IL-10Pred, the ACC of our model was improved by 12.1% (Figure 4A), MCC improved by 39% (Figure 4B), AUC improved by 4.5% (Figure 4C), P improved by 33.7% (Figure 4D), Sp improved by 8.1% (Figure 4E), and Sn improved by 17.1% (Figure 4F). Moreover, relative to the existing method, ILeukin10Pred, our model showed improvements of 2.4% in ACC (Figure 4A), 4.9% in MCC (Figure 4B), and 11.6% in Sn (Figure 4F). Our results demonstrate that IL10-Stack is one of the cutting-edge IL-10-induced peptide prediction techniques based on unified deep representation learning. Compared with existing non-stacking methods, IL10-Stack provides a more reliable and stable prediction of IL-10-induced peptides.

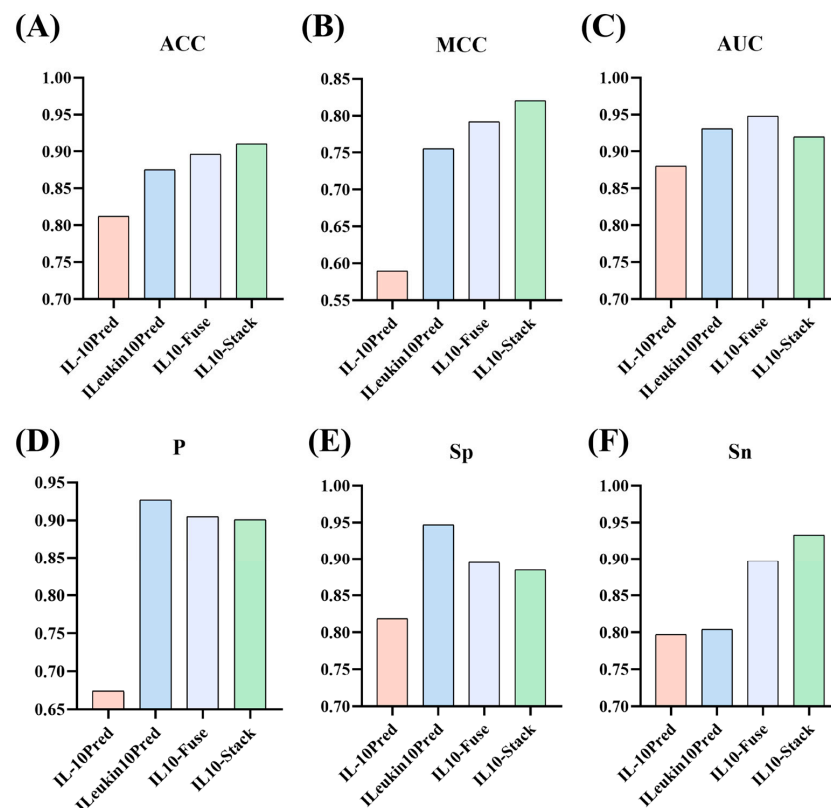


Figure 4. Comparison of independent testing performance between the IL-10Pred, ILeukin10Pred, IL10-Fuse, and IL10-Stack methods. (A) ACC, accuracy; (B) MCC, Matthews correlation coefficient; (C) AUC, the area under the receiver operating characteristic curve; (D) P, precision; (E) Sp, specificity; (F) Sn, recall.

3.4. Web Server Development

To enable more researchers to use our IL-10-induced peptide prediction algorithm, we developed a user-friendly IL10-Stack web server that will be available for free online at <https://servers.aibiochem.net/soft/IL10-Stack/> (accessed on 19 July 2023). The server is designed to be easy to use; users simply need to input a peptide sequence, and subsequently await the results [50,51]. The webpage will display whether the peptide sequence is predicted to be an IL-10-induced peptide, along with a corresponding confidence level. The output includes the input sequence, the prediction result, and the confidence level. Please refer to the web server interface on our website for more information.

4. Conclusions

Here, we developed IL10-Stack, a stacking ML approach based on unified deep representation learning for IL-10-induced peptide recognition. We first employed SMOTE to address the imbalanced dataset and then extracted potential IL-10-induced peptide information using AAIndex, UniRep, and a fusion of both feature extraction methods. We then compared the predictive performance of three non-stacking ML algorithms (LR, SVM, and LGBM) and stacking algorithms, with and without feature selection, resulting in two optimal models: IL10-Fuse and IL10-Stack. After testing and optimization, we found that utilizing 1900D UniRep features as the feature set and developing the model based on the stacking algorithm yielded the best performance. Our results demonstrate that IL10-Stack provides more reliable and accurate predictions of IL-10-induced peptides than the most recently reported methods (AUC = 0.920, ACC = 0.910, MCC = 0.820). Ultimately, we established an IL10-Stack web server to make it convenient for scientists to employ this algorithm.

One advantage of this study is that we used the latest AAIndex and UniRep methods to extract features, which is an improvement over existing methods like ILeukin10Pred. Additionally, our model construction using unified deep representation learning and the stacking algorithm significantly enhanced the accuracy of IL-10-induced peptide prediction. However, our research also had some limitations, such as a relatively small dataset and the slower running speed of the web server. In the future, we plan to validate the model with more data. We envisage that the use of IL10-Stack to predict IL-10-induced peptides can contribute to the development of immunosuppressive drugs.

Supplementary Materials: The following supporting information can be downloaded at: <https://www.mdpi.com/article/10.3390/app13169346/s1>, Table S1: Comparison of independent testing results between IL10-Stack and other existing methods.

Author Contributions: Conceptualization, Z.L.; Data curation, J.L. and Z.L.; Formal analysis, J.L. and Z.L.; Funding acquisition, Z.L.; Writing—review and editing, J.L., H.P., J.J. and Z.L. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the National Natural Science Foundation of China (No. 62001090), and Fundamental Research Funds for the Central Universities of Sichuan University (No. YJ2021104, No. 20826041G4189).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data used to support the findings of this study can be made available by the corresponding author upon request.

Conflicts of Interest: The authors declare no conflict of interest.

References

- Hervas-Salcedo, R.; Fernandez-Garcia, M.; Hernando-Rodriguez, M.; Suarez-Cabrera, C.; Bueren, J.A.; Yanez, R.M. Improved efficacy of mesenchymal stromal cells stably expressing CXCR4 and IL-10 in a xenogeneic graft versus host disease mouse model. *Front. Immunol.* **2023**, *14*, 1062086.
- Maynard, C.L.; Weaver, C.T. Diversity in the contribution of interleukin-10 to T-cell-mediated immune regulation. *Immunol. Rev.* **2008**, *226*, 219–233. [[PubMed](#)]
- Mannino, M.H.; Zhu, Z.W.; Xiao, H.P.; Bai, Q.; Wakefield, M.R.; Fang, Y.J. The paradoxical role of IL-10 in immunity and cancer. *Cancer Lett.* **2015**, *367*, 103–107. [[PubMed](#)]
- Fiorentino, D.F.; Bond, M.W.; Mosmann, T.R. Two types of mouse T helper cell, I.V. Th2 clones secrete a factor that inhibits cytokine production by Th1 clones. *J. Exp. Med.* **1989**, *170*, 2081–2095.
- Tanaka, Y.; Nakai, T.; Suzuki, A.; Kagawa, Y.; Noritake, O.; Taki, T.; Hashimoto, H.; Sakai, T.; Shibata, Y.; Izumi, H.; et al. Clinicopathological significance of peritumoral alveolar macrophages in patients with resected early-stage lung squamous cell carcinoma. *Cancer Immunol. Immunother.* **2023**, *72*, 2205–2215.
- Zuurveld, M.; Ayechu-Muruzabal, V.; Folkerts, G.; Garssen, J.; Van't Land, B.; Willemsen, L.E.M. Specific Human Milk Oligosaccharides Differentially Promote Th1 and Regulatory Responses in a CpG-Activated Epithelial/Immune Cell Coculture. *Biomolecules* **2023**, *13*, 263.
- Geladaris, A.; Hausser-Kinzel, S.; Pretzsch, R.; Nissimov, N.; Lehmann-Horn, K.; Hausler, D.; Weber, M.S. IL-10-providing B cells govern pro-inflammatory activity of macrophages and microglia in CNS autoimmunity. *Acta Neuropathol.* **2023**, *145*, 461–477.
- Riquelme-Neira, R.; Walker-Vergara, R.; Fernandez-Blanco, J.A.; Vergara, P. IL-10 Modulates the Expression and Activation of Pattern Recognition Receptors in Mast Cells. *Int. J. Mol. Sci.* **2023**, *24*, 9875.
- Ahmed, A.; Kohler, S.; Klotz, R.; Giese, N.; Hackert, T.; Springfield, C.; Jager, D.; Halama, N. Sex Differences in the Systemic and Local Immune Response of Pancreatic Cancer Patients. *Cancers* **2023**, *15*, 1815. [[PubMed](#)]
- Ao, C.; Jiao, S.; Wang, Y.; Yu, L.; Zou, Q. Biological Sequence Classification: A Review on Data and General Methods. *Research* **2022**, *2022*, 0011.
- Wang, H.; Guo, F.; Du, M.; Wang, G.; Cao, C. A novel method for drug-target interaction prediction based on graph transformers model. *BMC Bioinform.* **2022**, *23*, 459.
- Zhang, Z.; Cui, F.; Wang, C.; Zhao, L.; Zou, Q. Goals and approaches for each processing step for single-cell RNA sequencing data. *Brief. Bioinform.* **2020**, *22*, bbaa314.
- Zhang, Z.; Cui, F.; Cao, C.; Wang, Q.; Zou, Q. Single-cell RNA analysis reveals the potential risk of organ-specific cell types vulnerable to SARS-CoV-2 infections. *Comput. Biol. Med.* **2021**, *140*, 105092. [[PubMed](#)]
- Chao, W.; Quan, Z. A Machine Learning Method for Differentiating and Predicting Human-Infective Coronavirus Based on Physicochemical Features and Composition of the Spike Protein. *Chin. J. Electron.* **2021**, *30*, 815–823.
- Cui, F.; Li, S.; Zhang, Z.; Sui, M.; Cao, C.; El-Latif Hesham, A.; Zou, Q. DeepMC-iNABP: Deep learning for multiclass identification and classification of nucleic acid-binding proteins. *Comput. Struct. Biotechnol. J* **2022**, *20*, 2020–2028.
- Zhang, Z.; Cui, F.; Zhou, M.; Wu, S.; Zou, Q.; Gao, B. Single-cell RNA Sequencing Analysis Identifies Key Genes in Brain Metastasis from Lung Adenocarcinoma. *Curr. Gene Ther.* **2021**, *21*, 338–348. [[CrossRef](#)]
- Mendes, M.; Mahita, J.; Blazeska, N.; Greenbaum, J.; Ha, B.; Wheeler, K.; Wang, J.Y.; Shackelford, D.; Sette, A.; Peters, B. IEDB-3D 2.0: Structural data analysis within the Immune Epitope Database. *Protein Sci.* **2023**, *32*, e4605.
- Tirziu, A.; Avram, S.; Mada, L.; Crişan-Vida, M.; Popovici, C.; Popovici, D.; Faur, C.; Duda-Seiman, C.; Paunescu, V.; Vernic, C. Design of a Synthetic Long Peptide Vaccine Targeting HPV-16 and -18 Using Immunoinformatic Methods. *Pharmaceutics* **2023**, *15*, 1798.
- Nagpal, G.; Usmani, S.S.; Dhanda, S.K.; Kaur, H.; Singh, S.; Sharma, M.; Raghava, G.P.S. Computer-aided designing of immunosuppressive peptides based on IL-10 inducing potential. *Sci. Rep.* **2017**, *7*, 42851. [[CrossRef](#)]
- Singh, O.; Hsu, W.-L.; Su, E.C.-Y. ILekin10Pred: A Computational Approach for Predicting IL-10-Inducing Immunosuppressive Peptides Using Combinations of Amino Acid Global Features. *Biology* **2022**, *11*, 5.
- Liu, H.; Xu, Y.; Chen, F. Sketch2Photo: Synthesizing photo-realistic images from sketches via global contexts. *Eng. Appl. Artif. Intell.* **2023**, *117*, 105608. [[CrossRef](#)]
- Liu, M.; Zhang, X.; Yang, B.; Yin, Z.; Liu, S.; Yin, L.; Zheng, W. Three-Dimensional Modeling of Heart Soft Tissue Motion. *Appl. Sci.* **2023**, *13*, 2493. [[CrossRef](#)]
- Yang, B.; Li, Y.; Zheng, W.; Yin, Z.; Liu, M.; Yin, L.; Liu, C. Motion prediction for beating heart surgery with GRU. *Biomed. Signal Process. Control* **2023**, *83*, 104641. [[CrossRef](#)]
- Yang, S.; Li, Q.; Li, W.; Li, X.; Liu, A.A. Dual-Level Representation Enhancement on Characteristic and Context for Image-Text Retrieval. *IEEE Trans. Circ. Syst. Video Technol.* **2022**, *32*, 8037–8050. [[CrossRef](#)]
- Waziry, S.; Wardak, A.B.; Rasheed, J.; Shubair, R.M.; Rajab, K.; Shaikh, A. Performance comparison of machine learning driven approaches for classification of complex noises in quick response code images. *Heliyon* **2023**, *9*, e15108. [[CrossRef](#)] [[PubMed](#)]
- Farooq, M.S.; Khalid, H.; Arooj, A.; Umer, T.; Asghar, A.B.; Rasheed, J.; Shubair, R.M.; Yahyaoui, A. A Conceptual Multi-Layer Framework for the Detection of Nighttime Pedestrian in Autonomous Vehicles Using Deep Reinforcement Learning. *Entropy* **2023**, *25*, 135. [[CrossRef](#)]

27. Le, H.D.; Lee, G.S.; Kim, S.H.; Kim, S.; Yang, H.J. Multi-Label Multimodal Emotion Recognition With Transformer-Based Fusion and Emotion-Level Representation Learning. *IEEE Access* **2023**, *11*, 14742–14751. [[CrossRef](#)]
28. Yang, L.; Yu, X.Y.; Zhang, S.P.; Zhang, H.H.; Xu, S.; Long, H.B.; Zhu, Y.W. Stacking-based and improved convolutional neural network: A new approach in rice leaf disease identification. *Front. Plant Sci.* **2023**, *14*, 1165940. [[CrossRef](#)] [[PubMed](#)]
29. Kalule, R.; Abderrahmane, H.A.; Alameri, W.; Sassi, M. Stacked ensemble machine learning for porosity and absolute permeability prediction of carbonate rock plugs. *Sci. Rep.* **2023**, *13*, 9855. [[CrossRef](#)] [[PubMed](#)]
30. Li, Y.J.; Ma, D.; Chen, D.; Chen, Y. ACP-GBDT: An improved anticancer peptide identification method with gradient boosting decision tree. *Front. Genet.* **2023**, *14*, 1165765. [[CrossRef](#)]
31. Mardikoraem, M.; Woldring, D. Protein Fitness Prediction Is Impacted by the Interplay of Language Models, Ensemble Learning, and Sampling Methods. *Pharmaceutics* **2023**, *15*, 1337. [[CrossRef](#)]
32. Bao, W.Z.; Gu, Y.J.; Chen, B.T.; Yu, H.P. Golgi_DF: Golgi proteins classification with deep forest. *Front. Neurosci.* **2023**, *17*, 1197824. [[CrossRef](#)]
33. Nath, A.; Subbiah, K. The role of pertinently diversified and balanced training as well as testing data sets in achieving the true performance of classifiers in predicting the antifreeze proteins. *Neurocomputing* **2018**, *272*, 294–305. [[CrossRef](#)]
34. Elreedy, D.; Atiya, A.F.; Kamalov, F. A theoretical distribution analysis of synthetic minority oversampling technique (SMOTE) for imbalanced learning. *Mach. Learn.* **2023**, *112*. [[CrossRef](#)]
35. Mursalim, M.K.N.; Mengko, T.L.E.R.; Hertadi, R.; Purwianti, A.; Susanty, M. BiCaps-DBP: Predicting DNA-binding proteins from protein sequences using Bi-LSTM and a 1D-capsule network. *Comput. Biol. Med.* **2023**, *163*, 107241. [[CrossRef](#)] [[PubMed](#)]
36. Kawashima, S.; Pokarowski, P.; Pokarowska, M.; Kolinski, A.; Katayama, T.; Kanehisa, M. AAIindex: Amino acid index database, progress report 2008. *Nucleic Acids Res.* **2008**, *36*, D202–D205. [[CrossRef](#)] [[PubMed](#)]
37. Alley, E.C.; Khimulya, G.; Biswas, S.; AlQuraishi, M.; Church, G.M. Unified rational protein engineering with sequence-based deep representation learning. *Nat. Methods* **2019**, *16*, 1315–1322. [[CrossRef](#)] [[PubMed](#)]
38. Gao, Y.; Lewis, N.; Calhoun, V.D.; Miller, R.L. Interpretable LSTM model reveals transiently-realized patterns of dynamic brain connectivity that predict patient deterioration or recovery from very mild cognitive impairment. *Comput. Biol. Med.* **2023**, *161*, 107005. [[CrossRef](#)]
39. Zhao, S.; Meng, J.; Wekesa, J.S.; Luan, Y. Identification of small open reading frames in plant lncRNA using class-imbalance learning. *Comput. Biol. Med.* **2023**, *157*, 106773. [[CrossRef](#)] [[PubMed](#)]
40. Cao, C.; Kossinna, P.; Kwok, D.; Li, Q.; He, J.; Su, L.; Guo, X.; Zhang, Q.; Long, Q. Disentangling genetic feature selection and aggregation in transcriptome-wide association studies. *Genetics* **2022**, *220*, iyab216. [[CrossRef](#)]
41. Ao, C.; Ye, X.; Sakurai, T.; Zou, Q.; Yu, L. m5U-SVM: Identification of RNA 5-methyluridine modification sites based on multi-view features of physicochemical features and distributed representation. *BMC Biol.* **2023**, *21*, 93. [[CrossRef](#)]
42. Imakura, A.; Kihira, M.; Okada, Y.; Sakurai, T. Another use of SMOTE for interpretable data collaboration analysis. *Expert Syst. Appl.* **2023**, *228*, 120385. [[CrossRef](#)]
43. Jia, H.C. Simulation of English part-of-speech classification based on artificial intelligence and additive logistic regression. *Soft Comput.* **2023**, *27*. [[CrossRef](#)]
44. Wang, J.F.; Huang, S.H.; Wang, Z.W.; Huang, D.; Qin, J.; Wang, H.; Wang, W.Z.; Liang, Y. A calibrated SVM based on weighted smooth GL(1/2) for Alzheimer's disease prediction. *Comput. Biol. Med.* **2023**, *158*, 106752. [[CrossRef](#)]
45. Zhang, H.; Zou, Q.; Ju, Y.; Song, C.; Chen, D. Distance-based Support Vector Machine to Predict DNA N6-methyladenine Modification. *Curr. Bioinform.* **2022**, *17*, 473–482. [[CrossRef](#)]
46. Zhang, Z.; Cui, F.; Lin, C.; Zhao, L.; Wang, C.; Zou, Q. Critical downstream analysis steps for single-cell RNA sequencing data. *Brief. Bioinform.* **2021**, *22*, bbab105. [[CrossRef](#)] [[PubMed](#)]
47. Li, Z.M.; Zhao, Y.; Duan, T.; Dai, J.Q. Configurational patterns for COVID-19 related social media rumor refutation effectiveness enhancement based on machine learning and fsQCA. *Inf. Process. Manag.* **2023**, *60*, 103303. [[CrossRef](#)] [[PubMed](#)]
48. Hurtado, M.; Mora-Marquez, F.; Soto, A.; Marino, D.; Goicoechea, P.G.; de Heredia, U.L. DEGoldS: A Workflow to Assess the Accuracy of Differential Expression Analysis Pipelines through Gold-standard Construction. *Curr. Bioinform.* **2023**, *18*, 296–309. [[CrossRef](#)]
49. Cevik, T.; Cevik, N.; Rasheed, J.; Abu-Mahfouz, A.M.; Osman, O. Facial Recognition in Hexagonal Domain—A Frontier Approach. *IEEE Access* **2023**, *11*, 46577–46591. [[CrossRef](#)]
50. Cao, C.; Wang, J.; Kwok, D.; Cui, F.; Zhang, Z.; Zhao, D.; Li, M.J.; Zou, Q. webTWAS: A resource for disease candidate susceptibility genes identified by transcriptome-wide association study. *Nucleic Acids Res.* **2022**, *50*, D1123–D1130. [[CrossRef](#)]
51. Zhang, Z.; Cui, F.; Su, W.; Dou, L.; Xu, A.; Cao, C.; Zou, Q. webSCST: An interactive web application for single-cell RNA-sequencing data and spatial transcriptomic data integration. *Bioinformatics* **2022**, *38*, 3488–3489. [[CrossRef](#)] [[PubMed](#)]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.