EAI Endorsed Transactions
on AI and Robotics
Research Article EAI.EU

Enhancing the Prediction of IL-4 Inducing Peptides Using Stacking Ensemble Model

Rajib Mia[1], Tawfiqul Hasan[1], Abu Kowshir Bitto[1], Mohammad Mahadi Hassan[2], Mohammed Shamsul Alam[2], Abdul Kadar Muhammad Masum[3*]

[1]Department of Software Engineering, Daffodil International University, Dhaka, Bangladesh
[2]Department of Computer Science and Engineering, International Islamic University Chittagong, Chittagong, Bangladesh
[3]Department of Computer Science and Engineering, Southeast University, Dhaka, Bangladesh

Abstract

Interleukin-4 (IL-4) plays a critical role in immune regulation and inflammation suppression, and therefore precise prediction is important in immunotherapy and vaccine design. In this work, we present an innovative stacking ensemble-based predictive model for IL-4-inducing peptide discovery. The method combines the group of feature extraction techniques, i.e., Amino Acid Composition (AAC), Amphiphilic Pseudo Amino Acid Composition (APAAC), and their combinations, and their pruning using SHAP (SHapley Additive exPlanations) with only the most relevant features being retained. To solve the class imbalance problem inherent in the peptide data, the ADASYN (Adaptive Synthetic Sampling) algorithm was applied for synthetic oversampling. We applied eight machine learning classifiers: Logistic Regression, Random Forest, Support Vector Classifier, Decision Tree, K-Nearest Neighbors, XGBoost, LightGBM, and a stacking ensemble model, enabling the strong prediction on both imbalanced and balanced datasets. Our evaluation demonstrates the stacking model's better performance on the imbalanced and balanced dataset. Surprisingly, with combined characteristics, the stacking model over the independent test set yielded accuracy of 89.97% and Matthew's Correlation Coefficient (MCC) as 0.79. Accurate comparisons of performance over AAC and APAAC feature spaces indicate that the stacking model performs better than other classifiers in all instances, albeit more so under balanced scenarios, referring to data rebalancing requirements. This research not only highlights the precision of stacking ensembles in peptide classification tasks but also urges the integration of interpretable feature selection and data balancing in future immunoinformatic pipelines.

Keywords: Immunoinformatic, Peptides, Interleukin-4, Artificial Intelligence, Machine learning, Stacking Ensemble.

Received on 04 August 2025, accepted on 11 October 2025, published on 27 October 2025

doi: 10.4108/airo.9867

1. Introduction

The immune system is a complex network of cells, molecules, and signaling pathways that function coordinately to defend the body against infection and to preserve homeostasis. Of the enormous repertoire of components in this system, cytokines are instrumental in conducting immune responses, such as inflammation, repair of tissue, and immune modulation [1]. Interleukin-4 (IL-4) is a critical cytokine that has been recognized for playing multifunctional roles in both immunomodulatory and anti-inflammatory cascades. Revealing the underlying mechanisms of IL-4 action is crucial to devising effective therapeutics for immune disorders [2]. It is one of its primary functions to induce the differentiation of naïve T helper (Th) cells to Th2 cells, which has a critical function in helping with humoral immunity and antibody response [3]. IL-4 stimulates B cell growth and the production of immunoglobulins such as IgE and IgG1—both important in the defence against extracellular pathogens. IL-

*Corresponding author. Email: akmmasum@yahoo.com

4 also enhances the expression of major histocompatibility complex (MHC) class II molecules on antigen-presenting cells, thus maximizing immune response efficiency. All these effects position IL-4 centrally in immune system regulation [4]. While it has protective roles, IL-4 also plays a role in the pathogenesis of allergic illnesses such as asthma, eczema, and allergic rhinitis [5]. This is since IL-4 may augment IgE production and cause Th2-polarized immune responses [6]. This contradictory characteristic of IL-4 being both a regulator and a possible etiology of disease renders it a slippery therapeutic target.

IL-4 impacts autoimmune diseases by modulating B cell function and inducing Th2 responses, and research has linked it with cancer development by its potential to block cell death and induce tumor cell survival [7]. Computational immunology advances have made way for new exploration of cytokine interactions and regulatory mechanisms [8]. IL-4 has been emphasized for its potential as a therapeutic target for disease modulation. Accurate prediction of IL-4-inducing peptides would form the basis for new treatments that selectively enhance or suppress IL-4 activity according to clinical needs. However, acquiring high predictive accuracy remains a problem due to the complexity of peptide-protein interactions and variety in immune reactions. We address these issues within this paper by focusing on the improvement of IL-4-inducing peptide predictive accuracy.

Despite numerous machine learning efforts for IL-4-inducing peptide prediction, most existing models suffer from limited generalizability, class imbalance sensitivity, and a lack of interpretability. Many previous works fail to incorporate ensemble learning or deep feature engineering that combines sequence-based and physicochemical properties. Moreover, comparative evaluations with various data balancing and encoding strategies remain underexplored.

Based on recent literature and advances in technology, we employ advanced machine learning techniques to improve peptide prediction models. Specifically, the study employs a stacking ensemble approach that integrates various amino acid feature encoding methods with the aim of achieving greater performance than to other predictive models [9]. The model has been exhaustively tested against recognized datasets and has demonstrated consistent improvement in accuracy. Such consistency is a measure of its viability as a valuable research tool as well as for possible clinical application. The implications of this work have the potential to contribute to the development of targeted therapies for diseases in which interleukin-4 (IL-4) is a central protagonist—e.g., allergic disease, autoimmune disorders, and some cancers.

In this study, our contributions include, we employed a comprehensive and diverse feature extraction strategy using iLearnPlus. This multi-perspective representation captures both sequence composition and physicochemical properties, enabling the model to learn intricate peptide patterns crucial for IL-4 induction. To address the significant class imbalance in the dataset, we applied advanced data balancing techniques like ADASYN, which dynamically generates synthetic samples based on learning difficulty. This approach improves the model's ability to accurately detect and classify minority class (IL-4-inducing) peptides, ensuring balanced learning and reducing bias. We propose a robust stacking ensemble learning framework that integrates diverse base classifiers—Random Forest, Support Vector Machine, XGBoost, and LightGBM—with Logistic Regression as a meta-learner. This hierarchical architecture leverages the unique strengths of each model to enhance predictive accuracy and generalization for IL-4-inducing peptide identification.

2. Related Work

IL-4 is one of the most prominent immune regulatory cytokines, and IL-4-inducing peptide prediction is a significant consideration in vaccine development. Various computational tools have been employed since the early days to construct IL-4 peptide prediction models. Some of the common tools used are motif-based search algorithms, quantitative matrix (QM) algorithms, and more recently, advanced machine learning models. The QM approach has proved important in delineating a distinct image of the function of individual amino acids in peptide recognition by different MHC loci. Conventional approaches to T-cell epitope prediction, often based on MHC class I binding assumptions, may be sub-optimal at times [10]. IL-4-inducing peptides identification using traditional lab techniques has been cumbersome, time-consuming, and labor-intensive in the past. Computational approaches offer a robust alternative by dramatically reducing the burden of experiments with improved prediction efficiency [11].

Machine learning (ML) is becoming a fundamental tool for biomedical science. By analyzing large and complex data sets, ML is capable of uncovering faint patterns that other strategies often do not identify. It has been widely applied across many disciplines, from disease diagnosis through drug discovery to personalized medicine. Its strength lies in the capacity to integrate heterogeneous data types—genomic, proteomic, and clinical data sets, for instance—providing greater understanding of biological processes [12]. Researchers have employed ML algorithms to enhance IL-4 peptide prediction accuracy.

Most of the existing models are overfitting-prone and are not interpretable, especially when trained on small or noisy biological datasets. Moreover, the lack of high-quality, well-labelled data even further limits the generalizability of such models [13]. To better handle these issues, this paper proposes a stacking ensemble model—a powerful meta-learning approach that integrates multiple classifiers with the aim of achieving optimal predictive accuracy and robustness in IL-4 peptide prediction. Machine learning provides unprecedented potential in biological study, especially using supervised algorithms such as support vector machines, random forests, and neural networks. All these approaches have unique contributions to biomedical data analysis. Nevertheless, model validation, interpretability, and quality of data are crucial for the success of ML applications. Through the suggestion of a strong stacking ensemble approach, this article adds an even more accurate and reliable

computer method for IL-4 peptide prediction—thus further pushing the boundary of rational vaccine design.

3. Methodology

The approach employed in the present research is a meticulous process of organized and adhered-to procedures shows in figure 1. First, the dataset includes 985 IL-4-inducing peptides and 744 non-inducing peptides. In the case of addressing the class imbalance, feature extraction is conducted, and both ADASYN (Adaptive Synthetic Sampling) is employed. The use of advanced resampling techniques is crucial in helping the dataset attain an equal split of the two classes, thereby reducing the potential bias from the original imbalance. The resampling process is followed by an efficient 5-fold cross validation method. The use of this cross-validation technique is crucial in model evaluation and in ensuring the model to perform adequately on various subsets of the data. The data are divided into five folds, and the model is rigorously tested iteratively, using four folds for training and a validation one. Feature selection is performed to find out the relevant features using SHAP. In the third step, the constructed model undergoes a critical evaluation and is found to be effective in making good-quality predictions for IL-4 inducing peptides. Predictability of the approach adopted is examined using various performance measures to make it reliable and practical for IL-4-inducing peptides identification. Research methodology employed here uses rich reproducing techniques, cross-validation processes, and selection processes to enable improvement in robustness and accuracy as far as model efficiency is concerned to predict inducing peptides IL-4.

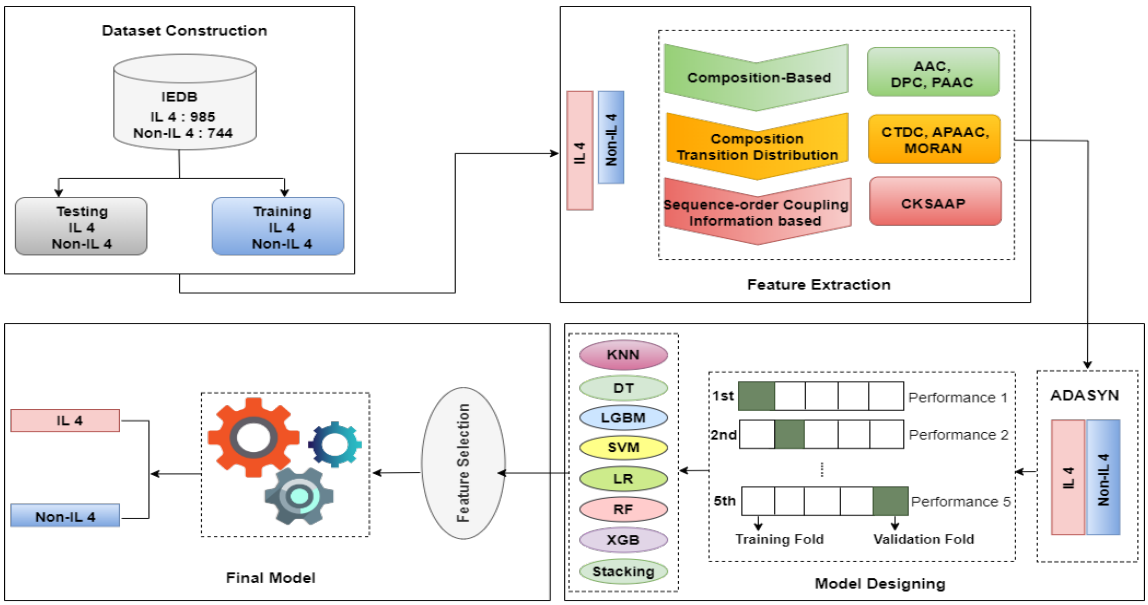

Figure 1. Our Proposed Methodology for prediction of IL 4 and Non-IL 4.

3.1. Dataset

The dataset was downloaded from the Immune Epitope Database (IEDB), a clinical research database of antibody and T cell epitopes. The dataset was produced by IEDB, which were experimentally confirmed for their ability to induce IL-4. Peptides that were not IL-4-inducing peptides were referred to as IL-4-non-inducing peptides. The final dataset included 744 IL-4-non-inducing peptide variants and 985 IL-4-inducing peptide variants.

3.2. Data Balancing

Class imbalance is a common challenge in machine learning that can lead to biased models which overestimate the majority class in most cases, with impaired minority class prediction accuracy. There is considerable class imbalance in the dataset used in this study: 985 peptides are IL-4-inducing, and 744 peptides are not IL-4-inducing. This class imbalance can reduce model performance, particularly in the correct identification of minority class instances. To fight against this issue, the Adaptive Synthetic Sampling (ADASYN) algorithm was used [14]. ADASYN is a strong oversampling method that seeks to reduce bias by generating synthetic samples for minority classes. ADASYN differs from the previous methods in that it dynamically adjusts the number of the generated synthesized samples based on the learning

difficulty of each minority instance. Synthetic example generation in ADASYN is governed using a weight parameter, λ, which controls the number of synthesized examples to be created for each minority instance. This weight is determined based on the minority class's density distribution in relation to how close it is to majority class neighbors. The closer a minority sample is to majority samples, the higher the λ value it is assigned, and consequently, more synthetic points are created around it. This guarantees that the new samples maintain the distributional characteristics of the original data but improve on class balance. Synthetic samples are prepared using the formula equation (1):

$$s_i = x_i + \lambda(x_1 - x_i) \quad (1)$$

3.3. Feature Extraction

Feature extraction is a central step in applying machine learning techniques to peptide sequence analysis [15] [16]. Feature extraction used in this study: Amino Acid Composition (AAC), Amphiphilic Pseudo Amino Acid Composition (APAAC), Composition of k-spaced Amino Acid Pairs (CKSAAP), Composition-Transition-Distribution (CTDC), Conjoint Triad of Codons (CTRAID), Dipeptide Composition (DPC) and Pseudo Amino Acid Composition (PAAC). These methods are categorized into three general classes: amino acid composition-based, composition-transition-distribution models, and sequence-order-based descriptors. To streamline and enhance the computational efficiency in feature encoding for the identification of IL-4-inducing peptides from sequences, the iLearnPlus platform was used [17]. The platform consolidates four functionalities of utmost significance into a streamlined, user-friendly interface, enhancing the feature encoding process.

1) Amino Acid Composition (AAC): The AAC method establishes the relative frequency of each amino acid in a protein or peptide sequence [17]. AAC provides a complete structural composition profile of the sequence and identifies significant patterns that are indicative of IL-4-inducing capacity. AAC is instrumental in enhancing model performance by identifying distinctive amino acid patterns that are vital to IL-4 induction.
2) Amphiphilic Pseudo Amino Acid Composition (APAAC): APAAC interacts both sequence information at local and global levels by examining the amphiphilic nature of amino acids their hydrophobicity, and hydrophilicity [18]. It considers the way these chemical features are distributed across amino acid pairs and yields a more advanced feature vector. This enables the model to better identify IL-4 inducing structural motifs.
3) Composition of k-spaced Amino Acid Pairs (CKSAAP): CKSAAP captures sequence-order information by analyzing the frequency of amino acid pairs separated by k residues [19]. This operation imputes spatial proximities between amino acids, enhancing the model's ability to identify sequential patterns that determine biological processes like IL-4 activation.
4) Conjoint Triad Descriptor of Codons (CTRAID): This method analyzes the frequency and organization of codon triplets based on their physicochemical attributes [20]. By grouping codons and identifying their frequency in sets of three, CTRAID provides a comprehensive characterization of gene sequences. It is particularly beneficial in separating patterns of gene expression and biological function.
5) Dipeptide Composition (DPC): DPC examines the sequence in terms of dipeptide frequency—two consecutive amino acids [21]. It reflects short-range interactions in the sequence and gives information about structural and functional elements accountable for IL-4 induction. It facilitates building more precise predictive models.
6) Pseudo Amino Acid Composition (PAAC): PAAC differs from standard AAC in that it uses sequence-order information and physicochemical characteristics of amino acids [22]. PAAC-based feature vectors consider both the frequency and distribution of amino acids, as well as their biochemical characteristics. PAAC is particularly adept at describing complex biological processes such as IL-4 production.

After feature extraction, SHAP (SHapley Additive exPlanations) [27] identified AAC, APAAC, and their combination as the most influential for IL-4-inducing peptide prediction [23]. The top retained features included AAC frequencies of leucine, lysine, and glycine, and APAAC descriptors related to hydrophobicity, polarity, and solvent accessibility. These features are optimal because AAC captures overall residue composition, while APAAC encodes key physicochemical properties, providing complementary information that improved prediction performance compared to CKSAAP or DPC.

3.4. Model Development

We observe from Table I, that all machine learning models contribute uniquely to IL-4-inducing peptide prediction. The Random Forest model uses an ensemble of decision trees based on the Gini Index to classify peptide sequences accurately. Logistic Regression is a probabilistic approach to IL-4 induction prediction using maximum likelihood estimation. Support Vector Machine (SVM) identifies optimal hyperplanes to distinguish inducing and non-inducing peptides. XGBoost and LightGBM, both of which are gradient boosting algorithms, learn complex nonlinear relationships and possess regularization for improvement in generalization. K-Nearest Neighbors (KNN) applies localized distance-based measurements for identifying similarities among peptides. The Decision Tree model provides interpretable rules to explain IL-4 induction. Finally, Stacking Classifier stacks a collection of base learners—KNN, Decision Tree, LightGBM, and SVM—upon a meta-

learner (Logistic Regression) to achieve better overall prediction accuracy through the combined strength of each model.

Table 1. Our Applied Model Description and Workflow of the Models in Term Peptide Prediction

Model	Model Work for Peptide Captures
Random Forest	This ensemble method utilizes a fusion of many decision trees to produce predictions by means of a majority voting procedure. The Gini index shows in equation 2, which quantifies the lack of purity in a dataset, is computed in the following manner where we use this metric to evaluate the effectiveness of decision trees in distinguishing between peptides that induce IL-4 and those that do not. $$\boldsymbol{Gini\ Index} = \sum_{i=1}^{n} \boldsymbol{p_i^2} \quad (2)$$
Logistic Regression	We use logistic regression to predict the probability that a peptide induces IL-4 based on its features. MLE-based training is used to the model so that it can optimize its coefficients optimally for the data. It uses the logistic function, Equation (3), to give a probability value between 0 and 1 for the induction of IL-4: $$\boldsymbol{Py(1|x) = \sigma(\beta 0 + \beta 1 x_1 + \cdots)} \quad (3)$$
Support Vector Machine	A discriminative method is used by SVM to find a hyperplane that effectively separates peptides that cause IL-4 from those that do not shows in equation 4. The weight vector, denoted as W, represents the weights assigned to the input characteristics, represented by x. The bias component, denoted as b, is a constant term. $$\boldsymbol{w^T x + b = 0} \quad (4)$$
XGBoost	XGBoost has ability to manage intricate data structures and handle missing information makes it highly effective at forecasting IL-4 induction. The model utilizes gradient boosting to optimize performance by combining numerous weak learners into a robust predictive model. In XGBoost, the objective function O(θ) consists of two components: the loss function L(θ) and the regularization term Ω(θ) shows in equation 5: $$\boldsymbol{O(\theta) = L(\theta) + \Omega(\theta)} \quad (5)$$
Light Gradient Boosting Machine	LightGBM well suited for larger dataset and its great option for IL-4 induction prediction. LightGBM's work fllow shows in equation 6 objective function O(t) is composed of the loss function L(t), a regularization term Ω(t), and an additional parameter C to control tree complexity and mitigate overfitting. $$\boldsymbol{0(T) = L(T) + \Omega(T) + C} \quad (6)$$
K-Nearest Neighbors	It is particularly useful for detecting local relationships within the feature space of IL-4-inducing peptides. K-nearest neighbors (KNN) algorithm assigns a class label to a new instance by examining the k closest data points in the feature space and determining the most common class among its neighbors. The distance metric shows in equation 7, often measured using the Euclidean distance formula: $$\boldsymbol{d(x,y) = \sqrt{\sum_{i=1}^{n}(x_i - \dot{y}_i)^2}} \quad (7)$$
Decision Tree	We use a decision tree to guess peptides that cause IL-4 by looking at different parts of the peptides and seeing how they affect IL-4 production. This makes an excellent model for understanding and predicting IL-4 induction patterns based on peptide properties.
Stacking Classifier	The Stacking Classifier is an ensemble technique that stacks a number of base learners in a hierarchical fashion to learn higher-order features in peptide data. In this work, we employed it to improve IL-4-inducing peptide prediction. KNN, Decision Tree, SVM, and LightGBM base classifiers were trained using stratified 5-fold cross-validation. For each fold, the out-of-fold probability scores from these classifiers were collected and concatenated to create a new meta-feature matrix. The meta-learner, a Logistic Regression model with L2 regularization (C=1), was trained on this matrix, stacking the probability-based outputs to arrive at the final prediction. The base models were tuned with default hyperparameters: KNN (k=3–15, Euclidean distance), Decision Tree (max depth 5–50, Gini/entropy criterion), SVM (RBF kernel, γ=0.001–0.1), and LightGBM (estimators 100–500, learning rate 0.01–0.2, depth 3–10). The whole workflow of this two-level stacking method is presented in Figure 2.

3.5. Performance Evaluation

Performance evaluation of predictive models is a critical step towards ensuring robustness and accuracy. In this study, a cautious selection of evaluation criteria for IL-4-inducing peptide prediction is practiced throughout. The modeling process follows a rigorous protocol with both 5-fold cross-validation. To compare the models in an exhaustive manner, several performance metrics are calculated like sensitivity, specificity, accuracy, and Matthews Correlation Coefficient (MCC) as shown in Equations 8 to 11. Area Under the Receiver Operating Characteristic Curve (AUC) is also calculated as a threshold-free measure to verify the overall discriminative ability of every model. Greater AUC shows more predictive capability. Specificity shows Equation 10, determines the rate of true negative cases correctly identified, gauging how well the model avoids false positives. Sensitivity, which is interested in how well the model correctly identifies IL-4-inducing peptides, gauges the effectiveness of the model at recognizing positive cases. Accuracy, shows in Equation 8, is the proportion of correctly predicted outcomes over the number of total predictions and gives a general indication of model accuracy. The Matthews Correlation Coefficient (MCC) is a balanced metric that considers true and false positives and negatives. In contrast with accuracy, MCC is particularly useful when dealing with imbalanced datasets since it provides a better understanding of the performance of the model over the two classes [24 - 26].

$$Accuracy = \frac{TP + TN}{TP+TN+FN+FP} \times 100\% \quad (8)$$

$$Sensitivity = \frac{TP}{TP+FN} \times 100\% \quad (9)$$

$$Specificity = \frac{TN}{TN + FP} \times 100\% \quad (10)$$

$$MCC = \frac{(TP \times TN)-(FP \times FN)}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}} \times 100\% \quad (11)$$

Here, TP = True Positive, TN = True Negative, FP = False Positive, and FN = False Negative.

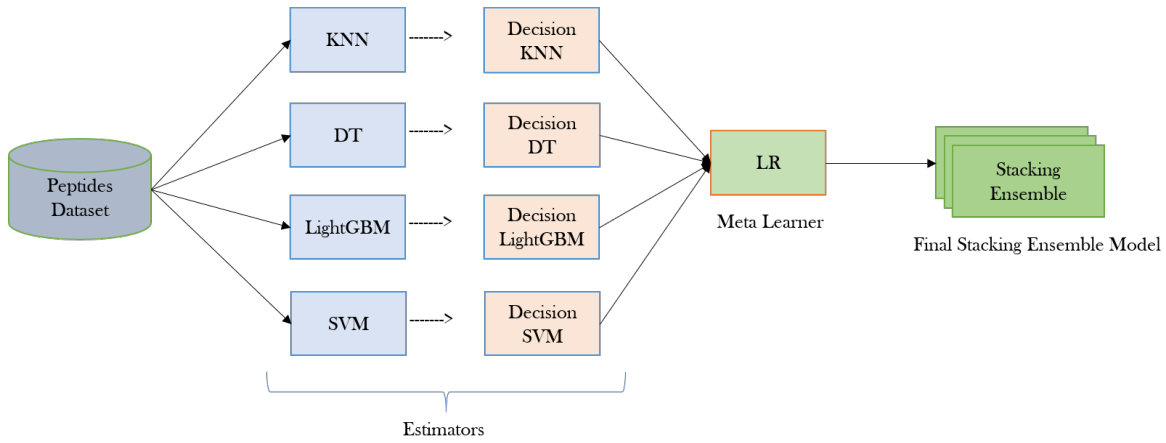

Figure 2. Our proposed stacking ensemble model architecture diagram.

4. Result and Discussion

Results of different machine learning models applied to various feature representations and dataset versions are reported and discussed below. Our aim was to investigate how stacking ensemble learning, together with ADASYN-balanced training data and chosen features, improves IL-4-inducing peptide prediction. Both imbalanced and ADASYN-balanced datasets are explored, tested on AAC, APAAC, and merged feature representations. Performance metrics such as accuracy, AUC, MCC, sensitivity, and specificity were compared between models and feature sets with particular interest in stacking model performance.

This study was directed towards the prediction of IL-4 producing peptides from an imbalanced APAAC features. Among the models experimented with, the Stacking classifier presented top scores, reaching accuracy of 88.74%, MCC of 0.7754, AUC of 0.8879, with sensitivity and specificity scores of 90.30% and 87.27%, respectively. These measures show that the Stacking model not only performs extremely well overall but is also extremely robust at making correct predictions of peptides that activate IL-4, even in the face of negative consequences of data imbalance. As a point of comparison, the Logistic Regression (LR) model fared the worst with an accuracy of 63.00%, MCC of 0.2629, AUC of 0.6252, sensitivity of 47.65%, and specificity of 77.40%. This means that LR struggles to capture the complex dependencies in the data as well as the other models. The other notable classifiers, including Random Forest (RF) and XGBoost (XGBClassifier), also did well with accuracy rates of 85.79% and 82.57%, respectively, as well as similar MCC and AUC values, which show that they are appropriate for this classification task. Stacking model in improving the accuracy

of IL-4-producing peptide prediction in imbalanced class distribution datasets.

Table 2. Performance Comparison of Diverse Classifiers Employing APAAC Features on Imbalanced Dataset

Classifier	Accuracy	MCC	AUC	Sensitivity	Specificity
LR	0.6300	0.2629	0.6252	0.4765	0.7740
RF	0.8579	0.7162	0.8568	0.8227	0.8909
SVC	0.7064	0.4164	0.7031	0.5983	0.8078
XGB	0.8257	0.6552	0.8235	0.7535	0.8935
DT	0.7802	0.5691	0.7767	0.6676	0.8857
KNN	0.6635	0.3258	0.6617	0.6039	0.7195
LGBM	0.8083	0.6222	0.8056	0.7202	0.8909
Stacking	**0.8874**	**0.7754**	**0.8879**	**0.9030**	**0.8727**

Table 3. Performance Comparison of Diverse Classifiers Employing AAC Features on Imbalanced Dataset

Classifier	Accuracy	MCC	AUC	Sensitivity	Specificity
LR	0.6046	0.2099	0.5995	0.4432	0.7558
RF	0.8472	0.6958	0.8456	0.7978	0.8935
SVC	0.6327	0.2722	0.6271	0.4515	0.8026
XGB	0.8097	0.6262	0.8067	0.7147	0.8987
DT	0.7761	0.5664	0.7719	0.6399	0.9039
KNN	0.6501	0.2988	0.6481	0.5845	0.7117
LGBM	0.8177	0.6393	0.8153	0.7424	0.8883
Stacking	**0.8807**	**0.7628**	**0.8815**	**0.9058**	**0.8571**

Table 4. Performance Comparison of Diverse Classifiers Employing AAC Features on Balanced Dataset

Classifier	Accuracy	MCC	AUC	Sensitivity	Specificity
LR	0.5685	0.1370	0.5685	0.5638	0.5732
RF	0.8782	0.7579	0.8780	0.8444	0.9116
SVC	0.6650	0.3299	0.6649	0.6531	0.6768
XGB	0.8388	0.6887	0.8384	0.7474	0.9293
DT	0.7830	0.5779	0.7825	0.6786	0.8864
KNN	0.7018	0.4042	0.7016	0.6684	0.7348
LGBM	0.8160	0.6356	0.8157	0.7602	0.8712
Stacking	**0.8959**	**0.7919**	**0.8960**	**0.9005**	**0.8914**

Table 5. Performance Comparison of Diverse Classifiers Employing APAAC Features on Balanced Dataset

Classifier	Accuracy	MCC	AUC	Sensitivity	Specificity
LR	0.5787	0.1573	0.5786	0.5663	0.5909
RF	0.8807	0.7625	0.8806	0.8520	0.9091
SVC	0.7018	0.4037	0.7017	0.6811	0.7222
XGB	0.8236	0.6534	0.8232	0.7526	0.8939
DT	0.8046	0.6284	0.8039	0.6786	0.9293
KNN	0.6954	0.3914	0.6953	0.6633	0.7273
LGBM	0.8211	0.6469	0.8207	0.7577	0.8838
Stacking	**0.8947**	**0.7903**	**0.8948**	**0.9184**	**0.8712**

Table 6. Performance Comparison of Diverse Classifiers Employing AAC+APAAC Features on Balanced Dataset

Classifier	Accuracy	MCC	AUC	Sensitivity	Specificity
LR	0.5647	0.1293	0.5647	0.5718	0.5575
RF	0.8604	0.7274	0.8609	0.7960	0.9258
SVC	0.6865	0.3754	0.6869	0.6398	0.7340
XGB	0.8261	0.6612	0.8267	0.7481	0.9054
DT	0.7944	0.6078	0.7953	0.6751	0.9156
KNN	0.7043	0.4102	0.7046	0.6675	0.7417
LGBM	0.8096	0.6260	0.8102	0.7406	0.8798
Stacking	**0.8997**	**0.7995**	**0.8997**	**0.9018**	**0.8977**

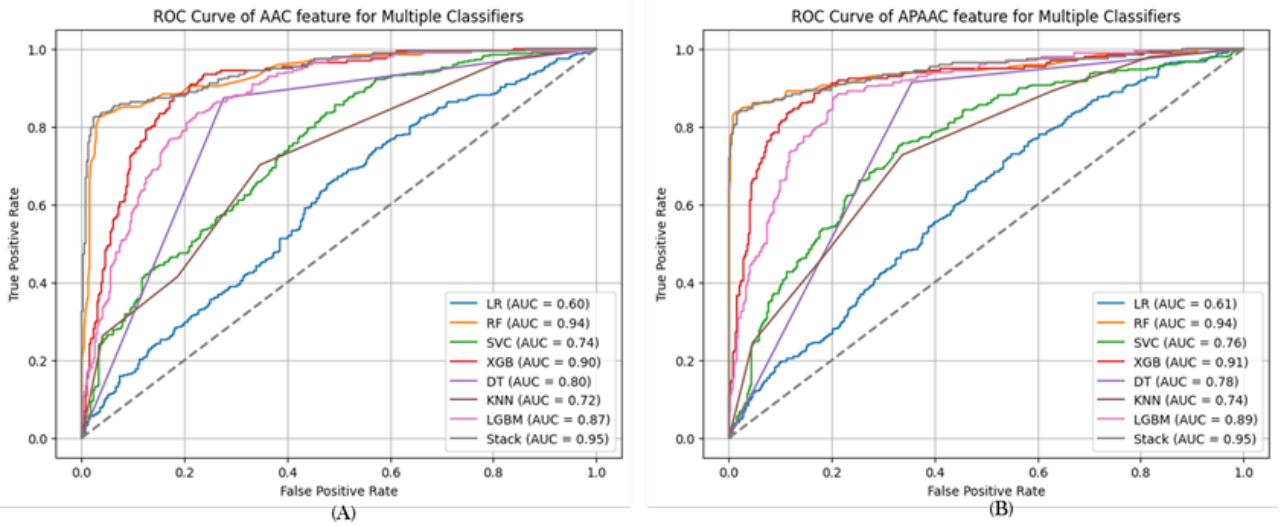

Figure 3. ROC Curve (A) AAC Feature (B) APAAC Feature

This study was directed towards the prediction of IL-4 producing peptides from an imbalanced AAC features. Performance comparison of several classifiers made, and the results are shown in Table 3. Of the models compared, the Stacking Classifier was the best, with accuracy of 88.07%, MCC of 0.7628, AUC of 0.8815, sensitivity of 90.58%, and specificity of 85.71%. These results refer to the model's strong ability to combine several base learners and improve predictive accuracy overall. In comparison, the Logistic Regression (LR) model was worst, achieving only 60.46% accuracy, MCC of 0.2099, AUC of 0.5995, 44.32% sensitivity, and 75.58% specificity, showing that LR is less appropriate for the task at hand.

This comparison underscores the valuable role of advanced ensemble methods like Stacking in handling complex biological data and enhancing the precision of IL-4 peptide prediction. After that, the classifiers were tested on the balanced dataset and the outcomes are shown in Table 4. The Stacking Classifier performed best among them with 89.59% accuracy, 0.7919 MCC, 0.8960 AUC, 90.05% sensitivity, and 89.14% specificity. In contrast, the Logistic Regression model also performed the worst here with an accuracy of 56.85%, MCC of 0.1370, AUC of 0.5685, sensitivity of 56.38%, and specificity of 57.32%. This clearly indicates that LR is not very useful for this feature. These findings also indicate the dominance of ensemble approaches like Stacking in handling complex biological datasets and improving predictive capability for IL-4 producing peptides.

We tested the classifiers on the balanced APAAC dataset, as shown in Table 5. The Stacking Classifier again gave improved results with accuracy of 89.47%, MCC of 0.7903, AUC of 0.8948, sensitivity of 91.84%, and specificity of 87.12%. Logistic Regression remained the poorest performing model with accuracy of 57.87%, MCC of 0.1573, AUC of 0.5786, sensitivity of 56.63%, and specificity of 59.09%. The results indicate the performance of advanced ensemble techniques, such as Stacking, in improving the accuracy and correctness of the detection of IL-4 producing peptides in balanced data sets. Among different combined feature settings examined in this study, the concurrent use of AAC and APAAC features yielded the optimal outcome. For context-independent IL-4 inducing peptides prediction with the AAC+APAAC balanced dataset shows in Table VI, the Stacking Classifier was better with 89.97% accuracy, 0.7995 MCC. AUC of 0.8997, sensitivity of 0.9018, and specificity of 0.8977, indicating that it is excellent at aggregating the predictions of multiple base models. The poorest performance was by the Logistic Regression (LR) model with

56.47% accuracy, an MCC of 0.1293, an AUC of 0.5647, sensitivity of 0.5718, and specificity of 0.5575.

The findings of this study validate that ensemble learning, in the form of stacking, provides a material performance advantage in IL-4-inducing peptides prediction. Comparison revealed that the stacking model performed better than each individual base learner on all feature types and dataset variants. In case of combined feature set (AAC + APAAC), the stacking model ranked as high as standalone testing accuracy of 89.97% and MCC of 0.7995 and was robust. When comparing performances between two situations of imbalanced and balanced datasets, it was evident that balancing the data with ADASYN substantially improved MCC as well as sensitivity, most importantly for minority class prediction. One of the advantages of the stacking model is that it can handle linear and nonlinear patterns using varied classifiers. This does, however, come at a computational complexity cost along with the risk of overfitting when the meta-learner is not well adjusted. The practical applications of our research reach as far as vaccine design, allergy research, and immunotherapy development, where it is crucial to accurately identify cytokine-inducing peptides. This model can assist researchers in pre-selecting candidates for wet-lab validation, thus streamlining experimental expenses and lead times. On a policy level, adoption of explainable and balanced machine learning pipelines in the study of peptides must be encouraged in biomedical research protocols. Granting bodies for immunoinformatics studies can incorporate feature transparency and fairness-aware data handling within grant proposal guidelines. Although the results are encouraging, there are limitations. The use of synthetic data generation might impose biases, and the model's performance on unseen peptides or noisy data is to be evaluated. Also, since experimentally verified IL-4-inducing peptides are limited in number, the dataset size limits the potential of deep learning models to excel beyond what classical ensembles can offer in bigger datasets. To provide the comparative performance graph, a bar diagram displays in figure 4 where we can see that the Stacking model's accuracy, MCC, AUC, sensitivity, and specificity over combined, imbalanced, and balanced datasets.

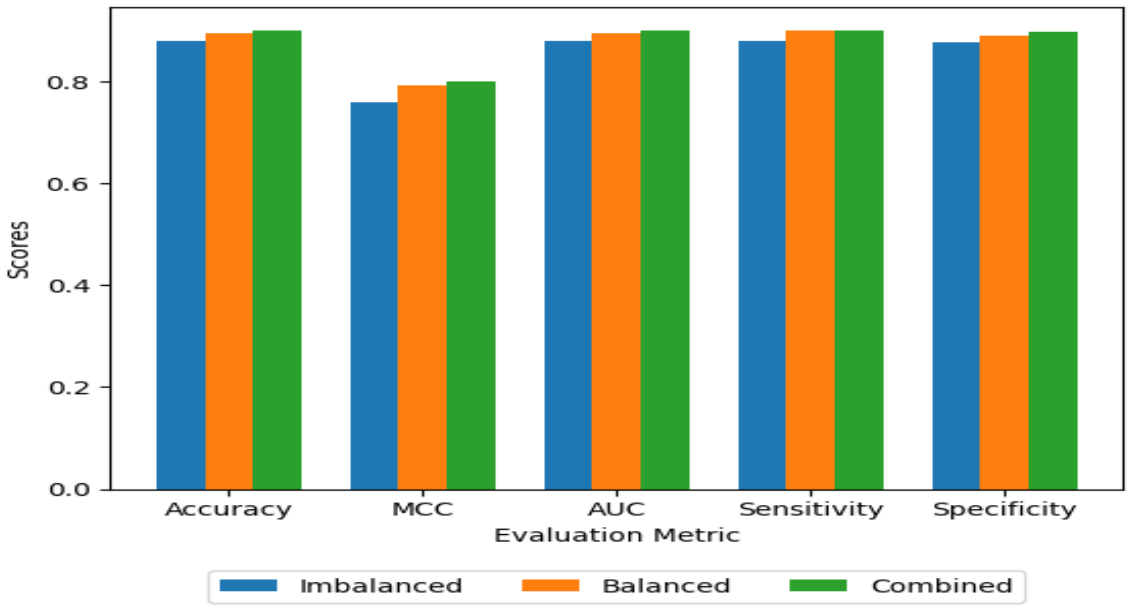

Figure 4. Comparison of Scores for Imbalanced, Balanced and Combined Feature

5. Conclusion

In this work, we have suggested a stacking ensemble-based prediction model for the prediction of IL-4-inducing peptides using extracted peptide features (AAC, APAAC, and their combination) and SHAP-guided feature selection. Our result suggests that the combination of interpretable feature engineering, effective sampling methods, and heterogeneous ensemble model produces more accurate, MCC, AUC, and overall better robustness. The stacking model, for example, handled both the imbalanced and balanced cases strongly and outperformed other machine learning algorithms across all types of features. This model not only illustrates academic significance but is also of promise to be applied in such real-world contexts as individualized immunotherapy, vaccine development, and immune disease diagnosis. By placing in the foreground, the predictive power of SHAP-selected features and balance strategies, this study makes a major contribution to the design of explainable and generalizable peptide prediction models. In the future, we would like to

build on this work by including more advanced embedding-based features, e.g., from protein language models. In addition, employing ensemble models based on deep learning on larger, experimentally verified datasets could also contribute to additional prediction improvements. Addition of uncertainty estimation and model calibration techniques could further enhance model trustworthiness in clinical or pharmaceutical decision-making scenarios. Overall, our work lays the foundation for more interpretable, accurate, and generalizable peptide prediction architectures in computational immunology.

References

[1] Sharma A, Rudra D. Emerging functions of regulatory T cells in tissue homeostasis. Frontiers in immunology. 2018 Apr 25;9:883.

[2] Bernstein ZJ, Shenoy A, Chen A, Heller NM, Spangler JB. Engineering the IL-4/IL-13 axis for targeted immune modulation. Immunological reviews. 2023 Nov;320(1):29-57.

[3] Romagnani S. Type 1 T helper and type 2 T helper cells: functions, regulation and role in protection and disease. International Journal of Clinical and Laboratory Research. 1992 Jun;21(2):152-8.

[4] Brown MA, Hural J. Functions of IL-4 and control of its expression. Critical Reviews™ in Immunology. 2017;37(2-6).

[5] Simbirtsev AS. Cytokines and their role in immune pathogenesis of allergy. Russian Medical Inquiry. 2021;5(1):32-7.

[6] León B. A model of Th2 differentiation based on polarizing cytokine repression. Trends in immunology. 2023 Jun 1;44(6):399-407.

[7] Accogli T, Bruchard M, Végran F. Modulation of CD4 T cell response according to tumor cytokine microenvironment. Cancers. 2021 Jan 20;13(3):373.

[8] Chakraborty AK. A perspective on the role of computational models in immunology. Annual review of immunology. 2017 Apr 26;35(1):403-39.

[9] Arif M, Ahmed S, Ge F, Kabir M, Khan YD, Yu DJ, Thafar M. StackACPred: Prediction of anticancer peptides by integrating optimized multiple feature descriptors with stacked ensemble approach. Chemometrics and Intelligent Laboratory Systems. 2022 Jan 15;220:104458.

[10] Riccio J, Presotto L, Doniza L, Inverso D, Nevo U, Chirico G. Predictive modeling and experimental control of macrophage pro-inflammatory dynamics.

[11] Farooq S, Khurshid J, Nazeer I. Targeted Immunization: Application of Machine Learning in Prediction of IL-4 Inducing Peptides. InComputational Techniques for Biological Sequence Analysis 2025 Jun 17 (pp. 148-170). CRC Press.

[12] Yetgin A. Revolutionizing multi-omics analysis with artificial intelligence and data processing. Quantitative Biology. 2025 Sep;13(3):e70002.

[13] Zhou X, Liu G, Cao S, Lv J. Deep Learning for Antimicrobial Peptides: Computational Models and Databases. Journal of Chemical Information and Modeling. 2025 Feb 10;65(4):1708-17.

[14] Musaazi IG, Liu L, Shaw A, Zaniolo M, Stadler LB, Vela JD. Optimizing models for the prediction of one step ahead extreme flows to wastewater treatment plants using different synthetic sampling methods. Journal of Environmental Management. 2025 Sep 1;392:126592.

[15] Xie C, Wei Y, Luo X, Yang H, Lai H, Dao F, Feng J, Lv H. NeXtMD: a new generation of machine learning and deep learning stacked hybrid framework for accurate identification of anti-inflammatory peptides. BMC biology. 2025 Jul 15;23(1):212.

[16] Miller B, de Souza EV, Pai VJ, Kim H, Vaughan JM, Lau CJ, Diedrich JK, Saghatelian A. ShortStop: a machine learning framework for microprotein discovery. BMC Methods. 2025 Aug 1;2(1):16.

[17] Tantoh DM, Yu JC, Chien CH, Yeh WY, Chu YW. Ubigo-X: Protein ubiquitination site prediction using ensemble learning with image-based feature representation and weighted voting. Computational and Structural Biotechnology Journal. 2025 Jul 14.

[18] Attanasio S, Kwasigroch J, Rooman M, Pucci F. SOuLMuSiC, a novel tool for predicting the impact of mutations on protein solubility. Scientific Reports. 2025 Jul 29;15(1):27531.

[19] Kao HJ, Weng TH, Chen CH, Yu CL, Chen YC, Huang CC, Huang KY, Weng SL. iDNS3IP: Identification and Characterization of HCV NS3 Protease Inhibitory Peptides. International Journal of Molecular Sciences. 2025 Jun 3;26(11):5356.

[20] Dholaniya PS, Rizvi S. Effect of various sequence descriptors in predicting human proteinprotein interactions using ANN-based prediction models. Current Bioinformatics. 2021 Oct 1;16(8):1024-33.

[21] Ullah F, Salam A, Nadeem M, Amin F, AlSalman H, Abrar M, Alfakih T. Extended dipeptide composition framework for accurate identification of anticancer peptides. Scientific Reports. 2024 Jul 29;14(1):17381.

[22] Esmaeili M, Mohabatkar H, Mohsenzadeh S. Using the concept of Chou's pseudo amino acid composition for risk type prediction of human papillomaviruses. Journal of theoretical biology. 2010 Mar 21;263(2):203-9.

[23] Badruzzaman Biplob KB, Sammak MH, Bitto AK, Mahmud I. COVID-19 and Suicide Tendency: Prediction and Risk Factor Analysis Using Machine Learning and Explainable AI. EAI Endorsed Transactions on Pervasive Health & Technology. 2024 Jan 1;10(1).

[24] Bitto, AK, Karim, R., Begum, MH, Khan, MFIK., Hassan, M M, & Masum, AKM. Explainable AI based deep ensemble convolutional learning for multi-categorical ocular disease prediction. EAI Endorsed Transactions on AI and Robotics, 4, Jul. 2025.

[25] Masum, AKM., Bitto, AK, Talukder, SI, Khan, MFI., Alam, MS, & Uddin, KMM. An explainable AI based deep ensemble transformer framework for gastrointestinal disease prediction from endoscopic images. EAI Endorsed Transactions on AI and Robotics, 4, Aug. 2025.

[26] Masum AKM, Khan MF, Hassan MM, Farid DM, Bitto AK, Rahman MA. Multi-Model Ensemble Approach for Accurate Classification of Ocular Disorders. In2025 International Conference on Quantum Photonics, Artificial Intelligence, and Networking (QPAIN) 2025 Jul 31 (pp. 1-6). IEEE.

[27] Masum AKM, Khan MF, Hassan MM, Bitto AK, Farid DM, Rahman T. Enhancing Dengue Fever Diagnosis: A Machine Learning Framework with Stacking Ensemble and SHAP Explainability. In2025 International Conference on Quantum Photonics, Artificial Intelligence, and Networking (QPAIN) 2025 Jul 31 (pp. 1-6). IEEE.