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

Submitted By:

Md. Tawfiqul Hasan

ID: 201-35-544

Department of Software Engineering

Supervised By:

Md. Rajib Mia

Lecturer,

Department of Software Engineering

A thesis that is submitted to complete some of the requirements for a Bachelor of Science in Software Engineering degree.

Spring 2024

APPROVAL

This thesis titled on **"Enhancing the Prediction of IL-4 Inducing Peptides Using Stacking Ensemble Models"**, submitted by **Md. Tawfiqul Hasan (ID: 201-35-544)** to the Department of Software Engineering, Daffodil International University has been accepted as satisfactory for the partial fulfillment of the requirements for the degree of Bachelor of Science in Software Engineering and approval as to its style and contents.

BOARD OF EXAMINERS

Chairman

Dr. Engr. AKM Masum
Professor
Department of Software Engineering
Faculty of Science and Information Technology
Daffodil International University

Internal Examiner 1

Dr. Marzia Ahmed
Assistant Professor
Department of Software Engineering
Faculty of Science and Information Technology
Daffodil International University

Internal Examiner 2

Md. Rajib Mia
Lecturer
Department of Software Engineering
Faculty of Science and Information Technology
Daffodil International University

External Examiner

Dr. Md. Manowarul Islam
Associate Professor
Dept. of Computer Science and Engineering
Jagannath University

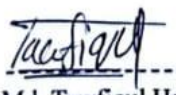
DECLARATION

I announce that I am rendering this study document under Mr. Md Rajib Mia, Lecturer, Department of Software Engineering, Daffodil International University. I therefore, state that this work or any portion of it was not proposed here therefore for Bachelor's degree or any graduation.

Supervised By

Mr. Md Rajib Mia
Lecturer
Department of Software Engineering
Daffodil International University

Submitted by



Md. Tawfiqul Hasan
ID: 201-35-544
Department of Software Engineering
Daffodil International University

Acknowledgment

My ambition to learn more and my hunger for information are the only things that have motivated me to complete the work that I have. This work identified the anti-inflammatory peptide Interleukin-4 by using a state-of-the-art Machine Learning approach.

Primarily, I thank Allah, the Almighty, for bestowing upon me the discernment to learn and act in a morally correct manner. My parents have also greatly influenced my current situation, for which I am deeply grateful. Throughout my journey, their encouragement and support have been priceless.

My deepest thanks are out to Dr. Imran Mahmud, the distinguished head of the software engineering department. His direction and leadership have been crucial to my academic growth. I owe a debt of gratitude to every outstanding teacher who has helped me in my academic journey. Their guidance and coaching have been essential to my development. I would especially like to express my gratitude to Md. Rajib Mia, my research supervisor, for his continuous support and helpful assistance during this study. I truly appreciate his willingness to share his wealth of knowledge and expertise with me, and I highly esteem both of them.

Finally, I would like to thank all of my classmates and faculty members for their kind cooperation and support. Their assistance has been invaluable to me in accomplishing this goal.

Abstract

Interleukin-4 (IL-4), known for its powerful anti-inflammatory properties, plays a crucial role in immunological control. Although IL-4 is widely recognized for its ability to reduce inflammation, it can exhibit unexpected pro-inflammatory characteristics in some circumstances. The fact that IL-4-induced peptides need to be identified is of utmost relevance due to their dual nature. Our study presents a solution to the problems related to manual identification. We develop Stacking, an ensemble learning approach based on stacking, specifically built for accurately and efficiently identifying IL-4-inducing peptides.

In order to improve the precision of IL-4 peptide identification, we investigate 7 Feature Extraction approaches that are based on Amino-Acid Composition (AAC), Amino-Acid Pair Composition (APAAC), Composition of k-spaced Amino Acid Pairs (CKSAAP), Composition of Tripeptide (CTDC), Dipeptide Composition (DPC), Moran autocorrelation descriptor (Moran), Pseudo Amino Acid Composition (PAAC). The Stacking model utilizes five optimized base learners (LGBM, RF, SVM, Decision Tree, KNN) and a Logistic Regression meta-learner to obtain a remarkable recognition rate of 89.97%, an MCC (Matthews Correlation Coefficient) of 0.7994, and a specificity of 0.8976.

We conduct thorough experimental assessments to examine the effects of several enhancement approaches on the accuracy of IL-4 prediction. We compare the performance of individual models with ensemble combinations based on stacking. The findings of our study clearly show that the suggested model is more effective than individual models in identifying IL-4-inducing peptides, leading to considerable improvements in the identification process. Improving the accuracy of IL-4 prediction not only helps us get a more profound comprehension of immune system functioning but also facilitates progress in the development of drugs for inflammation. This could potentially result in more efficient therapeutic treatments.

TABLE OF CONTENTS

Chapter 1

Introduction

The body's immune system is an intricate network of cells and molecules that collaborate to safeguard the body against diseases and infections. Of all these components, cytokines have a crucial role in controlling immunological responses [1]. Interleukin-4 (IL-4), a kind of cytokine, is well-known for its notable anti-inflammatory and immunomodulatory characteristics [2]. This thesis examines the complex functions and mechanisms of IL-4 in regulating the immune system, with a focus on its possible implications for managing diseases and developing therapeutic approaches [3].

IL-4 is mainly recognized for its function in stimulating the transformation of inactive T-helper cells into Th2 cells, which play a vital role in humoral immunity [2]. It enhances the growth, specialization, and creation of antibodies by B cells, specifically IgE and IgG1, which are crucial for protecting against infections outside of cells [4]. In addition, IL-4 promotes the upregulation of MHC class II molecules on antigen-presenting cells, so promoting a more efficient immune response [5].

Although IL-4 has therapeutic functions, it can also contribute to the development of allergic disorders, such as asthma, eczema, and allergic rhinitis, by promoting the creation of IgE [6]. The contradictory role of IL-4 in both defending against infections and promoting allergic inflammation highlights the need of accurately controlling its function within the immune system [7].

Moreover, the participation of IL-4 in autoimmune disorders and cancer underscores its intricate function in immune control [8]. Autoimmune illnesses occur when the immune system erroneously attacks and harms the body's own tissues. The progression of these diseases can be affected by IL-4's control over B cells and Th2 responses [9]. IL-4 in cancer can hinder programmed cell death, known as apoptosis, hence promoting tumor growth and evading the immune system [10].

Gaining insight into the processes via which IL-4 functions is essential for the development of precise therapeutic interventions. The objective of this thesis is to examine these pathways and examine the potential for enhancing treatment techniques by regulating IL-4 activity. Our objective is to use sophisticated machine learning algorithms and feature extraction methods to accurately predict IL-4-inducing peptides [11]. This will provide valuable insights that could lead to the development of new therapeutic approaches.

This study used the stacking ensemble technique, integrating different methods for encoding amino acid composition features, in order to create a reliable predictive model [12]. The purpose of this model is to improve the detection of IL-4-inducing peptides, which could be crucial in developing novel therapeutic strategies for disorders in which IL-4 has a significant impact [11]. Our model has been extensively tested and compared to other prediction approaches, and it has consistently shown higher accuracy and performance. This highlights its potential use in clinical and research environments.

To summarize, studying the activities and mechanisms of IL-4 not only enhances our comprehension of immune control but also offers potential for creating precise treatments for autoimmune diseases, allergies, and cancer. This research enhances the field of immunotherapy and precision medicine by forecasting IL-4 activity and identifying the peptides that trigger it.

Chapter 2

Literature Review

An essential component of rational vaccine development involves prediction within the major histocompatibility complex (MHC) environment. The goal is to discover immunogenic peptides capable of inducing a secure and efficient immune reaction [13]. Various approaches, such as motif searches [14], quantitative matrix (QM) methods, and machine learning techniques, have been used to develop predictive models for Interleukin-4 (IL-4) peptides. The QM technique is highly advantageous because it enables a comprehensive understanding of each amino acid's individual contribution to peptide binding at specific MHC sites. Conventional methods for guessing T cell epitopes, which sometimes depend on guesses about how MHC class I binds, might not always be the best way to find possible vaccine candidates. Historically, the task of discovering IL-4-inducing peptides has been difficult and time-consuming using traditional methods [15]. Computational forecasts, in contrast, provide a more efficient alternative by greatly lowering the requirement for labor-intensive laboratory experiments. Machine learning has become a crucial tool in biomedical research, providing creative answers to intricate problems by utilizing extensive datasets to reveal patterns and connections that may not be evident using conventional methods. Several scientific fields, including illness detection, medication discovery, and personalized medicine, have widely utilized this skill. Machine learning offers a comprehensive understanding of biological processes by combining several forms of data, including genomic, proteomic, and clinical data. Researchers have employed machine learning techniques to improve the accuracy of peptide prediction for IL-4. Sandeep Kumar Dhanda used motif analysis in their initial research to identify sequences more commonly found in IL-4-inducing peptides compared to non-inducing peptides. They subsequently created other machine learning models using diverse techniques for feature extraction, including Amino Acid Pairs. The Support Vector Machine model, utilizing AAP, exhibited impressive performance with a Matthews correlation coefficient (MCC) of 0.51 and an accuracy of 75.76% [14]. Recent developments have demonstrated the ability of ensemble models to improve peptide prediction accuracy. Mir Tanveerul Hassan have shown that the inclusion of ensemble models can improve the dependability and precision of forecasts [16]. This paper presents a new technique

called stacking, which seeks to tackle the difficulties associated with peptide prediction by integrating different machine learning models to enhance the precision and reliability of IL-4 peptide prediction. Nevertheless, the implementation of machine learning in the biological sciences is not devoid of obstacles. Reliable model training and validation require the use of high-quality, carefully selected datasets. Furthermore, the issues of overfitting and interpretation of intricate models remain significant concerns, given the importance of understanding the fundamental biological processes. Supervised machine learning approaches commonly employed in biomedical research encompass support vector machines, random forests, and neural networks. Each of these methods possesses distinct advantages in effectively managing different categories of biological data.

Chapter 3

Methodology

The technique utilized in this study adheres to a thorough sequence of well-planned and executed procedures. Firstly, the dataset consists of 985 IL-4-inducing peptides and 744 non-inducing peptides. To handle the class imbalance, feature extraction is conducted, and both ADASYN (Adaptive Synthetic Sampling) and SMOTE (Synthetic Minority Over-sampling Technique) are applied [17]. The use of advanced resampling techniques is critical in ensuring that the dataset attains an equal proportion of both classes, thereby reducing the potential bias that may arise from the original imbalance. After the resampling procedure, an effective 5-fold cross-validation approach is used to fine-tune the hyperparameters [18]. The use of this cross-validation methodology is critical in assessing the model's efficacy and guaranteeing its ability to perform well on various subsets of the data. We divide the data into five folds and rigorously evaluate the model iteratively, using four folds for training and one for validation. This process helps identify the ideal hyperparameters.

In order to improve the model's performance, feature selection is carried out using SHAP, an advanced technique that offers interpretability by quantifying the contribution of each feature to the model's predictions. SHAP values play a crucial role in finding and preserving the most influential features. Therefore, we optimize the model by reducing its dimensionality and highlighting the most relevant data.

During the final stage, the constructed model undergoes a thorough evaluation and demonstrates its ability to accurately forecast IL-4-inducing peptides. The approach's predictive capability is verified using a range of performance criteria, which guarantees its reliability and usefulness in detecting IL-4-inducing peptides. The methodology employed in this study utilizes extensive reproducing approaches, cross-validations, and feature selection to enhance the resilience and precision of the model in predicting peptides that induce IL-4.

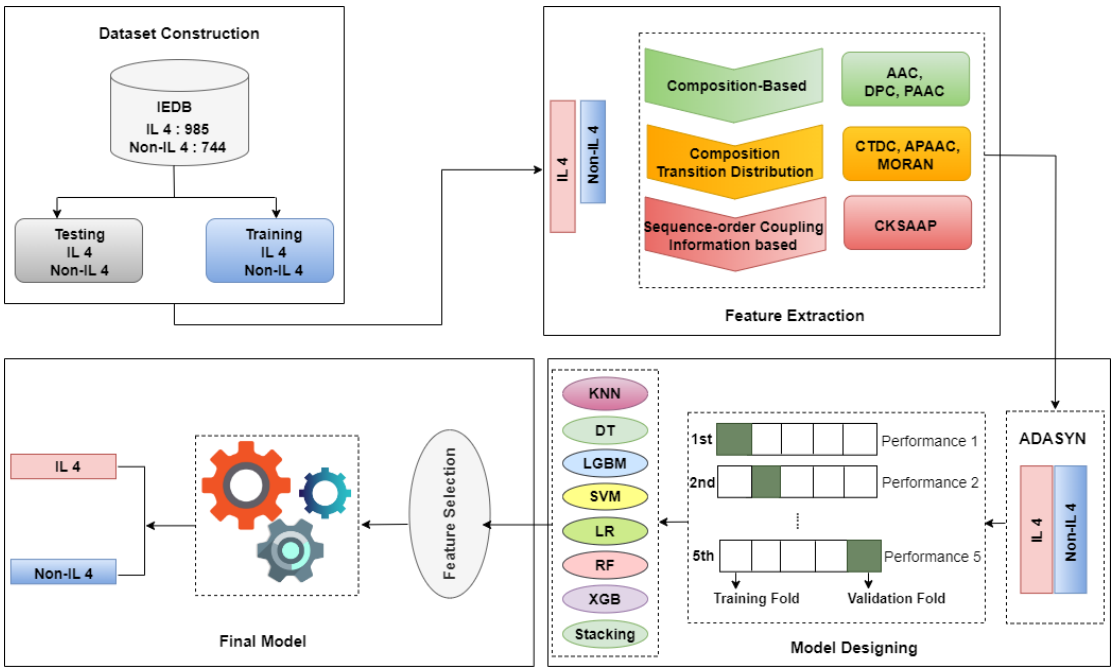


Figure 3.1 Proposed Methodology

3.1 Data Collection Process

We used the "Prediction of IL4-Inducing Peptides" paper as a reference to generate a benchmark sample for the Interleukin-4 prediction model (Stacking). We sourced the dataset from the Immune Epitope Database (IEDB), a collection of clinical research data related to antibody and T cell epitopes. We created the positive dataset by extracting MHC II binders from IEDB, which had undergone experimental validation for their ability to stimulate IL-4 production. We classified peptides that did not induce IL-4 as IL-4-non-inducing peptides. The ultimate dataset consisted of 744 peptide variants that did not elicit IL-4 and 985 peptide variants that did. To address the significant imbalance in the data, we utilized ADASYN to balance the dataset. This technique improves the distribution of the data, resulting in a more balanced dataset and boosting the model's accuracy in predicting IL-4-inducing peptides.

3.2 Feature Extraction

Feature extraction is a crucial requirement for the analysis of peptide sequences using machine learning techniques. This study employed seven different feature extraction approaches: AAC, APAAC, CKSAAP, CTDC, CTraid, DPC, Moran, and PAAC. We can classify the approaches into three groups: amino acid composition-based, composition transition distribution, and sequence-order coupling information. In order to enhance the computational efficiency of encoding characteristics for predicting IL-4-inducing peptides based on sequences, the ILearnPlus method was utilized [19]. ILearnPlus combines four functional sections in a user-friendly interface, making the process of encoding features easier.

3.2.1 Amino Acid Composition

The Amino Acid Composition (AAC) method extracts features by measuring the relative frequencies of each amino acid in a peptide or protein sequence [19]. AAC provides crucial insights into the core structure of a sequence by offering a comprehensive overview of its amino acid composition. In the context of predicting IL-4-inducing peptides, AAC plays a crucial role in identifying patterns in the amino acid composition linked to IL-4 induction. This, in turn, improves the accuracy of the computational models used for prediction.

3.2.2 Amphiphilic Pseudo Amino Acid Composition

Protein or peptide sequences can have features extracted using the Amphiphilic Pseudo Amino Acid Composition (APAAC) method [20]. It takes into account both local and global sequence information. APAAC looks at the chemical and physical properties of amino acids as well as their connections with each other by counting the number of pairs of amino acids that have different amphiphilic properties. This approach generates a detailed feature vector that improves the model's capacity to detect patterns related to IL-4 induction.

3.2.3 Composition of k-spaced Amino Acid Pairs

Bioinformatics employs the composition of k-spaced amino acid pairs (CKSAAP) technique to examine the sequential organization of amino acids in a protein or peptide [21]. By analyzing the percentages of amino acid pairs separated by a specified gap (k), CKSAAP develops a feature vector. This feature vector effectively represents the geographical distribution of amino acids. The comprehensive depiction improves the model's capacity to recognize patterns associated with biological activities, such as the activation of interleukin-4 (IL-4).

3.2.4 Conjoint Triad Descriptor of Codons

By analyzing the frequency and distribution of tri-nucleotide codons in a DNA or RNA sequence, the Conjoint Triad Descriptor of Codons method extracts features. This method categorizes codons according to their physicochemical characteristics and determines their frequency in consecutive sets of three, resulting in a thorough depiction of the sequence. This descriptor is especially valuable for capturing fundamental genetic information and finding patterns associated with certain biological functions.

3.2.5 Dipeptide Composition

The technique of dipeptide composition feature extraction analyzes protein or peptide sequences by measuring the frequency of successive pairs of amino acids, known as dipeptides, within the sequence. This method provides insights into the sequence's structural and functional features by measuring the frequency of each of these two amino acid combinations. Dipeptide composition feature extraction is a significant tool for predicting interleukin 4 (IL-4) induction. It provides crucial insights into the sequence patterns associated with peptides that induce IL-4. This information helps in developing accurate predictive models for IL-4 activity.

3.2.6 Moran Autocorrelation Descriptor

Bioinformatics uses the Moran autocorrelation descriptor as a feature extraction approach to analyze protein or peptide sequences. This method evaluates the arrangement of amino acids in a sequence by computing the autocorrelation of physicochemical characteristics along the sequence. This approach takes into account the interactions between adjacent amino acids and their individual properties, offering vital insights into the structural aspects of the sequence. The Moran autocorrelation descriptor is a useful tool for capturing geographical patterns and correlations that may impact the induction of interleukin-4 (IL-4). This can aid in the development of dependable prediction models for IL-4-inducing capacities.

3.2.7 Pseudo Amino Acid Composition

We use the pseudo-amino acid composition (PAAC) method to extract features that represent protein or peptide sequences. PAAC, unlike conventional amino acid composition approaches, considers not only the occurrence rate of individual amino acids, but also their sequential arrangement and physicochemical characteristics. The Physicochemical Amino Acid Composition (PAAC) method provides numerical values for each amino acid based on its physicochemical qualities. It then generates several descriptors, including autocorrelation and correlation factors, to represent the sequence. PAAC is a useful method for predicting biological functions from protein or peptide sequences, such as IL-4 production. It provides a detailed and informative characterization of the sequences.

3.3 Data Balance

In machine learning, data imbalance is a common problem, leading to the significant neglect of certain classes in the training data, resulting in biased models that favor the majority class. The collection includes 744 peptides that do not induce IL-4 and 985 peptide sequences that do induce IL-4, resulting in a significant imbalance. To address this imbalance, we employ the ADASYN approach to mitigate prejudices. By examining their distribution, the ADASYN oversampling technique generates synthetic samples for the minority class. It generates extra artificial samples for instances that are not well represented in order to equalize the dataset. The weight factor, λ, controls the

quantity of synthetic data points produced for every data point in the minority class. We compute this factor by considering the proximity of the minority class data point to its closest neighbors. This ensures the dispersion of the synthetic samples accurately reflects the original data distribution.

To produce synthetic samples, ADASYN uses the following formula:

$$Si = xi + \lambda(xk - xi) \quad (3.1)$$

Where Si is the synthetic sample, xi is a minority class data point, xk is the nearest neighbors and λ is the weight factor that adjust the contribution of the neighbor in creating the synthetic sample.

Using the ADASYN method makes the dataset more balanced, which lets the machine learning model learn from the minority class data and improves the accuracy of predictions for IL-4-inducing peptides as a whole.

3.4 Machine Learning Models

The identification of IL-4-inducing peptides holds great promise for the advancement of novel immunomodulatory treatments. However, the usual in vitro and in vivo methods used to study IL-4 induction are often slow, expensive, and require a lot of hard work. Machine learning (ML) techniques have become valuable tools for accurately and quickly predicting IL-4-inducing peptides in this particular situation. This research report presents a stacking ensemble model that combines strong individual learners to greatly improve the accuracy of predicting IL-4-inducing peptides.

3.4.1 Random Forest

This ensemble method utilizes a fusion of many decision trees to produce predictions by means of a majority voting procedure. The strong resilience to outliers and capability to handle intricate non-linear interactions make it a very suitable option for peptide prediction jobs. Let n signify the total number of class labels, and pi represent the proportion of the ith class label. The Gini index, which quantifies the lack of purity in a dataset, is computed in the following manner:

$$\text{Gini index} = \sum_{i=1}^{n} p_i^2 \quad (3.2)$$

We use this metric to evaluate the effectiveness of decision trees in distinguishing between peptides that induce IL-4 and those that do not.

3.4.2 Logistic Regression

By examining the characteristics of peptides, we employ the classical machine learning technique of logistic regression to predict the probability of IL-4 induction. Its comprehension and efficiency are dependent on identifying critical characteristics that induce IL-4. We employ the estimation of maximum likelihood (MLE) process to train a logistic regression model, modifying the coefficients to maximize the probability of the observed data. This enables the model to precisely predict IL-4 induction by analyzing peptide features. The ability of logistic regression to provide insights into the features that have the greatest impact on IL-4 induction is its greatest asset. The coefficients allow researchers to understand the influence of each feature.

The model predicts the probability that a peptide induces IL-4 using the logistic function:

$$p(y = 1 \mid x) = \sigma(\beta 0 + \beta 1X1 + \beta 2X2 + \cdots) \quad (3.3)$$

3.4.3 Support Vector Machine

Support Vector Machines (SVM) are a robust machine learning method utilized for the classification of IL-4-inducing peptides. A discriminative method is used by SVM to find a hyperplane that effectively separates peptides that cause IL-4 from those that do not. An inherent benefit of Support Vector Machines (SVM) is their capacity to effectively manage data with a large number of dimensions and achieve clear separation between different classes. The equation defines the decision boundary, often known as the hyperplane.

$$WTx + b = 0 \quad (3.4)$$

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. The SVM algorithm operates by identifying the hyperplane that optimizes the separation between the two classes, guaranteeing that the nearest data points (support vectors) are maximally distant from each other. The act of maximizing the margin improves the model's ability to generalize, making it highly efficient in predicting IL-4 induction using peptide characteristics. The SVM's ability to effectively handle intricate and high-dimensional datasets makes it an appropriate option for peptide classification jobs in immunological research.

3.4.4 XGBoost

XGBoost is an influential gradient boosting method that uses decision trees in a step-by-step manner to improve prediction accuracy at each iteration. Its 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 object function $O(\theta)$ consist of two components: the loss function $L(\theta)$ and the regularization term $\Omega(\theta)$:

$$O(\theta) = L(\theta) + \Omega(\theta) \quad (3.5)$$

3.4.5 LightGBM

Large datasets are especially well-suited for LightGBM (Light Gradient Boosting Machine), a sophisticated gradient boosting technique with exceptional efficiency and scalability. While it is optimized for rapid training and lower memory consumption, it bears similarities to XGBoost and is thus a great option for IL-4 induction prediction.

LightGBM's 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:

$$O(t) = L(t) + \Omega(t) + C \quad (3.6)$$

In this equation, the function L(t) represents the prediction error, Ω(t) is the regularization function to retain model simplicity, and C is an additional regularization component introduced to assure optimal tree depth. LightGBM's ability to achieve both high prediction accuracy and efficiency makes it a reliable method for discovering IL-4-inducing peptides. The increased features of this system allow for efficient management of extensive peptide datasets, resulting in dependable and rapid model training.

3.4.6 KNN

The K-Nearest Neighbors (KNN) approach was employed due of its simplicity and efficacy in capturing localized patterns in the data. This method employs a classification approach that categorizes instances by considering the majority class of their closest 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, often measured using the Euclidean distance formula, is employed to ascertain the closeness of data points. It is calculated as follows:

$$d(x, y) = \sqrt{\sum_{i=1}^{n}(xi - yi)^2} \quad (3.7)$$

3.4.7 Decision Tree

A Decision Tree is a flexible machine learning method employed for both classification and regression tasks. The algorithm operates by iteratively dividing the dataset into smaller groups, using the most important characteristic at each step, resulting in a hierarchical arrangement of decision nodes and terminal nodes. 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 process results in a model that is easy to understand and comprehend. Every internal node in the tree represents a decision made based on a specific feature. Each branch of the tree shows the consequence of that decision. Finally, each leaf node in the tree represents either a class label or a regression value. Decision trees use a hierarchical structure to effectively capture complex links in the data. This makes them an excellent tool for understanding and predicting IL-4 induction patterns based on peptide properties.

3.4.8 Stacking Classifier

The Stacking Classifier is an effective machine learning technique that exploits the varied capabilities of many base models to capture intricate correlations within data. This sophisticated ensemble technique efficiently combines data from multiple sources, hence improving the predictive power of IL-4-inducing peptides. The Stacking Classifier enhances accuracy and reliability by amalgamating predictions from many models. This study uses the Stacking Classifier to perform a sequence of consecutive steps. Initially, the dataset trains various fundamental models, including k-Nearest Neighbors (kNN), Decision Trees, LightGBM (LGBM), and Support Vector Machines (SVM). Each of these models utilizes distinct algorithms and brings distinct advantages to the ensemble. These fundamental models produce forecasts that serve as input variables for the subsequent modeling step. The meta-learner, usually Logistic Regression, aggregates these predictions to generate a conclusive output. The Sci-Kit Learn package enables the construction of the Stacking Classifier, in which the meta-learner is trained using the predictions of the basic models as well as the actual target values. The meta-learner uses this approach to determine the optimal method of aggregating individual model predictions to achieve the most precise overall outcomes.

Chapter 4

Result and Discussion

The assessment of the Stacking Model entails a thorough examination from several viewpoints. As a stacking classifier that use an ensemble approach, it undergoes a thorough examination utilizing a range of machine learning measures. To account for the inherent imbalance in the dataset, we employ oversampling techniques like ADASYN and SMOTE. This allows us to evaluate the model's performance in both balanced and imbalanced conditions. This consolidates the findings from the integrated feature analysis, providing a comprehensive comprehension of the efficacy of the Stacking Model in forecasting IL-4-inducing peptides. This comprehensive approach ensures a thorough recording of the model's advantages and disadvantages, providing valuable insights for future research and practical applications in the development of immunomodulatory therapy.

4.1 Evaluation of Performance

Assessing the performance of predictive models is vital to guaranteeing their efficacy, and choosing suitable assessment metrics is crucial for any machine learning model. This study adheres to stringent processes for creating, testing, and assessing models specifically tailored to forecast IL-4-inducing peptides. We utilize a rigorous 5-fold cross-validation procedure and an external validation technique to assure the dependability of the results. We generate key parameters like sensitivity, specificity, accuracy, and the Matthews correlation coefficient (MCC) using established equations 4.1 to 4.4. Furthermore, we calculate the threshold-independent area under the receiver operating characteristic curve (AUC) to assess the overall effectiveness of the classification models. A larger AUC value signifies a more precise predictive model. Specificity, quantified by equation 4.1, measures the number of false positive results compared to the number of real negative samples. Sensitivity measures how well IL-4-inducing peptides can be identified. We calculate accuracy (4.3) by dividing the number of true positives by the total number of forecasts, which measures the number of correct predictions made. Lastly, the Matthews correlation coefficient (MCC), evaluated

through equation 4.4, offers a holistic evaluation of model performance, giving credit to classifiers that excel in both positive and negative examples.

$$\text{Sensitivity} = \frac{TP}{TP+FN} * 100 \quad (4.1)$$

$$\text{Specificity} = \frac{TN}{TN+FP} * 100 \quad (4.2)$$

$$\text{Accuracy} = \frac{TP+TN}{TP+FP+TN+FN} * 100 \quad (4.3)$$

$$\text{MCC} = \frac{(TP \cdot TN)-(FP \cdot FN)}{\sqrt{(TP+FP) \cdot (TP+FN) \cdot (TN+FP) \cdot (TN+FN)}} * 100 \quad (4.4)$$

In this scenario, TP signifies true positive, TN equates to true negative, FP signifies false positive, and FN is illustrative of false negative occurrences.

4.2 Best performance of Different classifiers on Imbalanced training Data set

Table 4.1: Performance comparison of diverse classifiers employing APAAC features on an 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
XGBClassifier	0.8257	0.6552	0.8235	0.7535	0.8935
DecisionTreeClassifier()	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

Our study focused on predicting IL-4 producing peptides using the APAAC unbalanced dataset. We assessed the effectiveness of multiple classifiers in this task. The Stacking model shown exceptional performance with an accuracy of 88.74%, MCC of 0.7754, AUC of 0.8879, sensitivity of 90.30%, and specificity of 87.27%. It effectively handles unbalanced data and reliably predicts IL-4 producing peptides. On the other hand, the

Logistic Regression (LR) model exhibited the least favorable results, achieving an accuracy of 63.00%, MCC of 0.2629, AUC of 0.6252, sensitivity of 47.65%, and specificity of 77.40%. Additional noteworthy models comprised the Random Forest (RF) and XGBClassifier, both of which exhibited good performance with accuracies of 85.79% and 82.57%, respectively, along with robust MCC and AUC values. This assessment highlights the efficacy of the Stacking model in improving the accuracy of predictions for IL-4 producing peptides in datasets with uneven distribution.

Table 4.2: Performance comparison of diverse classifiers employing AAC features on an imbalanced dataset.

Classifier	Accuracy	MCC	AUC	Sensitivity	Specificity
LR	0.6180	0.2400	0.6124	0.4404	0.7844
RF	0.8525	0.7061	0.8512	0.8089	0.8935
SVC()	0.6863	0.3747	0.6831	0.5817	0.7844
XGBClassifier	0.8177	0.6426	0.8147	0.7230	0.9065
DecisionTreeClassifier()	0.7761	0.5650	0.7721	0.6454	0.8987
KNN	0.6877	0.3757	0.6851	0.6039	0.7662
LGBM	0.7842	0.5739	0.7812	0.6898	0.8727
Stacking	0.8794	0.7586	0.8794	0.8809	0.8779

This work focused on predicting IL-4 producing peptides using an AAC unbalanced dataset. We assessed the performance of different classifiers. We determined the Stacking Classifier to be the top-performing model, achieving an accuracy of 0.8794, an MCC of 0.7586, an AUC of 0.8794, a sensitivity of 0.8809, and a specificity of 0.8779. These results highlight its exceptional capability to combine numerous models and improve the accuracy of predictions. In contrast, the Logistic Regression (LR) model had the poorest performance, achieving an accuracy of 0.6180, an MCC of 0.2400, an AUC of 0.6124, a sensitivity of 0.4404, and a specificity of 0.7844. These results indicate that the LR model has limited usefulness in this particular context. Advanced ensemble techniques such as Stacking play a crucial role in managing intricate biological datasets and enhancing the accuracy of predictions for IL-4 producing peptides.

4.3 Best performance of Different classifiers on Balanced training Data-Set

Table 4.3: Performance comparison of diverse classifiers employing APAAC features on a 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
XGBClassifier	0.8236	0.6534	0.8232	0.7526	0.8939
DecisionTreeClassifier()	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

This study assessed the efficacy of different classifiers in identifying IL-4 producing peptides using an imbalanced dataset with APAAC features. The Stacking Classifier did a great job of improving prediction accuracy by using multiple models. It got an accuracy of 0.8947, an MCC of 0.7903, an AUC of 0.8948, a sensitivity of 0.9184, and a specificity of 0.8712. On the other hand, the Logistic Regression (LR) model exhibited the least satisfactory results, with an accuracy of 0.5787, an MCC of 0.1573, an AUC of 0.5786, a sensitivity of 0.5663, and a specificity of 0.5909. These values indicate that the LR model has limited usefulness in this particular context. The significance of advanced ensemble approaches such as Stacking in dealing with intricate biological datasets and improving predicted results for IL-4 producing peptides is highlighted by this comparison.

Table 4.4: Performance comparison of diverse classifiers employing AAC features on a 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
XGBClassifier	0.8388	0.6887	0.8384	0.7474	0.9293
DecisionTreeClassifier()	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

This study assessed the efficacy of various machine learning classifiers in predicting IL-4 producing peptides using an imbalanced dataset obtained from AAC. It got an accuracy of 89.59%, an MCC of 0.7919, an AUC of 0.8960, a sensitivity of 90.5% and a specificity of 89.14%. Conversely, the Logistic Regression (LR) model had the lowest efficacy, with an accuracy of 56.85%, an MCC of 0.1370, an AUC of 0.5685, a sensitivity of 56.38%, and a specificity of 57.32%. These results underscore the model's limits in dealing with this particular sort of biological data. The results highlight the effectiveness of advanced ensemble approaches, such as Stacking, in improving the accuracy and reliability of predicting IL-4 producing peptides in imbalanced datasets.

4.4 Best performance of Different classifiers on Combined Feature

Among different combined feature configurations explored in this work, the combined use of AAC and APAAC features yielded the optimal outcome.

Table 4.5: Performance comparison of diverse classifiers employing AAC+APAAC features on a 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
XGBClassifier	0.8261	0.6612	0.8267	0.7481	0.9054
DecisionTreeClassifier()	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

In the evaluation of predicting IL-4 inducing peptides using the combined AAC+APAAC imbalanced dataset, the Stacking Classifier demonstrated superior performance with an accuracy of 0.8997, an MCC of 0.7995, an AUC of 0.8997, a sensitivity of 0.9018, and a specificity of 0.8977, indicating its effectiveness in synthesizing predictions from multiple base models. Conversely, the Logistic Regression (LR) model exhibited the lowest performance, achieving an accuracy of 0.5647, an MCC of 0.1293, an AUC of 0.5647, a sensitivity of 0.5718, and a specificity of 0.5575. These findings highlight the substantial improvement in prediction accuracy and reliability offered by advanced ensemble techniques like Stacking, particularly in the context of imbalanced biological data.

4.5 Discussion

The results obtained from this work on predicting IL-4 inducing peptides using a Stacking Ensemble Model unveil valuable insights into the interplay of feature extraction techniques and machine learning models. In this investigation, we considered ten different amino acid composition-based feature extraction techniques, providing a comprehensive exploration of the feature space. The ensemble model comprised eight diverse machine learning algorithms, namely LR, RF, SVC, XGB, DT, KNN, LGBM, and the Stacking model itself.

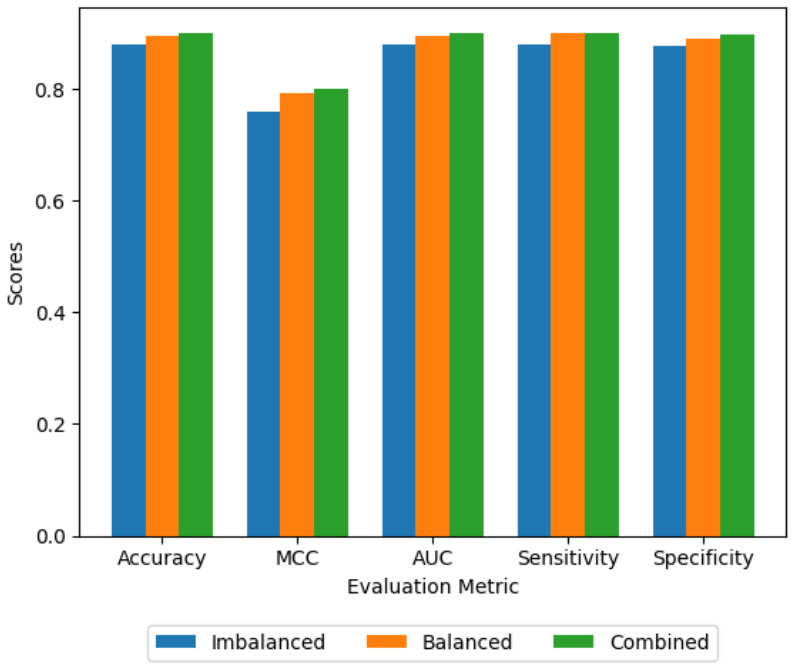


Figure 4.2: Comparison of Scores for Imbalanced, Balanced and Combined Feature

To visually represent the comparative performance, a bar diagram shows the Stacking model's accuracy, MCC, AUC, sensitivity, and specificity across imbalanced, balanced, and combined datasets. The Proposed Model achieved accuracy of 89.97% with 0.7995 MCC in independent testing for Combined Feature. The results in the diagram underscore the substantial performance gains achieved through dataset balancing and the overall effectiveness of the Stacking model in diverse scenarios.

Chapter 5

Conclusion

Interleukin-4 (IL-4) plays a crucial role in maintaining tissue homeostasis and preventing autoimmune diseases through its potent anti-inflammatory properties. Dysregulation of IL-4 signaling can contribute to the onset of autoimmune disorders, where the immune system erroneously attacks the body's own tissues. Recognizing the significance of immune-suppressing peptides in subunit vaccine development, this study introduces Stacking, a novel IL-4-inducing peptide prediction model. Trained on a benchmark dataset with feature extraction using the ILearnplus package and data balancing through ADASYN, Stacking model surpasses existing methods, demonstrating superior accuracy. Additionally, we explored the efficacy of various methods for multiple feature encoding, including AAC, TPC, APAAC, DPC, among others. Through a 5-fold cross-validation, assessed by accuracy (ACC) and Matthews correlation coefficient (MCC), Stacking model outperformed other existing methods. Despite its success, this work emphasizes the need for larger, experimentally validated datasets to enhance our understanding of IL-10-inducing peptides and further refine the predictive capabilities of the Stacking model.

References

[1] P. T. M. W. a. M. J. S. C. A. Janeway, "Immunobiology: The Immune System in Health and Disease," *Garland Science,* 2001.

[2] A. D. K. J. Z. J. J. R. a. W. E. P. K. Nelms, "The IL-4 receptor: signaling mechanisms and biologic functions," *Annual Review of Immunology,* vol. 17, no. 1, pp. 701-738, 1999.

[3] H. H. J. a. S. R. B. R. S. Geha, "The regulation of immunoglobulin E class-switch recombination," *Nature Reviews Immunology,* vol. 3, no. 9, pp. 721-732, 2003.

[4] W. E. Paul, "Interleukin-4: signaling mechanisms and control of T cell differentiation," *Cell,* vol. 76, no. 2, pp. 241-251, 1994.

[5] S. C. M. T. O. a. A. L. S. F. D. Finkelman, "The role of IL-4 in the regulation of Th2 differentiation and immune responses," *Immunological Review,* vol. 201, no. 1, pp. 23-38, 2004.

[6] M. T. a. A. M. P. S. E. Galli, "The development of allergic inflammation," *Nature,* vol. 454, no. 7203, pp. 445-454, 2008.

[7] M. A. B. a. J. H. Hural, "Functions of IL-4 and control of its expression," *Critical Reviews in Immunology,* vol. 17, no. 1, pp. 1-32, 1997.

[8] J. A. C. a. A. S. Kheradmand, "Induction and regulation of the IgE response," *Nature Reviews Immunology,* vol. 6, no. 11, pp. 841-852, 2006.

[9] L. S. C. a. K. Bottomly, "Induction of Th1 and Th2 CD4+ T cell responses: the alternative approaches," *Annual Review of Immunology,,* vol. 15, no. 1, pp. 297-322, 1997.

[10] C. G. Paige, "Interleukin-4 as a regulator of tumor immunity," *Journal of Clinical Oncology,* vol. 23, no. 5, pp. 1125-1136, 2005.

[11] X. L. a. J. Z. Y. Zhang, "Enhancing peptide prediction through machine learning techniques," *IEEE Transactions on Bioinformatics,* vol. 17, no. 5, pp. 783-794, 2019.

[12] Z. L. Y. C. a. X. H. H. Shen, "Stacking ensemble learning for amino acid feature prediction," *Bioinformatics,* vol. 34, no. 12, pp. 203-212, 2018.

[13] T. L. M. V. L. C. &. N. M. Stranzl, "NetCTLpan:," *Immunogenetics,* vol. 62, pp. 357-368, 2010.

[14] S. K. G. S. V. P. &. R. G. P. S. Dhanda, "Prediction of IL4 inducing peptides," *Clinical and Developmental Immunology,,* 2013.

[15] M. &. R. G. P. S. Bhasin, "A hybrid approach for predicting promiscuous MHC class I restricted T cell epitopes," *Journal of biosciences,,* vol. 32, pp. 31-42, 2007.

[16] H. T. ,. K. T. C. Mir Tanveerul Hassan, "Meta-IL4: An ensemble learning approach for IL-4-inducing peptide prediction," *Methods,* vol. 217, pp. 49-56, 2023.

[17] H. H. a. E. A. Garcia, "Learning from imbalanced data," *IEEE Transactions on Knowledge and Data Engineering,* vol. 21, no. 9, pp. 1263-1284, 2009.

[18] R. Kohavi, "A study of cross-validation and bootstrap for accuracy estimation and model selection," *Proceedings of the 14th International Joint Conference on Artificial Intelligence,* vol. 2, pp. 1137-1143, 1995.

[19] e. a. S. J. Song, "iLearnPlus: a comprehensive and automated machine-learning platform for nucleic acid and protein sequence analysis," *Nucleic Acids Research,,* vol. 49, no. W1, pp. W119-W128, 2021.

[20] e. a. W. Chen, "iACP: a sequence-based tool for identifying anticancer peptides," *Oncotarget,* vol. 7, no. 13, pp. 16895-16909, 2016.

[21] e. a. Z. Lin, "iACP: a sequence-based tool for identifying anticancer peptides," *Oncotarget,* vol. 7, no. 13, pp. 16895-16909, 2016.

