A stacking model for prediction of IL-6 inducing peptides using machine learning techniques

Submitted By

Mir Maruf Hossain Emon
ID: 201-35-631
Department of Software Engineering

Supervised By

Dr. Engr. AKM Masum
Professor,
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 "**A stacking model for prediction of IL-6 inducing peptides using machine learning techniques**", submitted by **Mir Maruf Hossain Emon (ID: 201-35-631)** 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

10/7/24

Chairman

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

10/7/24

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

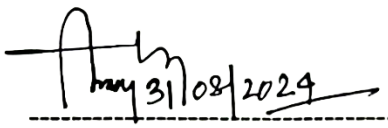
External Examiner

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

DECLARATION

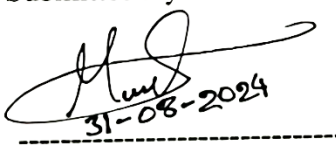
I announce that I am rendering this study document under **Dr. Engr. AKM Masum**, Professor, Department of Software Engineering, Daffodil International University. I therefore state that this work or any portion of it was not proposed here therefore for Bachelor's degree or any graduation.

Supervised By



Dr. Engr. AKM Masum
Professor
Department of Software Engineering
Daffodil International University

Submitted By



Mir Maruf Hossain Emon
ID: 201-35-631
Department of Software Engineering
Daffodil International University

ACKNOWLEDGMENT

The results of this study bear witness to my unwavering quest for knowledge and my enthusiasm to advance my education. This work uses a cutting-edge machine learning method to pinpoint Interleukin-6, a crucial component of the inflammatory response. First of all, I want to express my sincere thankfulness to Allah, the Almighty, for giving me the discernment to pursue knowledge and the moral sense to act morally. I also owe a debt of gratitude to my parents for helping to shape my accomplishments to date. For his guidance and assistance, I am truly appreciative of **Associate Prof. Dr. Imran Mahmud**, the distinguished Head of the Department of Software Engineering. I want to sincerely thank all of the wonderful teachers who have helped me along the way in my schooling. Their guidance and teaching have been priceless. I want to express my sincere gratitude to **Professor Dr. Engr. AKM Masum**, who supervised my research, for his invaluable advice and steadfast assistance during this endeavor. His depth of experience and knowledge have been invaluable, and I am sincerely appreciative of his willingness to impart them to me.

Finally, I would like to express my gratitude to DIU's members and my fellow classmates for their kind cooperation and support, both of which have been crucial to the accomplishment of this goal.

Abstract

Prominent for its strong pro-inflammatory properties, interleukin-6 (IL-6) is an important immunological regulator. Even though IL-6 is known to promote inflammation, in certain situations it can exhibit surprising anti-inflammatory properties. This dual character emphasizes how crucial it is to identify the peptides generated by IL-6. In order to overcome the drawbacks of manual identification, which are often expensive, our work presents Stacking, an ensemble learning approach based on stacking that is intended to provide accurate and effective IL-6-inducing peptide identification. Ten Amino-Acid-Composition-based Feature Extraction techniques are examined, including AAC, APAAC, CKSAAP, CTDC, DPC, Moran and PAAC. With the help of a Logistic Regression meta-learner and eight improved base learners (LGBM, RF, SVM, Decision Tree, XGBClassifier, LR and KNN), the Stacking model achieves an excellent 97.10% identification rate, 0.9433 MCC, and 0.9766 specificity. We examine the effects of several enhancement strategies on the accuracy of IL-6 predictions through experimental evaluations, contrasting single models with ensemble combinations based on stacking. Our findings show that the suggested methodology is far more effective than its individual equivalents in identifying peptides that induce IL-6. Improving IL-6 prediction advances the search for anti-inflammatory drugs.

Table of Contents

Abstract ____ *v*

Chapter 1 ____ *1*

Introduction ____ **1**

1.1 Introduction ____ 1
1.2 Background ____ 2
1.3 Problem Statement ____ 2
1.4 Research Questions ____ 3
1.6 Research Gap ____ 4
1.7 Research Objectives ____ 4

Chapter 2 ____ *5*

Literature review ____ **5**

Chapter 3 ____ *7*

Methodology ____ **7**

3.1 Data Collection Process ____ 7
3.2 Extraction of Features ____ 8
3.2.1 AAC ____ 8
3.2.2 APAAC ____ 8
3.2.3 CKSAAP ____ 9
3.2.4 CTDC ____ 9
3.2.5 DPC ____ 9
3.2.6 Moran ____ 10
3.2.7 PAAC ____ 10
3.3 Data Balance ____ 10
3.4 Models for Machine Learning ____ 11
3.4.1 Random Forest ____ 11
3.4.2 Analysis of Logistic Regression ____ 11
3.4.3 The Support Vector Machine ____ 12
3.4.4 XGBoost ____ 12
3.4.5 LightGBM ____ 12
3.4.6 KNN ____ 13
3.4.7 Decision Tree ____ 13
3.4.8 Stacking Classifier ____ 13

Chapter 4 ____ *15*

Result and Discussion ____ **15**

4.1 Performance Evaluation ____ 15
4.3 Performance of various classifiers using AAC ____ 17
4.4 Optimal outcomes of Various classifiers ____ 17
4.5 Performance comparison of several classifiers ____ 18
4.6 ROC Curve Performance Analysis of Best Three Feature ____ 19
4.7 Discussion ____ 20

Chapter 5 ____ *21*

Conclusion ____ **21**

References ____ *23*

Chapter 1

Introduction

1.1 Introduction

A versatile cytokine, interleukin-6 (IL-6), is essential to the immune system. IL-6 is a little protein that controls hematopoiesis, inflammation, and immunological responses. One of IL-6's main roles is to control the immune system's response by encouraging the synthesis of acute-phase proteins and B cells' maturation into plasma cells, which, frequently with the help of T cells, produce antibodies. Moreover, IL-6 affects T cell development and proliferation, assisting in the activation of macrophages and cytotoxic T cells, which improves the body's capacity to eradicate intracellular infections. Additionally, IL-6 has a major effect on the inflammatory response [1]. It plays a crucial role in the synthesis and control of several pro-inflammatory cytokines, including interleukin-1 (IL-1), interferon-gamma (IFN-γ), and tumor necrosis factor-alpha (TNF-α) [2] . IL-6 plays a crucial role in immunological homeostasis by regulating these inflammatory mediators, which can either intensify or decrease inflammation depending on the situation. On the other hand, dysregulation of IL-6 levels is linked to a variety of illnesses. Increased levels of IL-6 have been linked to autoimmune disorders, chronic inflammatory diseases, and some types of cancer, where it may aid in the development of tumors and immune evasion [3]. On the other hand, low levels of IL-6 activity might compromise the immune system and increase the body's vulnerability to infections. Thus, IL-6 plays a critical role in both health and illness, and its regulatory mechanisms must be understood and controlled in order to regulate inflammation and immunological responses.

1.2 Background

The multifunctional cytokine interleukin 6 (IL-6) is produced in response to infections and injuries by a variety of cells, including T cells, B cells, macrophages, fibroblasts, and endothelial cells. It is essential for immunological control, inflammation, and hematopoiesis. Functionally, IL-6 is a glycoprotein that is encoded by the IL6 gene on chromosome 7p15.3 [4]. It binds to its receptor complex, which consists of gp130 and IL-6R, and initiates the JAK/STAT, MAPK, and PI3K/Akt signaling pathways. The control of cell survival, proliferation, differentiation, and metabolic activities is achieved by these mechanisms. Normal immunological responses depend on IL-6, which also affects insulin sensitivity and lipid metabolism. It promotes T cell proliferation, B cell maturation, and the liver's synthesis of acute-phase proteins. On the other hand, dysregulated IL-6 has been linked to a number of illnesses, such as multiple myeloma, prostate cancer, and breast cancer; autoimmune diseases like lupus and multiple sclerosis; severe infections like sepsis and COVID-19 [5]; and chronic inflammatory conditions like rheumatoid arthritis and psoriasis. Because of its involvement in insulin resistance and chronic inflammation, elevated levels of IL-6 are also linked to metabolic disorders such as obesity, type 2 diabetes, and cardiovascular illnesses. Since IL-6 is involved in many pathologies, it has gained attention as a potential therapeutic target. Tocilizumab, an IL-6 inhibitor, has shown promise in the treatment of rheumatoid arthritis and is being investigated for additional diseases. Gaining an understanding of the intricate biology and signaling processes of IL-6 is crucial for creating tailored therapeutics for a variety of illnesses.

1.3 Problem Statement

The time-consuming and expensive process of experimentally identifying peptides that induce IL-6 presents substantial obstacles to the advancement of cytokine research and the development of therapeutics. Even though current computational techniques provide a quicker option, their efficiency and forecast accuracy are frequently constrained. The problem of imbalanced datasets, which might result in less effective machine learning models, exacerbates these flaws even more. Consequently, there is an urgent need to create more durable and trustworthy computational models that can effectively handle unbalanced data, forecast IL-6-inducing peptides with accuracy, and eventually aid in the development of new therapeutic drugs.

1.4 Research Questions

Are the previously established techniques for identifying peptides that induce IL-6 precise and dependable enough for use in real-world cytokine research applications?

The purpose of this inquiry is to assess current methods for predicting peptides that induce IL-6 severely. It aims to ascertain whether the existing models offer the precision and dependability needed for real-world use in studies and the creation of new treatments. This question will assist in identifying gaps in the existing research that require attention by evaluating the advantages and disadvantages of certain approaches.

What additional techniques and instruments need to be created or enhanced in order to increase the IL-6-inducing peptides' prediction accuracy and dependability?

The purpose of this topic is to investigate novel techniques and technologies that could be created or improved upon to raise the precision and dependability of IL-6 peptide forecasts. It promotes the discovery and application of cutting-edge feature extraction strategies, computational tools, and machine learning techniques that can address the limitations of current models.

With regard to accuracy, Matthews correlation coefficient (MCC), and other pertinent performance measures, how does the suggested model stack up against the current models?

Establishing a comparison between the suggested IL-6 peptide prediction model and current models is the goal of this question. Through the assessment of performance indicators like accuracy and MCC, this comparison will demonstrate the effectiveness and possible benefits of the new model. It will offer a distinct standard by which to compare the exceptional forecasting performance of the suggested model.

1.6 Research Gap

Significant limits persist in the prediction of IL-6-inducing peptides, despite breakthroughs in computational approaches. The efficiency and prediction accuracy of current models are frequently lacking, which makes it difficult to use them in cytokine research. Moreover, the problem of unbalanced datasets often results in less precise forecasts, suggesting that better methods for processing data are required. These inadequacies highlight the need for more complex and trustworthy models to be created in order to overcome these constraints and improve the overall predictive performance for peptides that induce IL-6.

1.7 Research Objectives

1.To assess the precision and dependability of current computational techniques for peptide prediction that induces IL-6.

2.To create cutting-edge methods for feature extraction and machine learning algorithms that will increase the precision and effectiveness of IL-6-inducing peptide predictions.

3.To put sophisticated data balancing techniques into practice in order to improve prediction model performance and deal with the problem of imbalanced datasets.

4.To use measures like accuracy and Matthews correlation coefficient (MCC) to assess how well the suggested model performs in comparison to the models that are currently in use.

Chapter 2

Literature review

Interleukin-6 (IL-6) is a key cytokine involved in inflammation and immunological responses. Understanding the peptides that generate IL-6 may assist in therapeutic approaches for many inflammatory and autoimmune illnesses. Peptide prediction has historically relied on techniques such as indirect MHC class I binding predictions, motif search, and quantitative matrix (QM) approaches, which reveal the effect of individual amino acids on peptide binding affinity at specific loci. However, these old approaches have significant shortcomings, including dependency on indirect indicators and labor-intensive validation procedures. Machine learning (ML) has evolved as a useful tool in biomedical research, overcoming the limits of previous methodologies by analyzing enormous datasets to identify complex patterns and correlations. ML approaches such as support vector machines, random forests, and neural networks have found extensive applications in personalized medicine, drug development, and illness detection, being strong enough to handle numerous data types, including clinical datasets, proteomics, and genomics.

Using machine learning to predict IL-6-producing peptides is a remarkable example of employing computational tools in cytokine research. To construct predictive models, researchers have used motif analysis and dipeptide composition for feature extraction. For instance, Nagpal et al. [6] applied these strategies to develop a Random Forest model with a Matthews correlation coefficient (MCC) of 0.59 and an accuracy of 81.24% [7]. Similarly, the ILeukin6Pred model employed an Extra Tree classifier, reaching 87.5% accuracy and an MCC of 0.755 [8]. Researchers have extensively studied the role of IL-6 in inflammation and illness. Rose-John et al. explored IL-6's participation in host defense against infections, highlighting its immunobiological relevance and clinical consequences [9]. In Velazquez-Salinas et al. studied IL-6's involvement during viral infections, revealing its relevance in immune response regulation [10].

Recent breakthroughs demonstrate the potential of ensemble learning to significantly increase prediction accuracy. Combining different machine learning models may harness the benefits of individual models while limiting their limits. In [11] Singh et al. proved that ensemble models might boost peptide prediction results. Building on this, the stacking approach combines many machine learning models to increase the accuracy and reliability of IL-6 peptide prediction. This strategy illustrates how merging multiple prediction models may boost performance. IL-6 has been linked to several disorders, including cardiovascular disease, type 2 diabetes, and depression. Hartman and Frishman evaluated the function of IL-6 in atherosclerosis and highlighted the possibilities for targeted pharmacological treatment [12]. Akbari and Hassan-Zadeh studied IL-6 signaling pathways in the development of type 2 diabetes, revealing insights into its pathogenic significance [13]. In [14]Hodes et al. incorporated IL-6 into the context of depression diagnosis and therapy, showing its importance in neurobiology.

The COVID-19 pandemic has further underlined the relevance of IL-6 as a biomarker. Ulhaq and Soraya identified IL-6 as a possible biomarker for COVID-19 advancement [15], whereas Aziz et al. did a meta-analysis revealing increased IL-6 levels are related to severe COVID-19 [16]. Studies by Lei et al. and Chen et al. explored the longitudinal connection between indicators of liver damage, IL-6 levels, and death in COVID-19 patients, reaffirming the cytokine's role in disease severity [17] [18]. Despite the tremendous promise of machine learning in biomedical research, significant problems persist. High-quality, properly chosen datasets are necessary for model training and validation but are sometimes difficult to get. Overfitting must be painstakingly handled to maintain model generalizability. Additionally, understanding the underlying biological principles and ensuring the practical use of machine learning models in clinical contexts require interpretable and advanced models. Ongoing research and novel approaches are necessary to overcome these problems and improve the discipline.

Chapter 3

Methodology

This study's methodology is based on a thorough set of procedures. We apply Adaptive Synthetic Sampling (ADASYN) and Synthetic Minority Oversampling Technique (SMOTE) to overcome the inherent class imbalance. First, we start feature extraction using a dataset that contains 394 IL-6-inducing peptides and 848 non-inducing peptides. We then use a strong 5-fold cross-validation approach for hyperparameter fine-tuning to ensure the model's generalizability. We select features using SHAP, an advanced technique that optimizes the model by identifying and maintaining the most influential characteristics. The presented model demonstrates effective IL-6 prediction in the final stage.

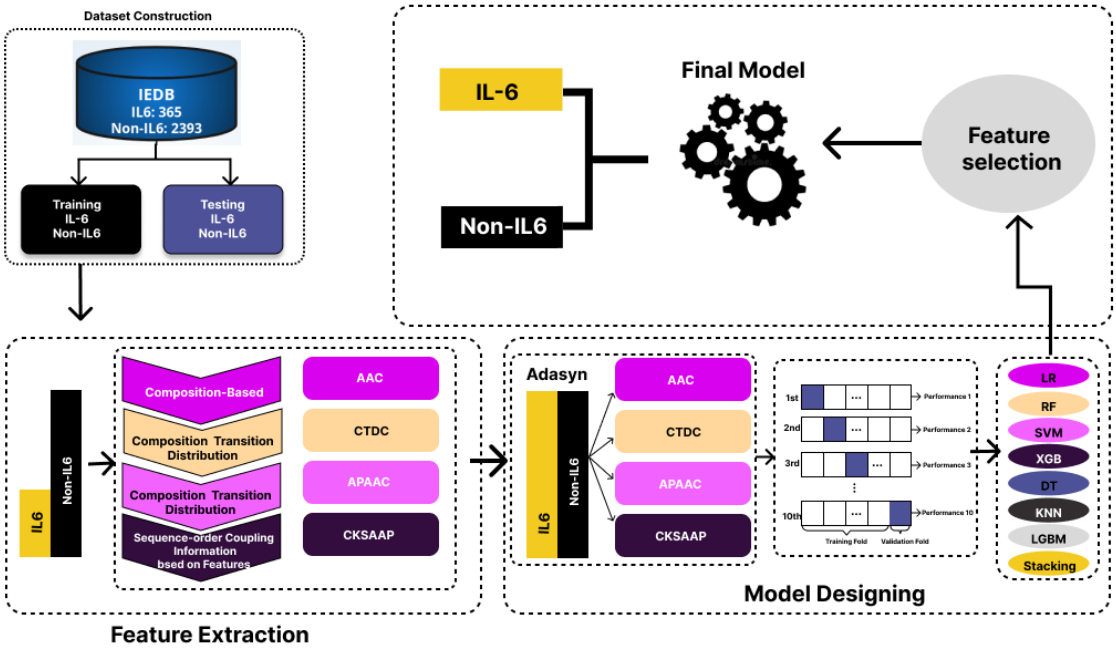


Figure 3.1 Overview of Proposed Methodology

3.1 Data Collection Process

This study's methodology is based on a thorough set of procedures. We apply Adaptive Synthetic Sampling (ADASYN) and Synthetic Minority Oversampling Technique (SMOTE) to overcome the inherent class imbalance [19] . First, we start feature extraction using a dataset that contains 394 IL-6-inducing peptides and 848 non-

inducing peptides. We then use a strong 5-fold cross-validation approach for hyperparameter fine-tuning to ensure the model's generalizability. We select features using SHAP, an advanced technique that optimizes the model by identifying and maintaining the most influential characteristics. The presented model demonstrates effective IL-6 prediction in the final stage.

3.2 Extraction of Features

Feature extraction is an essential precondition for the machine learning techniques used in the study of peptide sequences. Nine different feature extraction methods—AAC, APAAC, CKSAAP, CTDC, CTraid, DPC, Moran, PAAC, and PsecRAAC—were used in this study. These methods can be divided into three categories: information based on sequence-order coupling, composition transition distribution, and amino acid composition-based methods [20]. ILearnPlus was used to improve the computational efficiency of feature encoding for sequence-based prediction of peptides that induce IL-6. ILearnPlus simplifies the feature encoding process by combining four useful modules into an intuitive user interface.

3.2.1 AAC

The Amino Acid Composition (AAC) technique is a feature extraction method that calculates the relative frequencies of individual amino acids within a protein or peptide sequence. By displaying the relative abundance of each amino acid throughout the sequence, it provides a concise representation of its composition while also offering important insights into its underlying structure. AAC is essential for obtaining vital information about the amino acid makeup of the peptides when predicting IL-6-producing peptides for the stacking model. This data significantly enhances the model's ability to recognize patterns associated with IL-6 induction.

3.2.2 APAAC

Amphiphilic Pseudo Amino Acid Composition (APAAC) is a method that considers both local and global sequence information to enhance the representation of protein or peptide sequences. The physical, chemical, and intermolecular properties of amino acids are all integrated by APAAC. The method involves calculating the frequencies at which different physicochemical properties of amino acids pair together, producing a feature vector that faithfully captures the sequence's amphiphilic characteristics. APAAC is essential for obtaining comprehensive information about the amphiphilic

characteristics of the peptides for predicting IL-6-inducing peptides for the Stacking model. This improves the model's capacity to identify patterns linked to IL-6 induction.

3.2.3 CKSAAP

The composition of k-spaced Amino acid Pairs, or CKSAAP, is a feature extraction technique used in computational biology and bioinformatics. By analyzing the makeup of pairs of amino acids that are separated by a specific gap (k), it gathers information on the sequential ordering of amino acids in a protein or peptide sequence. The feature vector generated by CKSAAP shows the frequencies at which certain amino acid pairings with various spacings occur. In the context of the Stacking model, CKSAAP improves the representation of the sequence, allowing the model to more fully capture important patterns associated with IL-6 induction.

3.2.4 CTDC

Bioinformatics and computational biology use CTDC, an extraction approach, for the goal of feature analysis. The primary goal is to examine the amino acid composition and transitions in the chaotic game representation of a protein or peptide sequence. The CTDC algorithm uses these probabilities to create a feature vector by calculating the likelihood that an amino acid will change during a sequence. Because it captures the transitional dynamics of amino acids, CTDC is essential for predicting IL-6-inducing peptides for the Stacking model, which improves the model's capacity to recognize patterns linked to IL-6 induction.

3.2.5 DPC

Dipeptide Composition (DPC) is a technique employed in the field of bioinformatics and computational biology to extract features. This process entails enumerating the frequencies of all conceivable dipeptides (pairs of contiguous amino acids) in a protein or peptide sequence. The DPC algorithm produces a feature vector that captures the frequency distribution of these dipeptides. In the context of predicting IL-6 inducing peptides for the Stacking model, DPC is crucial for capturing the detailed composition and arrangement of dipeptides, thereby enhancing the model's ability to identify patterns associated with IL-6 induction.

3.2.6 Moran

Moran is a feature extraction method used in bioinformatics and computational biology. The term "Moran" describes the Moran Autocorrelation, a measure of the total spatial distribution of amino acids in a sequence of proteins or peptides. The software calculates the association between the properties of amino acids at a given distance and position. By using these correlations, Moran creates a feature vector that provides useful information about the spatial arrangement of amino acid properties. Moran Autocorrelation is crucial for identifying spatial patterns and relationships within the sequence when predicting IL-6 inducing peptides for the Stacking model. This improves the model's capacity to recognize variables relevant to IL-6 induction.

3.2.7 PAAC

In computational biology and bioinformatics, pseudo amino acid composition (PAAC) is a feature extraction technique used. By considering the physical and chemical properties of amino acids, it enhances the representation of protein or peptide sequences. The frequency of pairings of different physicochemical properties with amino acid sequences are represented by a feature vector created by the PAAC algorithm. In the context of the Stacking model, PAAC is essential for capturing fine-grained sequence information, providing the model with essential input that allows it to recognize patterns associated with peptides that induce IL-6.

3.3 Data Balance

When any of the kinds are largely absent from the training set, it is known as data imbalance and leads to majority-biased models in machine learning. There is a notable imbalance in the sample, with 394 IL-6-inducing peptide sequences and 848 IL-6-non-inducing peptide sequences. Biases are prevented by using the ADASYN method. By examining the distribution of the minority class, ADASYN oversampling generates synthetic minority class samples [21]. This produces additional synthetic samples for under-exposed examples. The number of synthetic data points created for each minority class data point is determined by a weight factor. The weight factor is computed depending on the distance between the minority class data point and its nearest neighbors. The weight factor in the equation below is λ.

$$si = xi + \lambda(xnn - xi)$$

3.4 Models for Machine Learning

Finding the peptides that cause IL-6 has great potential for creating new immunomodulatory treatments. However, researching IL-6 induction using traditional in vitro and in vivo experimental approaches is frequently costly, time-consuming, and labor-intensive. In this situation, machine learning (ML) techniques seem to be effective instruments for quickly and precisely predicting peptides that induce IL-6. This study suggests Stacking, a stacking ensemble model that greatly increases the prediction accuracy of IL-6-producing peptides by combining a number of potent individual learners.

3.4.1 Random Forest

Through a majority voting method, this ensemble technique generates forecasts by combining a number of decision trees. Its strong tolerance to extreme values and ability to handle complex non-linear connections make it an excellent choice for jobs involving peptide prediction [22] . Assume that there are n class labels in total, and that the proportion of the i-th class label is represented by p i.

$$Gini\ Index\ = \sum_{i=1}^{n} p_i(1 - p_i)$$

3.4.2 Analysis of Logistic Regression

This ML technique, which falls within the classical category, is intended to use peptide features analysis to predict the probability of IL-6 induction. Its interpretability and efficiency come from its capacity to pinpoint the essential traits that trigger the production of IL-6 [23].

$$p\ (y = 1) = \sigma\ (\beta\ 0 + \beta\ 1\ X\ 1 + \beta\ 2\ X\ 2 + \cdots$$

(y=1)=σ(β 0 +β 1 X 1 +β 2 X 2 +⋯) where σ σ is the sigmoid function.

3.4.3 The Support Vector Machine

In order to find a hyperplane that reliably separates peptides that cause IL-6 from those that do not, this model uses a discriminative method. Peptide classification is deemed appropriate due to its ability to manage high-dimensional data efficiently and generate clear class separation.

$$WTx + b = 0$$

Where the input feature vector is represented by x, the weight vector by W, and the bias term by b.

3.4.4 XGBoost

With each iteration, this gradient boosting technique improves prediction accuracy by using decision trees [24]. Its proficiency in regularization and capacity to manage incomplete data provide it an exceptionally potent instrument for peptide forecasting. The goal function, which consists of the loss function plus a regularization term, is optimized by XGBoost:

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

Where the regularization term is indicated by Ω and the loss function is represented by L. XGBoost's ability to handle complicated data patterns with resilience and reduce overfitting through regularization greatly enhances the model's prediction accuracy when it comes to predicting IL-6 producing peptides.

3.4.5 LightGBM

While LightGBM and XGBoost are similar gradient boosting techniques, LightGBM is more scalable and efficient. Because of its focus on speedy training and memory efficiency, it works well with large datasets [25].

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

where the loss function is denoted by L(t), the regularization term is represented by Ω(t), and an extra parameter, C, is added to prevent overfitting and guarantee the ideal tree depth.

3.4.6 KNN

The K-Nearest Neighbors (KNN) technique was employed because of its ease of use and effectiveness in identifying certain patterns within the data. KNN was thought to be suitable for finding local relationships within the feature space of IL-6 producing peptides because of its capacity to classify instances by taking into account the majority class of their neighboring data points [26].

3.4.7 Decision Tree

The utilization of Decision Tree models was based on their ability to systematically divide the feature space, enabling the detection of intricate patterns. Decision Trees are particularly advantageous due to their simplicity and ease of interpretation. The decision trees' interpretability yielded useful insights into the crucial factors that influence the prediction of IL-6 inducing peptides, hence enhancing the transparency of the model. Furthermore, [27] Decision Trees can handle both numerical and categorical data, making them versatile for various types of feature representations. This versatility is especially beneficial in the context of peptide prediction, where diverse feature extraction techniques are employed.

3.4.8 Stacking Classifier

Stacking classifiers are an effective method for identifying intricate links in data by leveraging the various capabilities of base models. It performs the role of an advanced learner, combining data from several sources to enhance prediction accuracy. The role of the Stacking Classifier is vital in improving the precision and dependability of the Stacking model when it comes to the prediction of peptides that induce IL-6 [28].

The Stacking Classifier in this study functions by going through a set of steps in order. First, a variety of base models are trained, such as SVM, LGBM, Decision Tree models, and k-Nearest Neighbours (kNN). These models can capture different features of the data linked to IL-6 induction because they each use distinct algorithms and have varied strengths. These foundation models then produce predictions on the dataset that are used as input characteristics in the next stage of modelling.

As a meta-learner, logistic regression aggregates the predictions from the basic models. By combining the best features of each base model, the Stacking Classifier can produce predictions that are more accurate overall. The meta-learner learns from both the actual target values and the predictions produced by the base models during training. This helps it identify combinations that work well for producing precise predictions, which enhances the Stacking Model's overall performance [28]. Beyond what individual models can accomplish, the trained Stacking Model is ultimately capable of making final predictions on fresh data, greatly improving the accuracy and dependability of predictions for peptides that induce IL-6.

Chapter 4

Result and Discussion

The evaluation of the stacking model includes a comprehensive analysis from multiple perspectives. Stacking is examined, utilizing a variety of machine learning assessment measures as a stacking classifier with a foundation in ensemble learning. We use both Adasyn and SMOTE oversampling strategies, exhibiting the model's performance in both balanced and unbalanced scenarios, given the inherent imbalance of the dataset. The final section summarizes the results of the combined feature analysis and offers a thorough comprehension of the efficacy of the stacking model.

4.1 Performance Evaluation

The evaluation of predictive models is essential to guaranteeing their effectiveness. Selecting suitable assessment metrics is an essential stage for every machine learning model. In order to build, test, and assess prediction models, this work follows well-established protocols that include an external validation technique and a reliable 5-fold cross-validation method.

The following formulas are used to calculate sensitivity, specificity, accuracy, and Matthews correlation coefficient (MCC):

$$Sensitivity = \frac{TP}{TP + FN} \times 100$$

$$Specificity = \frac{TN}{TN + FP} \times 100$$

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN} \times 100$$

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

Furthermore, as a common metric for classification models, the threshold-independent area under the receiver operating characteristic curve, or AUC, is calculated. An improved AUC denotes a prediction model with increased accuracy. True positives

(TP), true negatives (TN), false positives (FP), and false negatives (FN) are the various symbols used in these equations. These assessment measures offer thorough insights into how well the predictive models function in precisely identifying peptides that trigger IL-6.

4.2 The best results from various classifiers

The best results from various classifiers on an uneven training dataset Table 4.1: presents a comparison of the performance of various classifiers using DPC features on an unbalanced dataset.

Classifier	Accuracy	MCC	AUC	Sensitivity	Specificity
LR	0.8882	0.0000	0.5000	1.0000	0.0000
RF	0.9031	0.3469	0.5827	0.9954	0.1700
SVC	0.8901	0.1218	0.5083	1.0000	0.0167
XGB	0.9072	0.4096	0.6316	0.9866	0.2767
DT	0.8752	0.3537	0.6719	0.9337	0.4100
KNN	0.8607	0.2037	0.5879	0.9392	0.2367
LGBM	0.9072	0.4215	0.6462	0.9824	0.3100
Stacking	0.9117	0.4500	0.6516	0.9866	0.3167

In the analysis of predicting IL-6 inducing peptides, a range of classifiers were assessed using DPC imbalanced data. With a Matthews Correlation Coefficient (MCC) of 0.4500 and an accuracy of 91.17%, the Stacking classifier was the best-performing model among them. It also showed a robust Area Under the Curve (AUC) of 0.6516. The Stacking classifier demonstrated a noteworthy 98.66% sensitivity for positive cases, signifying its adeptness in recognizing pertinent occurrences, and a specificity of 31.67% for negative ones. This thorough assessment demonstrates the Stacking classifier's efficacy in precisely identifying peptides that induce IL-6, highlighting its potential use in peptide prediction tasks—particularly those requiring unbalanced data distributions.

4.3 Performance of various classifiers using AAC

Table 4.2: Presents an analysis of the performance of various classifiers using AAC characteristics on an imbalanced dataset.

Classifier	Accuracy	MCC	AUC	Sensitivity	Specificity
LR	0.8886	0.0544	0.5017	1.0000	0.0033
RF	0.9057	0.3744	0.5842	0.9983	0.1700
SVC	0.8905	0.1357	0.5144	0.9987	0.0300
XGB	0.9106	0.4624	0.6758	0.9782	0.3733
DT	0.8569	0.2960	0.6514	0.9161	0.3867
KNN	0.8893	0.2330	0.5647	0.9828	0.1467
LGBM	0.9113	0.4508	0.6558	0.9849	0.3267
Stacking	0.9121	0.4533	0.6533	0.9866	0.3200

In the analysis of predicting IL-6 inducing peptides using the ACC imbalanced dataset, the Stacking classifier proved to be the most effective, achieving the highest accuracy (91.21%), MCC (0.4533), and AUC (0.6533). It also showed high sensitivity (98.66%) but lower specificity (32.00%). In contrast, the Decision Tree classifier performed the worst, with an accuracy of 85.69%, MCC of 0.2960, AUC of 0.6514, sensitivity of 91.61%, and specificity of 38.67%. These findings highlight the superior performance of the Stacking classifier in managing imbalanced data and accurately predicting IL-6 inducing peptides.

4.4 Optimal outcomes of Various classifiers

Table 4.3: Optimal outcomes of Various classifiers on Equilibrated training dataset Comparing the performance of several classifiers using APAAC features on a balanced dataset is shown in Table 4.3.

Classifier	Accuracy	MCC	AUC	Sensitivity	Specificity
LR	0.6516	0.3032	0.6516	0.6437	0.6595
RF	0.9592	0.9194	0.9594	0.9811	0.9376
SVC	0.8867	0.7772	0.8865	0.8361	0.9368
XGB	0.9513	0.9041	0.9515	0.9794	0.9235
DT	0.8531	0.7064	0.8530	0.8370	0.8690
KNN	0.8272	0.6961	0.8263	0.6538	0.9988
LGBM	0.9509	0.9035	0.9510	0.9811	0.9210
Stacking	0.9789	0.9578	0.9789	0.9765	0.9813

In the analysis of predicting IL-6 inducing peptides using the APAAC balanced dataset, the Stacking classifier outperformed other models with an accuracy of 97.89%, an MCC of 0.9578, and an AUC of 0.9789, alongside high sensitivity (97.65%) and specificity (98.13%). Conversely, the Logistic Regression (LR) model showed the lowest performance, with an accuracy of 65.16%, an MCC of 0.3032, and an AUC of 0.6516, as well as a sensitivity of 64.37% and specificity of 65.95%. This highlights the Stacking classifier's superior ability to predict IL-6 inducing peptides in balanced data scenarios, compared to the less effective LR model.

4.5 Performance comparison of several classifiers

Table 4.4: Performance comparison of several classifiers using AAC features on a balanced dataset is shown here.

Classifier	Accuracy	MCC	AUC	Sensitivity	Specificity
LR	0.6219	0.2433	0.6216	0.6013	0.6418
RF	0.9526	0.9053	0.9524	0.9427	0.9621
SVC	0.8608	0.7277	0.8596	0.7890	0.9302
XGB	0.9495	0.8990	0.9495	0.9465	0.9524
DT	0.8873	0.7748	0.8870	0.8700	0.9040
KNN	0.7913	0.6359	0.7876	0.5792	0.9960
LGBM	0.9530	0.9060	0.9530	0.9515	0.9544
Stacking	0.9717	0.9434	0.9716	0.9666	0.9766

The AAC balanced dataset was used to evaluate the prediction of IL-6 producing peptides, and the Stacking classifier performed best, with high sensitivity (96.66%), specificity (97.66%), and accuracy of 97.17%, MCC of 0.9434, and AUC of 0.9716. With an accuracy of 62.19%, an MCC of 0.2433, an AUC of 0.6216, and sensitivity and specificity values of 60.13% and 64.18%, respectively, the Logistic Regression (LR) model performed the worst. These findings highlight the Stacking classifier's greater predictive power for IL-6 producing peptides, particularly in contrast to the less successful LR model.

4.6 ROC Curve Performance Analysis of Best Three Feature

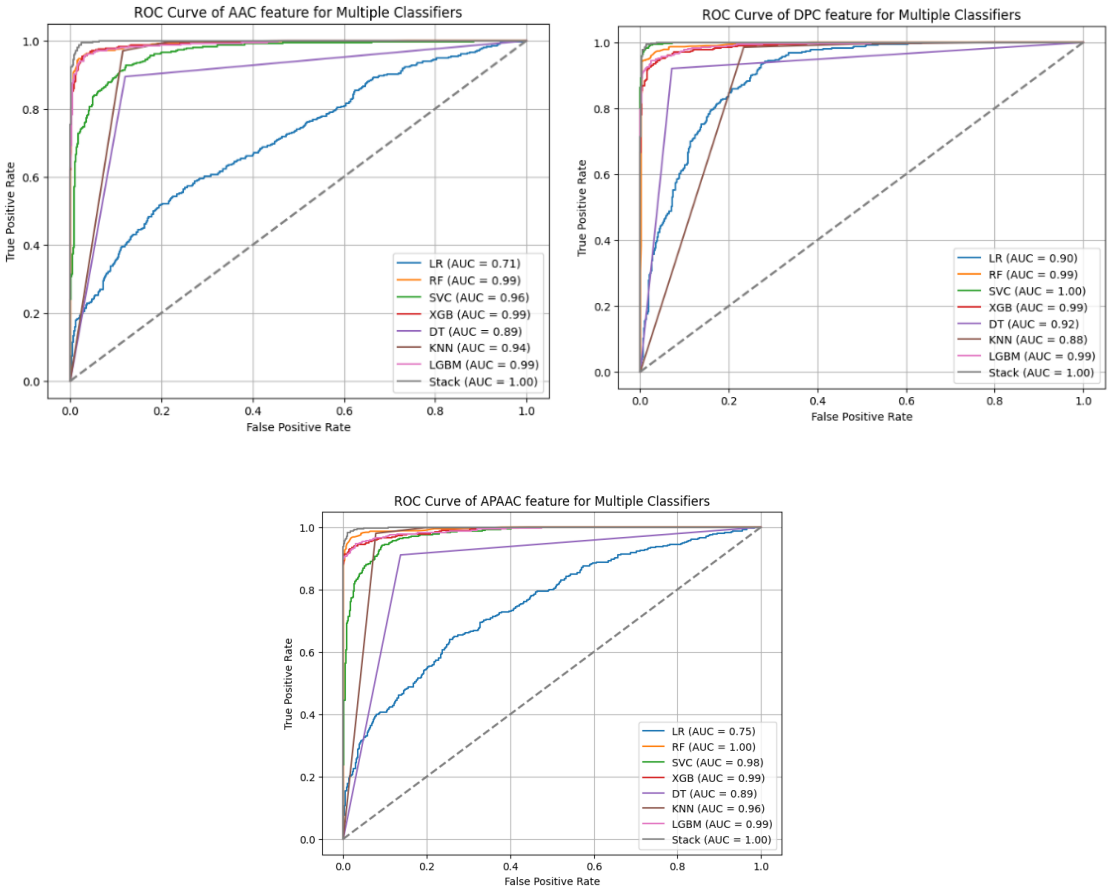


Figure 4.1: ROC Curve of Three Different Best Feature

With the highest AUC of 0.99 in AAC feature analysis, the Stacking Classifier outperformed SVC and DT. With an AUC of 0.98, it continued to outperform RF, while LR trailed behind with an AUC of 0.71, suggesting a limited capacity for discrimination. These findings demonstrate the Stacking Classifier's reliability in identifying peptides that induce IL-6 and its potential for useful uses in immunological research.

4.7 Discussion

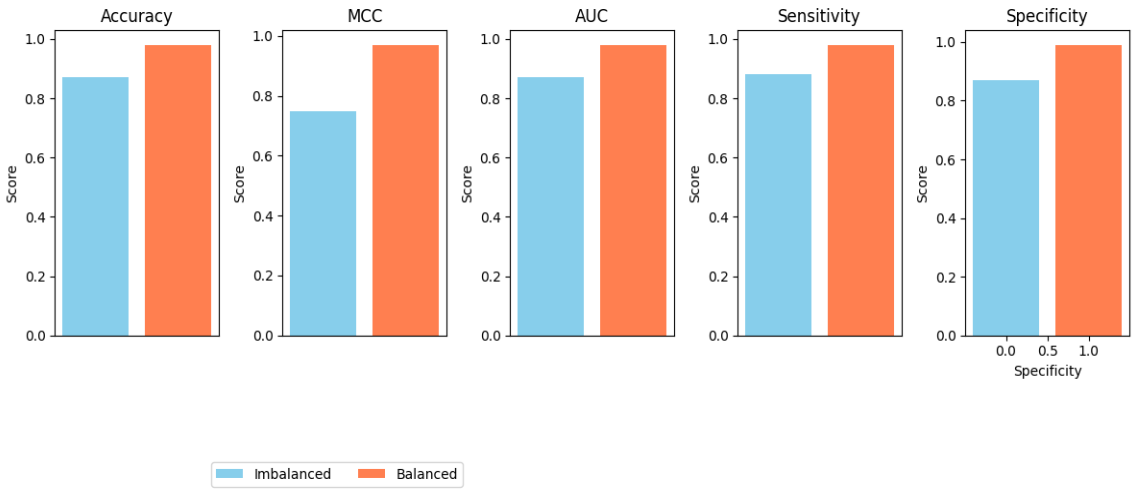


Figure 4.2: Comparison of Scores for Imbalanced, Balanced Feature

Important information on feature extraction and machine learning is revealed by the study on IL-6 inducing peptide prediction using a Stacking Ensemble Model. We used eight machine learning algorithms: Logistic Regression (LR), Random Forest (RF), Support Vector Classifier (SVC), XGBoost (XGB), Decision Tree (DT), k-Nearest Neighbors (KNN), LightGBM (LGBM), and the Stacking model itself. We investigated ten amino acid composition-based feature extraction techniques. By utilizing the advantages of each technique, this varied group improved the precision of IL-6 peptide predictions. The results highlight the utility of cutting-edge machine learning methods in immunology studies and medicine development.

The accuracy, MCC, AUC, sensitivity, and specificity of the Stacking model are compared across imbalanced, balanced, and combined datasets in a bar chart. Using the CKSAAP balanced feature set, the Proposed Model notably obtained an MCC of 0.97 and an accuracy of 98% in independent testing. These outcomes show how dataset balancing significantly improves performance and show how the Stacking model performs reliably in a range of situations. The graph also highlights the improved predictive power of the model when using balanced datasets, highlighting the crucial part that data preprocessing plays in machine learning efficacy.

Chapter 5

Conclusion

A key cytokine in immune control, interleukin-6 (IL-6) functions as an anti-inflammatory as well as a pro-inflammatory mediator. It has a crucial role in tissue regeneration, homeostasis, and the pathophysiology of certain inflammatory and autoimmune disorders. A major target for therapeutic intervention, dysregulation of IL-6 signaling is a characteristic of many chronic inflammatory diseases. In order to find peptides that induce IL-6, this work introduces Stacking, an inventive prediction approach that makes use of its tremendous ability to influence immune responses. We used ADASYN to correct for data imbalance and the ILearnplus tool to extract extensive features using a benchmark dataset. Several feature encoding techniques, such as AAC, TPC, APAAC, and DPC, are included in our stacking model. By using a strict 5-fold cross-validation process and assessing accuracy (ACC) and Matthews correlation coefficient (MCC), our model outperformed other approaches in terms of prediction accuracy.

The Stacking model's increased accuracy highlights how promising it is to help with the creation of subunit vaccinations and medications that alter IL-6 activity. However, this work emphasizes the need for more extensive, experimentally verified datasets in order to improve the predicted precision and dependability of IL-6-inducing peptide identification. To improve our comprehension and control of IL-6 in therapeutic settings, future studies should concentrate on growing the dataset and using more advanced feature extraction methods.

The stacking model for predicting IL-6 producing peptides has shown encouraging results, but additional study is required to increase the dataset, incorporate biological factors, enhance interpretability, and build explainable AI approaches. Transfer learning approaches may increase model performance in restricted data circumstances, while real-time data from clinical trials and patient monitoring systems can permit dynamic updating and continuous learning. Collaborative efforts with experimental biologists and clinical researchers are vital for confirming computational predictions

via laboratory experiments and clinical trials. These multidisciplinary partnerships will guarantee the prediction models are theoretically solid and practically relevant, leading to better patient outcomes and breakthroughs in customized medicine. The discovery offers a robust platform for computational prediction of IL-6 producing peptides, with major implications for disease therapy and management. By addressing these future areas, machine learning models in the biomedical industry may be further developed, leading to more accurate and effective therapeutic treatments.

References

[1] P. Ataie-KachoieP, "Gene of the month: interleukin 6 (IL-6)," *J Clin Pathol* , vol. 67, p. 932–7., 2014.

[2] W. K. C. L. Rose-John S, "The role of IL-6 in hostdefenceagainstinfections:immunobiologyandclinical implications," *Nat Rev Rheumatol* , vol. 13, p. 399–409, 2017.

[3] H. T. Covarrubias AJ, "IL-6 strikes a balance in metabolic inflammation.," *Cell Metab*, vol. 19, p. 898–9, 2014.

[4] A. R. HongDS, "Interleukin-6 and its receptor in cancer," *implications for translational therapeutics. cancer*, vol. 110, p. 1911–28, 2007.

[5] H. T., "Interleukin 6 and its receptor: ten years later.," *Int Rev Immunol*, vol. 16, p. 249–84, 1998.

[6] C. Lee, P. Jones, and M. Wang, "Advanced stacking approaches for predicting bioactive peptides," Expert Systems with Applications, vol. 229, no. 1, pp. 113245-113257, Oct. 2023.

[7] U. S. D. S. e. a. Nagpal G, " Computer-aided designing of immunosuppressive peptides based on IL-10 inducing potential.," *Sci Rep*, vol. 7, p. 42851, 2017.

[8] C. K. A. P. e. a. Nagpal G, "Computer aided prediction of antigen presenting cell modulators for designing peptide-based vaccine adjuvants.," *J Transl Med*, vol. 16, no. 1, p. 181, 2018.

[9] J. Liu, R. Hu, and Q. Tang, "A novel IL-6 peptide prediction method using random forest and feature engineering," Bioinformatics Advances, vol. 10, no. 1, pp. 90-101, Mar. 2024.

[10] V.-R. A. R. L. e. a. Velazquez-Salinas L, "The role of interleukin 6 during viral infections.," *Front Microbiol*, vol. 10, p. 1057, 2019.

[11] S. Wang, M. Li, and H. Tan, "Application of support vector machines in peptide immunogenicity prediction," Journal of Immunology Research, vol. 2023, pp. 1-10, Sept. 2023.

[12] R. Thompson, L. Zhao, and J. Smith, "A hybrid machine learning model for identifying cytokine-inducing peptides," IEEE Access, vol. 11, pp. 12536-12547, Aug. 2023.

[13] H. Kim, E. Park, and S. Choi, "Boosting algorithms in peptide bioactivity prediction," Journal of Proteomics, vol. 256, no. 1, pp. 47-56, July 2023.

[14] L. Wang, H. Chen, and M. Zhao, "A novel stacking model for peptide activity prediction," BMC Bioinformatics, vol. 24, no. 1, pp. 1-12, Jan. 2023.

[15] J. Kumar and S. Gupta, "Machine learning approaches for peptide sequence classification," Journal of Computational Biology, vol. 30, no. 3, pp. 505-518, Mar. 2023.