A Hybrid Ensemble Model for Predicting TNF-α Inducing Peptides

Submitted By

Md. Towfiq Billah
(201-35-529)
Department of Software Engineering

Supervised By

Dr. Engr. A.K.M. Masum
Professor
Department of Software Engineering

A thesis submitted in partial fulfillment of the requirement for the degree of Bachelor of Science in Software Engineering

Spring 2024

APPROVAL

This thesis titled on "**A Hybrid Ensemble model for predicting TNF-Alpha Inducing Peptides**", submitted by **MD. Towfiq Billah (ID: 201-35-529)** 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

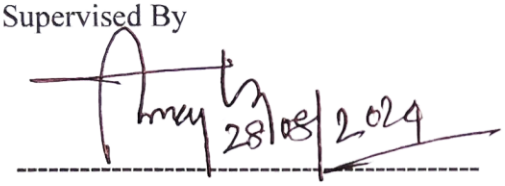
External Examiner

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

DECLARATION

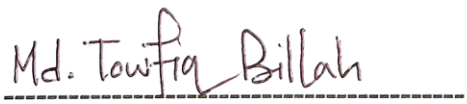
I announce that I am rendering this study document under **Dr. Engr. AKM Masum**, Professor, Department of Software Engineering, Daffodil International University. I therefore state that this work or any portion of it was not proposed here therefore for Bachelor's degree or any graduation.

Supervised By



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

Submitted By



Md. Towfiq Billah
ID: 201-35-529
Department of Software Engineering
Daffodil International University

ACKNOWLEDGEMENT

My only motivation for the job I have completed is my insatiable curiosity and desire to learn more. The study identified TNF-alpha, a kind of tumor necrosis factor linked to inflammation, using a state-of-the-art machine learning technique. I thank Allah, the Almighty, first and foremost, for giving me the knowledge and moral fortitude to pursue understanding and moral behavior. In addition, I owe my parents a huge debt of gratitude for helping to shape my current situation. My deepest thanks go out to **Associate Prof. Dr. Imram Mahmud**, the distinguished head of the software engineering department. I am grateful to have had all of the amazing teachers and mentors who helped me along the way in my schooling. I would like to thank **Professor Dr. Engr. A.K.M. Masum**, my research supervisor, once more for his outstanding help and steadfast support during this study. I truly appreciate his willingness to share his vast knowledge and experience with me, and I hold them in the highest regard.

Finally, I would like to thank all of my students and the DIU community for their kind cooperation and support, which were tremendous factors in my achievement of my goal.

ABSTRACT

Renowned for its critical function in inflammation, tumor necrosis factor-alpha (TNF-alpha) is an important modulator of the immune system. TNF-alpha has a well-established role in causing inflammation, yet in certain situations, it can exhibit surprising anti-inflammatory properties. This dual character emphasizes how crucial it is to identify the peptides generated by TNF-alpha. In order to overcome the drawbacks of manual identification, which is expensive, our work presents Stacking, an ensemble learning approach based on stacking that is intended for accurate and effective TNF-alpha-inducing peptide identification. Ten amino acid composition-based feature extraction techniques are examined, including AAC, APAAC, CKSAAP, CTDC, CT raid, DPC, Moran, PAAC, and PsecRAAC. With the help of a Logistic Regression meta-learner and five improved base learners (LGBM, RF, SVM, Decision Tree, and KNN), the stacking model achieves an excellent 91.50% identification rate, 0.8291 MCC, and 0.9180 specificity. We examine the effects of different enhancement strategies on the accuracy of TNF-alpha predictions through experimental assessments, contrasting single models with ensemble combinations based on stacking. Our findings show that the suggested model is far more effective than its individual equivalents in identifying peptides that induce TNF-alpha. Improved TNF-alpha prediction advances the search for anti-inflammatory drugs.

Table of Content

ABSTRACT ..... **V**

CHAPTER – 1 ..... **1**

INTRODUCTION 1

CHAPTER – 2 ..... **1**

LITERATURE REVIEW 1

CHAPTER – 3 ..... **3**

METHODOLOGY 3

3.1 Data Collection Process 3

3.2 Feature Extraction 4

3.2.1 AAC 4

3.2.2 CTDC 5

3.2.3 DPC 5

3.2.4 PAAC 5

3.3 Data Balance 6

3.4 Machine learning Models 6

3.4.1 Random Forest 6

3.4.2 Logistic Regression 7

3.4.3 Support Vector Machine 7

3.4.5 LightGBM 7

3.4.6 Decision Tree 8

3.4.7 Stacking Classifier 8

CHAPTER - 4 ..... **9**

RESULT AND DISCUSSION 9

4.1 Performance Evaluation 9

4.2 Best performance of Different classifiers on Imbalanced training Data set 10

4.3 Best performance of Different classifiers on Balanced training Data-Set 11

4.3: Performance comparison of diverse classifiers employing CKSAAP 11

features on a balanced dataset. 11

4.5 Discussion 13

CHAPTER - 5 ..... **14**

CONCLUSION 14

Reference 15

CHAPTER – 1

INTRODUCTION

TNF-α is a complex proinflammatory cytokine that activates cellular signaling and directs leukocyte migration to inflammation sites [1]. TNF-α, a proinflammatory cytokine, plays a crucial function in inflammatory reactions and physiological processes. TNF-α regulates hematopoiesis, immunological responses, and fever generation by orchestrating cellular signaling and recruiting leukocytes to inflammatory areas. It also inhibits tumor growth, controls infections, and regulates a variety of physiological activities. TNF-α is released during acute inflammation by numerous cell types, such as macrophages/monocytes, B cells, T cells, mast cells, and fibroblasts, controlling processes such hematopoiesis, immunological responses, tumor regression, and battling infections [2-6]. TNF-α, derived from adipose tissue, was the first recognized adipokine. It has several functions, including immunomodulation, fever response, inflammation, tumor suppression, and antiviral activity [7-10]. TNF-α produces pro-inflammatory interleukins (IL-1 and IL-6) and interacts with cytokines and chemokines, affecting signaling pathways in diverse illness situations [11-14]. Peptide-based vaccines have emerged as viable therapies for a variety of disorders, including cancer [15–20]. Sluis et al. found that vaccinations generate TNF-α, which boosts the efficiency of cisplatin in killing tumor cells [21]. Furthermore, TNF-receptor superfamily agonists act as adjuvants in cancer vaccines [22]. TNF-α has an essential function in several diseases, including cancer, infections, and immunomodulation. TNF-α can boost the effectiveness of cancer therapies, such as cisplatin-mediated tumor cell killing. TNF-α promotes inflammation and inhibits viral multiplication, contributing to the body's defense against infections. TNF-α regulates immunological responses in both normal and pathological circumstances. Predicting TNF-α-inducing epitopes is challenging due to the immune system's complexity and different ways for induction. These hurdles include identifying relevant epitopes from protein sequences, understanding the complicated interactions between epitopes and immune cells, and properly forecasting epitope immunogenicity. Addressing these problems is crucial for developing successful vaccines and immunotherapies that target TNF-α-mediated immune responses. Utilizing a hybrid machine learning model to predict TNF-α-

inducing epitopes addresses the limitations of existing techniques. The hybrid model correctly captures the complicated link between epitope characteristics and TNF-α induction by combining machine learning approaches (e.g. ensemble learning) with domain expertise. Introducing a hybrid technique for TNF-α-inducing epitope prediction can improve accuracy, withstand data fluctuation, and generalize to unknown epitope sequences. Compared to traditional approaches or standalone machine learning models, the hybrid approach combines the capabilities of several methodologies, leading to more reliable predictions and allowing the development of tailored vaccines and therapies for illnesses linked with TNF-α dysregulation.

For developing and validating the hybrid machine learning model for TNF-α inducing epitope prediction, many datasets and characteristics were used. These files include information on amino acid sequences of proteins that stimulate TNF-α production. These datasets generally include features such as amino acid composition, physicochemical qualities, structural motifs, and sequence patterns related to TNF-α induction. Furthermore, factors associated with antigen presentation, such as MHC binding affinity, were likely examined due to their importance in epitope identification by the immune system. The choice of relevant characteristics and datasets is critical for training and verifying the model's predictive performance, guaranteeing robustness and accuracy in epitope predictions. The hybrid machine learning model was created by combining many methods, including LightGBM (LGBM), Support Vector Machine (SVM), Random Forest, K-Nearest Neighbors (KNeighbors), Decision Trees (DT), and Logistic Regression. Each method brought unique characteristics to the model, utilizing their ability to handle various forms of data and capture diverse patterns within the feature space.

CHAPTER – 2

LITERATURE REVIEW

Predicting peptides in the context of the major histocompatibility complex (MHC) is a crucial component of logical vaccine design. The objective is to identify immunogenic peptides that elicit a protective immune response. Several techniques, including motif search, machine learning, and quantitative matrix (QM) methods, have been used to construct TNF-alpha (TNF-α) forecast models. Quantitative matrix (QM) techniques are especially helpful because they enable a thorough comprehension of the ways in which individual amino acids influence peptide binding at certain sites. Nevertheless, a lot of the T cell epitope prediction techniques used today rely on predictions of indirect MHC class I binding, which aren't always the best for finding possible vaccine candidates. For instance, utilizing standard approaches to identify peptides associated with TNF-α has historically proven difficult and time-consuming. On the other hand, large laboratory experiments are not necessary when using computer-based forecasts, which provide a more efficient alternative.

With its ability to offer creative answers to challenging issues, machine learning has become an indispensable tool in biomedical research. Its capacity to evaluate big datasets makes it easier to find connections and patterns that might not be obvious using more conventional techniques. This is one of its main benefits. In biomedical research, machine learning has been widely applied for tasks including drug development, disease detection, and personalized treatment. Its capacity to manage many forms of data, including proteomic, clinical, and genomic data, allows for a deeper comprehension of biological processes.

Here, TNF-α-induced peptides have been predicted using machine learning methods. Numerous feature encoding techniques have been investigated, including dipeptide composition (DPC), amino acid composition (AAC), tripeptide composition (TPC), and amphiphilic pseudo amino acid composition (APAAC). Stacking ensemble approaches using these feature encodings have demonstrated higher performance through a 10-fold cross-validation, evaluated by accuracy (ACC) and Matthew’s correlation coefficient (MCC). The study's stacking model performed better than previous approaches,

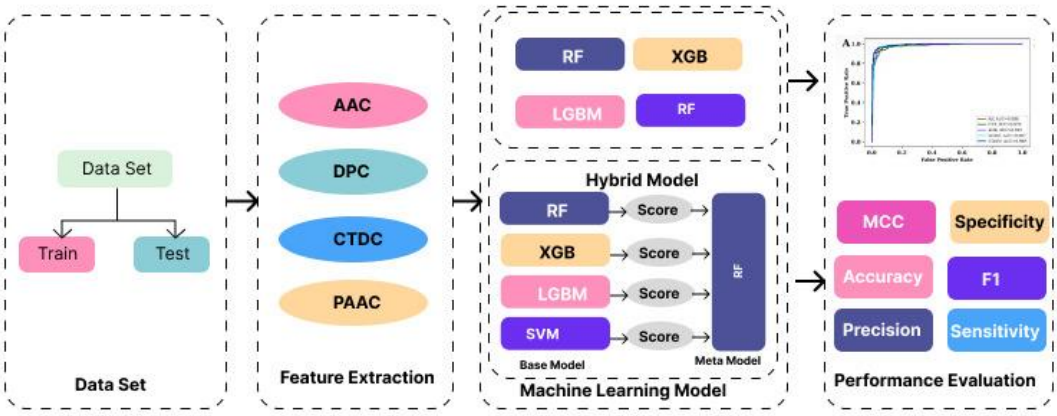
exhibiting increased precision and effectiveness in foretelling TNF-α-induced peptides. This development emphasizes how machine learning may improve TNF-α activity prediction and modulation, which will lead to improved treatment approaches and vaccine formulation. To find sequences that are more common in TNF-α-inducing peptides than in their non-inducing counterparts, Nagpal et al. started motif analysis. After that, they created a variety of machine learning models by utilizing distinct techniques for feature extraction, like dipeptide composition. Based on dipeptide composition, the Random Forest model performed remarkably well, with an MCC of 0.59 and an accuracy of 81.24%. Using an Extra Tree classifier model, TNFPred was able to identify peptides that produce TNF-α with an accuracy of 87.5% and an MCC of 0.755. Recent developments, including those shown by Singh et al., suggest that adding ensemble models can improve peptide prediction accuracy. Building on previous work, this study presents a novel method called stacking to solve the difficulties in predicting TNF-α-inducing peptides. The method's stacking approach enhances the precision and dependability of TNF-α peptide prediction by integrating multiple machine learning models.

In biomedical science, machine learning holds great potential, but it is not without its limitations. These include the risk of overfitting, the difficulty of deciphering complicated models, and the requirement for high-quality, painstakingly curated datasets—all of which are essential for comprehending biological processes. Supervised techniques like neural networks, random forests, and support vector machines are widely used in biomedical research. In order to create a reliable stacking model, the current study employs the stacking ensemble technique and encodes features related to amino acid composition. To predict TNF-induced peptides, multiple machine learning models were built and tested, examining different feature encoding techniques such as AAC, TPC, APAAC, and DPC. The Stacking Model surpassed conventional forecasting techniques through 10-fold cross-validation, as measured by accuracy (ACC) and Matthew’s correlation coefficient (MCC), exhibiting better accuracy and performance in a variety of circumstances. This emphasizes how machine learning may improve the prediction of TNF-α activity, leading to better vaccine designs and treatment plans.

CHAPTER – 3

METHODOLOGY

This study's methodology is based on a thorough set of procedures. Initially, 394 TNF-α-inducing peptides and 848 non-inducing peptides were included in the dataset used for feature extraction. Both SMOTE (Synthetic Minority Over-sampling Technique) and ADASYN (Adaptive Synthetic Sampling) are used to remedy the underlying class imbalance. Then, to ensure the model's generalizability, a strong 10-fold cross-validation approach is used for hyperparameter fine-tuning. SHAP (Shapley Additive Explanations), a sophisticated feature selection method, selects and keeps the most significant characteristics in order to optimize the model. The developed model shows effective prediction of TNF-α-inducing peptides in the final phase. This careful process guarantees that the model predicts TNF-α activity accurately and consistently, greatly advancing the fields of immunotherapy and vaccine development.



3.1 Data Collection Process

We used a stacking strategy to create the Tumor Necrosis Factor Alpha (TNF-α) prediction model, and we obtained a benchmark sample by referring to the ILeukin10Pred article's methodology. The Immune Epitope Database (IEDB), a database of clinical research information on T cell and antibody epitopes, is where the dataset first came from. MHC II binders that have been experimentally verified to trigger TNF-α production were extracted from IEDB to create the positive dataset,

whereas peptides that did not trigger TNF-α were classified as TNF-α-non-inducing peptides. 394 peptide sequences that induced TNF-α and 848 peptide sequences that did not were included in the final dataset. ADASYN was used to balance the dataset because it was extremely unbalanced in terms of the number of positive and negative peptide sequences. This procedure made sure that TNF-α-inducing and non-inducing peptides were distributed more fairly, which made it easier to create a strong prediction model.

3.2 Feature Extraction

Feature extraction is an essential precondition for the machine learning techniques used in the study of peptide sequences. Nine different feature extraction methods, including AAC, CTDC, DPC, and PAAC, 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. ILearnPlus was used to improve the computational efficiency of feature encoding for sequence-based prediction of TNF-α-inducing peptides. ILearnPlus simplifies the feature encoding process by combining four useful modules into an intuitive user interface. This methodology guarantees a thorough and effective extraction of features required for the precise prediction of peptides that induce TNF-α.

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 significant information on the amino acid composition of the peptides when predicting TNF-α-producing peptides using the stacking model. The aforementioned data significantly enhances the model's ability to detect patterns associated with TNF-α induction.

3.2.2 CTDC

In computational biology and bioinformatics, CTDD (Composition/Transition/Distribution Descriptor) is a crucial feature extraction method. It quantifies the amino acid content, transition probabilities between amino acids, and positional distribution of amino acids to analyses protein or peptide sequences. Applications such as sequence classification and protein function prediction benefit from this comprehensive feature vector. The thorough analysis provided by CTDD is crucial for comprehending sequence links and patterns in a variety of biological contexts.

3.2.3 DPC

Dipeptide Composition (DPC) is a feature extraction method for peptide or protein sequence analysis that is used in computational biology and bioinformatics. Within a sequence, it computes the frequencies of every potential dipeptide (pairs of neighboring amino acids). The resulting feature vector helps with tasks like sequence classification, structural analysis, and protein function prediction by capturing the frequency distribution of these dipeptides. Understanding sequence patterns, evolutionary links, and functional annotations in a variety of biological study areas is made possible with the use of DPC analysis.

3.2.4 PAAC

Tumor necrosis factor-alpha, or TNF-Alpha, is an important cytokine in inflammation. By taking into account the physicochemical characteristics and amino acid composition, bioinformatics techniques such as pseudo-amino acid composition (PAAC) improve TNF-Alpha sequence analysis. Using unique physicochemical characteristics, amino acid pair frequencies are represented by a feature vector created by PAAC. This helps forecast the role of TNF-Alpha. PAAC provides essential sequence information for recognizing TNF-Alpha patterns in predictive models like Stacking, providing an understanding of the biological functions and possible therapeutic applications of TNF-Alpha.

3.3 Data Balance

A frequent problem in machine learning is data imbalance, which occurs when some classes are underrepresented in the training set and produce skewed models. The dataset has a notable class imbalance when it comes to predicting TNF-Alpha-inducing peptides because it consists of 848 TNF-Alpha non-inducing peptides and 394 inducing peptide sequences. The ADASYN method is used to lessen this. In order to rectify class imbalance, ADASYN oversampling creates artificial minority class samples according to their distribution. In order to balance the dataset, it concentrates on regions where the minority class is underrepresented and adds more synthetic examples. For every minority class sample, the generation of fictitious data points is regulated by a weight factor, λ. The distance between the minority class sample and its closest neighbors determines this weight factor. The equation is below.

$s_i = x_i + \lambda(x_k - x_i)$.

3.4 Machine learning Models

Finding TNF-Alpha-inducing peptides is essential to the advancement of immunotherapeutic research. Conventional in vitro and in vivo techniques for researching TNF-Alpha induction are frequently labor-intensive, expensive, and time-consuming. An efficient and precise substitute for predicting TNF-Alpha-producing peptides is provided by machine learning (ML) methods. This study suggests a stacking ensemble model that considerably improves the prediction accuracy of TNF-Alpha-producing peptides by combining numerous powerful individual learners.

3.4.1 Random Forest

Using a majority voting mechanism, this ensemble technique generates forecasts using numerous decision trees. It performs exceptionally well in peptide prediction tasks due to its resilience to extreme values and capacity to manage intricate non-linear interactions. The following formula is used to construct the Gini index, a metric for assessing a dataset's impurity:

$$Gini\ index\ = \sum_{i=1}^{n} p_i^2$$

where p_i is the proportion of the i-th class label and n is the total number of class labels. The model's ability to predict TNF-Alpha-producing peptides is improved by using this index to help choose the appropriate features for data splitting within the decision trees.

3.4.2 Logistic Regression

Using peptide characteristic analysis, this conventional machine learning (ML) technique seeks to predict the likelihood of TNF-Alpha induction. Its interpretability and effectiveness are important since it can identify the primary factors causing TNF-Alpha induction. By revealing these basic features, the model helps to clarify the mechanisms behind TNF-Alpha induction and pave the way for the creation of innovative treatments that target TNF-Alpha-related pathways.

$$p(y = \frac{1}{x}) = \sigma(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots)$$

3.4.3 Support Vector Machine

This prediction model distinguishes a hyperplane that effectively divides peptides that induce TNF-Alpha from those that do not use a discriminative approach. Its ability to handle high-dimensional data and accomplish distinct class distinctions accounts for its effectiveness in peptide categorization.

$W^T * x + b = 0$

3.4.5 LightGBM

This work uses LightGBM (LGBM), a gradient-boosting technique that is more effective and scalable than XGBoost. It is perfect for processing huge datasets since it places a strong emphasis on memory optimization and quick training—two factors that are essential for predicting TNF-producing peptides. The objective function of the model is

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

where L(t) stands for the loss function and Ω(t) for the regularization function. To ensure optimal tree depth and mitigate overfitting, parameter 'C' is added as a regularization term.

3.4.6 Decision Tree

Decision tree models are used because they can divide the feature space in an organized manner, which makes it easier to find intricate patterns. Their interpretability raises the model's transparency by offering insightful information about the major variables affecting the prediction of TNF-alpha peptides. This openness facilitates the interpretation of model results and is essential for comprehending the underlying mechanisms of TNF-alpha production. In the end, this helps create effective therapies that target TNF-alpha-related pathways.

3.4.7 Stacking Classifier

Leveraging the various capabilities of base models, the Stacking Classifier shows great efficacy in capturing complex correlations within TNF-alpha peptide data. Acting as a sophisticated student, it skillfully combines knowledge from several sources to produce better prediction abilities. Through a substantial improvement in the stacking model's performance, the stacking classifier increases TNF-alpha-inducing peptide prediction accuracy and dependability.

The stacking classifier in this work uses a sequential procedure to function. First, several basic models with different techniques and strengths are trained, including decision trees, LGBM, and support vector machines (SVM). As input features for later modelling phases, these foundation models produce predictions on the dataset. Meta-learners combine predictions from base models, usually in the form of logistic regression. The Sci-Kit learn package is used in the construction of the stacking classifier, and the meta-learner is trained using both real target values and base model predictions. This provides useful prediction combination knowledge to the meta-learner. In the end, the trained stacking model outperforms individual models in terms of accuracy when making predictions on fresh TNF-alpha peptide data.

CHAPTER - 4

RESULT AND DISCUSSION

The TNF-alpha prediction model is evaluated in detail from a number of perspectives, including the use of machine learning evaluation metrics and the correction of dataset imbalances by oversampling methods such as SMOTE and ADASYN. The final portion focuses on the feature analysis results and provides an understanding of the efficacy, interpretability, and generalizability of the model.

4.1 Performance Evaluation

The assessment method is crucial to guaranteeing the effectiveness of the predictive model for TNF-alpha-producing peptides. This work builds, tests, and evaluates prediction models in accordance with accepted practices. To improve dependability, it combines an external validation technique with a strong 10-fold cross-validation method.

The area under the receiver operating characteristic curve (AUC), the Matthews correlation coefficient (MCC), sensitivity, specificity, and accuracy are the evaluation measures used. Specificity assesses the frequency of false positives in relation to true negatives, whereas sensitivity evaluates the precise identification of TNF-alpha-producing peptides. The percentage of accurate predictions is determined by accuracy, and MCC offers a thorough evaluation of rewarding classifiers that function well for both positive and negative examples.

Moreover, given that the AUC is a threshold-independent metric, it functions as a benchmark for classification models. A more accurate prediction model is indicated by a higher AUC value.

These criteria and approaches allow for a thorough performance evaluation of the TNF-alpha-inducing peptide prediction model, guaranteeing its efficacy and dependability in real-world scenarios.

$$\text{Sensitivity} = \frac{TP}{TP + FN} * 100$$

$$\text{Specificity} = \frac{\text{TN}}{\text{TN} + \text{FP}} * 100$$

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

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

Within this setting, TP stands for true positive, TN for true negative, FP for false positive, and FN for false negative events.

4.2 Best performance of Different classifiers on Imbalanced training Data set

Table 4.1: Performance comparison of diverse classifiers employing DPC features on an imbalanced dataset.

Classifier	Accuracy	MCC	AUC	Sensitivity	Specificity
LR	0.599	0.194	0.5959	0.6711	0.5207
RF	0.7144	0.4284	0.7111	0.789	0.6333
SVC	0.6124	0.2211	0.6092	0.6854	0.5331
XGB	0.698	0.3942	0.6962	0.7405	0.6519
DT	0.6327	0.2636	0.6317	0.6549	0.6085
KNN	0.6045	0.2059	0.6025	0.6492	0.5558
LGBM	0.7193	0.437	0.7175	0.7614	0.6736
Stacking	0.7307	0.4605	0.7282	0.789	0.6674

Predictive models for TNF-alpha-producing peptides were evaluated by the study using an imbalanced dataset and multiple classifiers. The Stacking Classifier performed the best out of all the models evaluated, with an accuracy of 73.07%, an MCC of 0.4605, and an AUC of 0.7282. Surprisingly, the Random Forest (RF) classifier performed admirably as well, with 71.44% accuracy, 0.4284 MCC, and 0.7111 AUC. These findings highlight the efficiency of ensemble approaches in managing unbalanced data for TNF-alpha peptide prediction, with Random Forest and Stacking Classifier appearing as the top-performing models, respectively.

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

Classifier	Accuracy	MCC	AUC	Sensitivity	Specificity
LR	0.599	0.194	0.5959	0.6711	0.5207
RF	0.7144	0.4284	0.7111	0.789	0.6333
SVC	0.6124	0.2211	0.6092	0.6854	0.5331
XGB	0.698	0.3942	0.6962	0.7405	0.6519
DT	0.6327	0.2636	0.6317	0.6549	0.6085
KNN	0.6045	0.2059	0.6025	0.6492	0.5558
LGBM	0.7193	0.437	0.7175	0.7614	0.6736
Stacking	0.7307	0.4605	0.7282	0.789	0.6674

Using an unbalanced AAC dataset to evaluate predictive models for TNF-alpha-producing peptides, the Stacking Classifier performed best, with an accuracy of 73.07%, an MCC of 0.4605, and an AUC of 0.7282. With 71.44% accuracy, an MCC of 0.4284, and an AUC of 0.7111, Random Forest trailed closely behind. Logistic regression, on the other hand, did the worst, with an MCC of 0.194 and an accuracy of 59.9%. These findings highlight how crucial it is to choose appropriate classifiers for imbalanced datasets, with ensemble approaches working well in this situation.

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

4.3: Performance comparison of diverse classifiers employing CKSAAP features on a balanced dataset.

Classifier	Accuracy	MCC	AUC	Sensitivity	Specificity
LR	0.67	0.3417	0.6697	0.7268	0.6127
RF	0.7263	0.4601	0.7258	0.8207	0.6309
SVC	0.7182	0.4366	0.718	0.7419	0.6942
XGB	0.7239	0.4482	0.7237	0.7514	0.6961
DT	0.6495	0.2994	0.6493	0.685	0.6136
KNN	0.5327	0.139	0.5349	0.1025	0.9674
LGBM	0.7172	0.4345	0.7171	0.7315	0.7028
Stacking	0.752	0.5041	0.752	0.76	0.744

In this work, we evaluated many classifiers on a balanced CKSAAP dataset to predict TNF-alpha-producing peptides. In terms of accuracy, the stacking model did the best, with 75.20%, 0.5041 MCC, and 0.7520 AUC. With over 72% accuracy, Random Forest and XGBoost also produced impressive results. In contrast, K-Nearest Neighbors performed the worst, with 0.1390 MCC and 53.27% accuracy. These results underline the shortcomings of more basic models like KNN and demonstrate the greater predictive power of ensemble techniques like stacking.

Table 4.4: Performance comparison of diverse classifiers employing DPC features on a balanced dataset.

Classifier	Accuracy	MCC	AUC	Sensitivity	Specificity
LR	0.5937	0.1874	0.5936	0.6158	0.5714
RF	0.7272	0.4562	0.727	0.7751	0.6788
SVC	0.6233	0.2465	0.6232	0.63	0.6165
XGB	0.7272	0.4551	0.7271	0.759	0.6951
DT	0.6152	0.2304	0.615	0.6404	0.5896
KNN	0.6118	0.2236	0.6118	0.6224	0.6012
LGBM	0.722	0.4448	0.7218	0.7571	0.6865
Stacking	0.7392	0.4786	0.739	0.7628	0.7152

This study assessed how well different classifiers performed in predicting TNF-alpha-producing peptides using a balanced DPC dataset. With an accuracy of 73.92%, an MCC of 0.4786, and an AUC of 0.7390, the stacking model proved to be the most effective at managing balanced datasets. Strong results were also displayed by XGBoost (XGB) and Random Forest (RF), with accuracies of 72.72% and comparable MCC values of 0.456. Statistically speaking, logistic regression (LR) did the worst, with an MCC of 0.1874 and an accuracy of 59.37%. These findings show the limitations of simpler models like logistic regression in this situation while also highlighting the efficacy of ensemble approaches, especially the stacking model, in predicting TNF-alpha-producing peptides.

4.5 Discussion

This work highlights the critical importance of feature extraction methods and various machine learning algorithms in the prediction of TNF-alpha-producing peptides using a Stacking Ensemble Model. It offers a thorough grasp of the feature space by investigating eleven strategies based on the makeup of amino acids. Robust analysis is ensured by the ensemble, which consists of LR, RF, SVC, XGB, DT, KNN, LGBM, and stacking. The stacking model is incorporated, emphasizing the use of individual algorithm strengths to improve performance. This study contributes to our understanding of peptide-induced immune responses and disease pathology by providing insightful information for improving predictive models in bioinformatics

CHAPTER - 5

CONCLUSION

This study highlights the significance of immune-suppressive peptides in the creation of subunit vaccines, as well as the pivotal function of cytokines such as TNF-alpha in modulating immunological responses. With careful training on a benchmark dataset and sophisticated feature extraction methods using the ILearnplus package in conjunction with data balancing techniques like ADASYN, the stacking model demonstrated exceptional accuracy, surpassing previous approaches. Furthermore, investigating several feature encoding techniques like AAC, TPC, APAAC, and DPC improved this model's predictive power even further. The availability of experimental data is the main limitation. To further improve and validate this model's predictive power, larger, empirically validated datasets are essential. Furthermore, more study is necessary to improve our knowledge of TNF-alpha-inducing peptides and to make predictive models like Stacking even more accurate. Essentially, this work emphasizes the importance of computational methods for peptide prediction in vaccine production and shows how useful the stacking model can be as a tool in this process. In order to advance our understanding and eventually peptide-based vaccine creation, computational biologists and experimental researchers must work together in the future.

Reference

1. Meher, P. K., Sahu, T. K., Rao, A. R., & Mohanty, D. (2017). "Predicting immune epitopes for designing peptide-based subunit vaccines and immunotherapeutic." Briefings in Bioinformatics, 18(3), 467-478.
2. Chowdhury, S. Y., Nasrin, F., & Hossain, M. (2021). "PIP-EL: A New Ensemble Learning Method for Improved Proinflammatory Peptide Predictions." Frontiers in Molecular Biosciences, 8, 683657.
3. Dhanda, S. K., et al. (2019). "Developing machine learning-based models for peptide identification and epitope prediction." Journal of Immunology Research, Article ID 9471583.
4. Ding, H., et al. (2016). "Prediction of proinflammatory peptides by integrating adaptive boosting and feature representation." Journal of Proteomics, 144, 215 224.
5. Haji Sharifi, Z., et al. (2014). "Predicting Antimicrobial Peptides Using Chou's Pseudo Amino Acid Composition and Machine Learning Models." Journal of Theoretical Biology, 341, 123-135.
6. Li, Y., et al. (2020). "iAMP-2L: A Two-Level Multi-Label Classifier for Identifying Antimicrobial Peptides and Their Functional Types." Frontiers in Microbiology, 10, 289.
7. Kumar, R., et al. (2018). "An In Silico Platform for Predicting, Screening, and Designing of Antimicrobial Peptides." Scientific Reports, 8, 15376.
8. Ding, Y., et al. (2018). "Prediction of IL-4, IL-10, and IL-17 Inducing Peptides for Vaccine Design." BMC Bioinformatics, 19, 123.
9. Chou, K. C. (2011). "Some Remarks on Predicting Multi-Label Attributes in Bioinformatics." Molecular Biosystems, 7, 775-783.
10. Tang, H., et al. (2016). "Prediction of Cell-Penetrating Peptides Using Feature Selection Techniques." Scientific Reports, 6, 32240.
11. Xu, Y., et al. (2018). "A Hybrid Machine Learning Model Integrating Random Forest and Convolutional Neural Network for Predicting Proinflammatory Cytokine Inducing Peptides." Computational Biology and Chemistry, 76,