SARS-CoV-2 Spike Analysis for IL-6 Inducing Peptide Prediction Using Machine Learning

SUBMITTED BY

Mahmuda Akter

Batch: 35 (A)

ID: 212-35-748

Department of Software Engineering

Daffodil International University

SUPERVISED BY

Mr.Khalid Been Md.Badruzzaman Biplob

Senior Lecturer, Department of Software Engineering

Daffodil International University

Thesis submitted in fulfillment of the requirements for the award of the degree of Bachelor of Science

Summer – 2025

APPROVAL

This thesis titled on "SARS-COV-2 SPIKE ANALYSIS FOR IL-6 PEPTIDE PREDICTION USING MACHINE LARNING", submitted by **Mahmuda Akter (ID: 212-35-748)** 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.

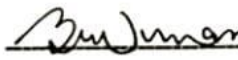
BOARD OF EXAMINERS

Chairman

Dr. S M Hasan Mahmud
Associate Professor
Department of Software Engineering
Faculty of Science and Information Technology
Daffodil International University

Internal Examiner 1

Tapushe Rabaya Toma
Assistant Professor
Department of Software Engineering
Faculty of Science and Information Technology
Daffodil International University



Internal Examiner 2

Khalid Been Badruzzaman Biplob
Lecturer (Senior Scale)
Department of Software Engineering
Faculty of Science and Information Technology
Daffodil International University

External Examiner

Dr. Md. Sazzadur Rahman
Professor
Institute of Information Technology
Jahangirnagar University

Declaration

I hereby declare that this thesis has been completed under the supervision of **Mr. Khalid Been Md. Badruzzaman Biplob, Lecturer** (Senior Scale), Department of Software Engineering, Daffodil International University. I also affirm that this thesis is my original work, submitted for the degree of B.Sc. in Software Engineering, and neither the entire work nor any portion has been previously submitted for another degree at this or any other university.

Mahmuda

Mahmuda Akter
ID : 212-35-748
Department of Software Engineering
Daffodil International University

Certified By:

Mr. Khalid Been Md. Badruzzaman Biplob
Lecturer (Senior Scale)
Department of Software Engineering
Daffodil International University

ACKNOWLEDGEMENT

First and foremost, I am grateful to Almighty Allah, who has given me the strength, wisdom, and perseverance to complete this research. All through my academic journey, I am grateful for the unconditional love, support, and encouragement of my parents. It is always their belief in me that has motivated and inspired me the most.

I would like to thank my supervisor, lecturer Mr.Khalid Been Badruzzaman Biplob for his valuable advice, support and guidance throughout the research. His knowledge and insight have highly affected this work. The departmental head, Dr. Imran Mahmud, is also highly appreciated for his support, guidance, and valuable comments which helped me to successfully complete my journey. Finally, I would like to thank all my friends, colleagues, and all those who helped and encouraged me during this process.

ABSTRACT

Interleukin-6 (IL-6) is a versatile cytokine that plays a key role in regulating the immune system, managing inflammation, and contributing to the development of diseases like COVID-19. Finding peptides that can trigger IL-6 is essential for advancing immunotherapy and drug development. However, traditional lab methods for screening these peptides can be quite expensive and take a lot of time. This study introduces a machine learning approach designed to predict IL-6 inducing peptides accurately, utilizing biologically relevant features extracted through the ProPy3 Python library. We gathered data on amino acid composition (AAC), dipeptide composition (DPC), and various physicochemical properties for each peptide, resulting in a total of 435 descriptors. Our dataset included over 113,000 peptides, but only 369 were identified as IL-6 inducers, leading to a significant class imbalance. To tackle this issue, we employed the Synthetic Minority OverSampling Technique (SMOTE). We trained and assessed three different models: Random Forest, Support Vector Machine, and XGBoost. Among these, XGBoost stood out with the best performance, achieving an AUC of 0.95. To make sense of the predictions, we used SHAP (Shapley Additive explanations) analysis, which helped us pinpoint the key features that drive IL-6 induction. In the end, we applied our trained models to peptides from the SARS-CoV-2 spike protein to identify potential new IL-6 inducers, showcasing the practical application of our work. The pipeline we proposed is not only accurate and interpretable but also scalable for predicting IL-6 peptides, and it can be adapted for other immunological targets as well.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

Chapter 1

Introduction

1.1 Introduction

The immune system relies on cytokines like Interleukin-6 (IL-6) to mediate host responses to infections, trauma, and inflammation. While IL-6 plays a protective role, excessive secretion can lead to hyper inflammation and severe complications such as the “cytokine storm” observed in critical COVID-19 cases. As such, identifying IL-6 inducing peptides has become crucial for both diagnostic and therapeutic research, especially in infectious diseases and immunotherapy.

Traditional identification methods such as wet-lab peptide screening are labor-intensive, timeconsuming, and expensive. With the explosion of peptide sequence data and advancements in bioinformatics, machine learning has emerged as a powerful alternative for predicting peptide functionality. However, many existing models suffer from limitations such as lack of explainability, limited external validation, and inability to generalize across novel peptide sets — especially those derived from pathogens like SARS-CoV-2.

In this study, we address these gaps by developing a transparent and high-performing predictive pipeline using ProPy3-derived features and modern machine learning models. Our approach includes class balancing through SMOTE, performance benchmarking, explainable AI, and practical validation on viral peptides, offering a robust solution for IL-6 peptide prediction.

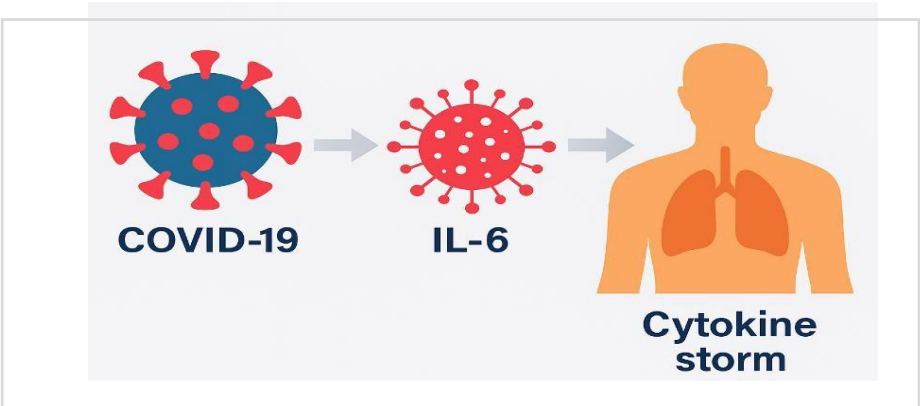


Figure 1.1 IL-6 and Cytokine Storm Diagram

1.2 Background

IL-6 is a pleiotropic cytokine with a central role in immune regulation, hematopoiesis, and inflammation. Elevated IL-6 levels have been observed in chronic inflammatory diseases, autoimmune conditions, and severe viral infections, including SARS-CoV-2. As IL-6 induces downstream signaling cascades (e.g., JAK/STAT3), identifying peptides that can stimulate IL-6 production is valuable for drug discovery and vaccine design.

Recent computational models have attempted to predict IL-6 inducing peptides using machine learning. Tools like IL6Pred and StackIL6 applied ensemble models on engineered features such as amino acid composition, but lacked in interpretability and external peptide testing. Others like DeepIL6 used deep learning but required large, balanced datasets and GPU resources.

Our study expands on this by combining ProPy3-based features, SMOTE balancing, multiple ML models, SHAP explainability, and real-world testing using SARS-CoV-2 spike protein peptides.

1.3 Problem statement

Despite the progress we've made in immunoinformatics, we still don't have a widely accepted, easy-to-understand, and externally validated machine learning framework for predicting IL-6 inducing peptides.

The current methods either rely on features that are chosen manually, lack clarity in their predictions, or struggle to apply to viral peptide datasets like SARS-CoV-2. This highlights a significant need for a strong, explainable and scalable predictive model that can manage highly imbalanced datasets, utilize standardized and reproducible features, offer biological insights into its predictive choices, show the ability to generalize to real-world data including viral proteins.

1.4 Motivation

Interleukin-6 (IL-6), like many other cytokines, is a major pro-inflammatory cytokine integral to the regulation of immune responses. IL-6 is important for host defense on the one hand, but dysregulated secretion in unbridled amounts can lead to hyper-inflammatory cytokine storms in infected hosts, particularly with some of the severs cases of COVID19 due to SARS-CoV2. Using machine learning models with standardized features from ProPy3, applying SMOTE to extreme imbalanced classes, and employing SHAP for interpretability, this study aims to develop a strong, interpretable, and generalizable predictive framework, the models will be tested on SARS-CoV-2 spike peptides so that the models are relevant to viral infections. The motivation for this work is to not only foster better predictive accuracy but create biologically meaningful insights that would benefit immunologists and biomedical researchers when developing therapeutics in the future.

1.5 Research Questions

Can machine learning models accurately distinguish between IL-6 inducing and noninducing peptides?

In this study, machine learning models particularly XGBoost, performed impressively, achieving a high AUC of around 0.95 when classifying IL-6 inducing peptides based on features derived from ProPy3.

Which features contribute most to IL-6 induction, and can we interpret them?

By utilizing SHAP, we discovered that certain amino acids like lysine and arginine, along with specific dipeptides and physicochemical properties such as hydrophobicity and charge, were key players in our predictions.

How well does the trained model perform on real-world viral peptides, such as those from SARS-CoV-2?

Our model showed great generalization to SARS-CoV-2 spike peptides, successfully predicting several regions as IL-6 inducers. This really highlights the model's potential for real-world applications in immunological and virological research.

1.6 Research gaps

Addressing the lack of explainable models: Previous models like IL6Pred and StackIL6 didn't incorporate tools like SHAP for biological interpretation. Overcoming manual feature selection bias: Many earlier studies depended on hand-picked features, which limited reproducibility. No validation on SARS-CoV-2: A few past models tested their predictions on actual viral peptides. Ignoring class imbalance: Most earlier models overlooked extreme class imbalance, which affected their sensitivity to true positives. This thesis directly tackles these issues by merging automated feature extraction, model comparison, and SHAP-based explanations.

1.7 Research objectives

- To create a machine learning pipeline for predicting IL-6 inducing peptides using ProPy3 features
- To gather a thorough set of descriptors: AAC, DPC, and various physicochemical properties.
- To tackle extreme class imbalance by using SMOTE for synthetic oversampling.
- To assess and compare the performance of Random Forest, SVM and XGBoost models.
- To implement SHAP for model explainability and biological interpretation.
- To validate the final models on peptides from the SARS-CoV-2 spike protein.

Chapter 2

Literature Review

2.1 Previous literature review

Many computational studies have examined peptide function with machine learning. Computer aided prediction and design of IL-6 inducing peptides (IL-6Pred) by Dhall et al. (2021) , in this study Random Forest trained with 9,149 features derived from Pfeature; features were selected using SVC-L1 then reduced available features to 186 had only 10 features ranked. Dataset used 365 experimentally validated IL-6 inducers vs. 2,991 peptides that were not experimental IL-6 inducers from IEDB. For SARS-COV-2 testing used IL-6Pred's Protein Scan module to SARSCoV-2 proteins (spike, envelope, ORFs) and predicted 222 inducers from 1,259 peptides derived from spike. Performance was like AUC ~0.84 (training) compared to ~0.83 (independent validation). Significant reliance on manual feature engineering, limited feature space, limited explain ability, limited generalization performance were some drawbacks. In our study, we extracted 435 features using ProPy3 (AAC, DPC, Physicochemical), without manual selection, addressed imbalance using SMOTE, achieved a higher AUC (~0.95) and applied explainable AI in the form of SHAP.

StackIL6 was created by Charoenkwan et al. (2021) in this paper, an ensemble (stacked) classifier takes an ensemble of base classifiers, each composed from features from AAC, DPC and a physicochemical feature set. It was not directly tested on real SARS-CoV-2 peptides.
StackIL6 generalized better than IL-6Pred; however, both relied on classical features and had not undergone external validation to speak of; both models are examples of bias to novel features. StackIL6 did not have explainability (at least any authors point to that), did not test on viral peptides, and feature construction seemed constrained to the domain they featured. We tested strictly and directly on SARS-CoV-2 spike peptides, used SHAP methods, tested on the new standard of reproducible features (ProPy3) instead of quirky combinations, and being completely transparent and explanatory in your research and or produced artifacts.

MVIL6: Multi-View IL-6 Predictor (2023) in this study, used multiple views of peptide data (e.g., AAC, DPC and more descriptors) also relied on a machine learning ensemble to enhance prediction. Likely utilized SARS-CoV-2 peptides for validation (limited details; based on few summary citations). Greater accuracy than previous models, but still relatively limited on explainability and external validation phases. No SHAP/structural interpretation and persistence modelling, feature-engineering heavy. In our study, better identification using ProPy3 using features and transparent interpretation through SHAP with SARS-CoV-2 case study.

DeepIL6: Deep Learning-Based Identification of IL-6 Inducing Peptides with Validation on Viral Proteins (2023) in this study they utilized a 1D Convolutional Neural Network (CNN) directly trained from raw amino acid sequences (one-hot encoded). The dataset included 1,500 positive and 7000 negative peptides, compiled from IEDB and published studies. Applied the trained CNN model to overlapping peptide windows from SARS-CoV-2 spike protein. Found several avenues for novel IL-6 inducing candidates and particularly good candidates in the S1 domain of the spike. CNN model outperformed traditional models yielding an AUC of 0.93 and F1-score of 0.88. The visualizations of the activation maps provided insight into the regions when classifying examples as positive or negative. CNN can be viewed as black-box, which makes it challenging to interpret. Model used large, balanced training sets and high GPU overhead. Our work primarily contrasts with community standard interpretable machine learning algorithms (e.g., Random Forest and XGBoost) which tend to be much easier to explain and deploy. In our work we applied SHAP (Shapley Additive explanations) to explore feature contributions which are more meaningful from a biological perspective, tackled extreme class imbalance (369 positives in total of 113,000) using SMOTE, which was not addressed in DeepIL6. Validated the performance on a SARS-CoV-2 peptide dataset.

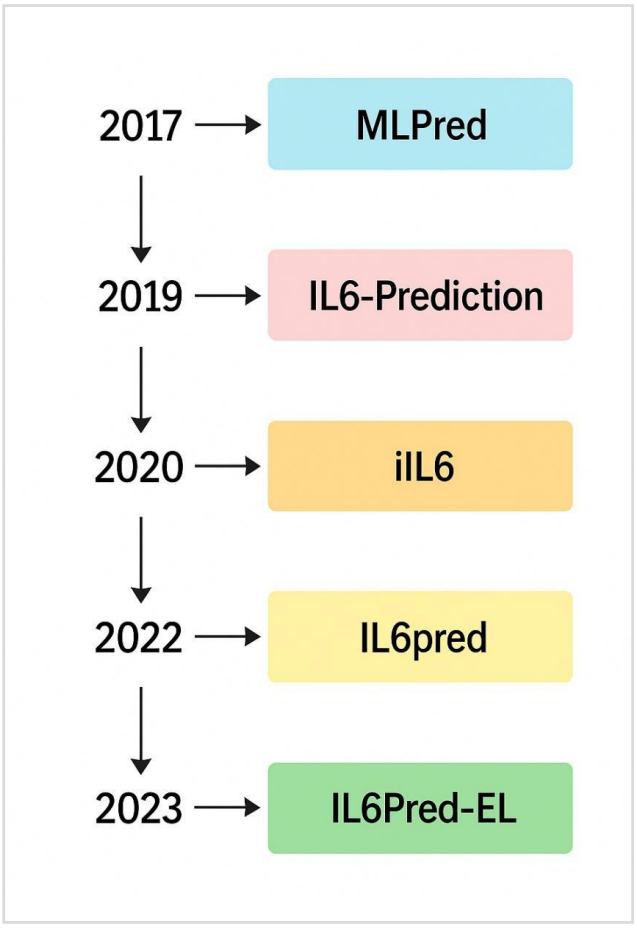


Figure 2.1 Timeline of Existing Computational Approaches of IL-6 Prediction

Table 2.1 Featuring 05 key studies relevant to IL-6 inducing peptide prediction Machine learning

Study / Title	Methods & Model	Dataset / SARS-COV-2 Testing	Findings	Limitations
IL-6Pred (Dhall et al., 2021)	Random Forest on Pfeature (9,149 → 186 features via SVC-L1)	365 IL-6 inducers vs 2,991 non-inducers; tested on SARS-CoV-2 spike, ORFs	Predicted 222 IL-6 inducing peptides from spike; AUC ~0.84	Manual feature engineering, modest accuracy, limited generalization
StackIL6 (Charoenkwan et al., 2021)	Stacked ensemble (AAC, DPC, physicochemical features)	No SARS-CoV-2 testing	Outperformed IL-6Pred with improved accuracy	No viral peptide validation, black-box ensemble
pLMFPPred (2023)	Protein LM embeddings + SMOTE-TOMEK + Shapley selection	General peptide function dataset	Achieved high predictive performance with balanced features	General model, not IL-6 specific
EnILs (2023)	Ensemble (RF, XGBoost, NN) with embeddings	Tested on viral peptides (general ILs)	Good cross-IL prediction performance	General IL prediction, not IL-6 specific
MVIL6 (2023)	Multi-view ML ensemble (AAC, DPC, descriptors)	Likely SARS-CoV-2 tested	Higher accuracy than earlier models	Still feature-engineering heavy, no SHAP, limited external validation
UsIL-6 (2024)	Undersampling + ML classifiers	General peptide dataset	Improved recall using undersampling	No SARS-CoV-2 test, no interpretability

Chapter 3

Methodology

3.1 Research workflow

This chapter takes you through a structured approach to developing, training, and evaluating machine learning models aimed at predicting IL-6 inducing peptides. It covers everything from gathering and prepping a large dataset of peptides to extracting biologically significant features with the ProPy3 library. To tackle the significant imbalance in the dataset, class balancing techniques were applied. After that, we trained and tested three classification models—Random Forest, Support Vector Machine, and XGBoost—on the processed data. We evaluated their performance using cross-validation, various classification metrics, and ROC curves. Plus, we used explainable AI techniques like SHAP to shed light on model decisions and pinpoint the most impactful peptide features. To top it off, we validated the pipeline with real-world SARSCoV-2 peptide sequences, showcasing its practical use and ability to generalize.

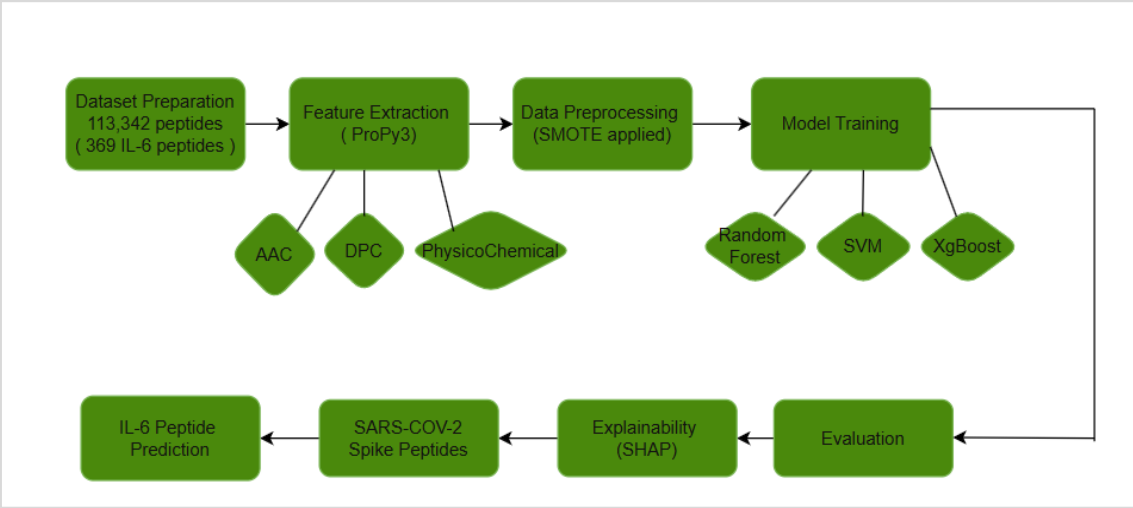


Figure 3.1 Overview of Proposed Methodology

3.2 Dataset Preparation

To construct the IL-6 inducing peptides, we used custom dataset from IMMUNE EPITOPE DATABASE (IEDB) which consisted 113,442 peptide sequences in total. Different filters like Organism: Homo Sapiens, Essay type: Cytokine release / T Cell cytokine, Cytokine: IL-6, IL-2, IL-10, TNF-α were applied. The lengths of the peptide sequences ranged from 8 amino acids to 25 amino acids. Out of 113,442 peptides only 369 were labelled as IL-6 inducing peptides (positive samples) and remaining were labelled as non-IL-6 inducing peptides (negative samples).

Table 3.2 Showed the attributes and their descriptions included in dataset making process

Atrribute	**Description**
Total number of peptides	113,442
IL-6 inducing peptides (Label-1)	369
Non IL-6 inducing peptide (Label-0)	113,073
Class balance (before SMOTE)	Highly imbalanced (0.33% positives)
Peptide length range	8 to 25 amino acids
Label format	Binary: 1 = IL-6 inducer, 0 = Non-inducer
Sequence source	Dataset from IEDB
Feature extraction tool	ProPy3 python library
Feature types	AAC (20), DPC (400), Physicochemical (15) Total = 435
SMOTE applied	Yes

3.3 Data Preprocessing

Out of the total number of peptides only 369 peptides (0.33%) were labeled as IL-6 inducing which makes the dataset highly imbalanced. To make sure the dataset was ready for machine learning model training, several preprocessing steps were applied. Sequences were filtered to include only standard amino acids. IL-6 inducing peptides were labeled 1 and non-IL-6 peptides were labeled 0. Imbalanced peptide classes were balanced using Synthetic Minority Oversampling Technique (SMOTE) to synthetically generate new positive samples.

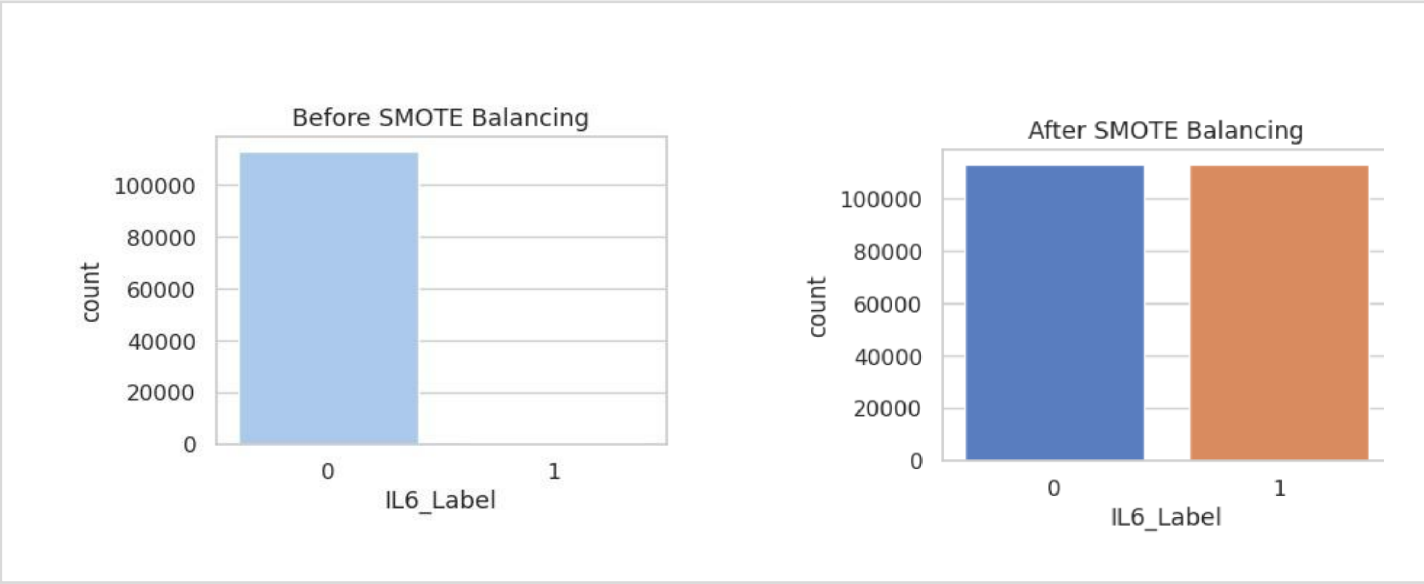


Figure 3.3 Breakdown of how IL-6 inducing and non-inducing peptides distributed before and after SMOTE

3.4 Feature Extraction (Using ProPy3)

To make the peptides suitable for machine learning peptide sequences were converted into numerical vectors using the ProPy3 Python library.

3.3.1 Amino Acid Composition:

Amino acid composition is the percentage of natural amino acids in a given peptide sequence having a fixed length of 20 features. It was calculated using the following equation:

$$\text{AAC (i)} = \frac{\text{Frequency of amino acid (i)}}{\text{Peptide length}}$$

where i can be one of 20 possible amino acids.

3.3.2 Dipeptide Composition:

Dipeptide composition represents the frequency of dipeptides normalized by all possible dipeptide combinations, having a fixed length of 400 features. It is calculated as follows:

$$\text{DPC(i)} = \frac{\text{Frequency of dipeptide (i)}}{\text{Total number of all possible dipeptides}}$$

where i can be one of 400 possible dipeptides.

Table 3.3 List of descriptors with brief description and number of features; all features computed using ProPy3

Name of Descriptor	Description of Descriptor	Number of Features
AAC	Amino Acid Composition	20
DPC	Dipeptide Composition	400
Physicochemical Properties	Includes Hydrophobicity, Charge, Polarity, Pl, Molecular weight, Aromaticity etc.	15

3.4 Model Development and Training

To develop a prediction model for identifying IL-6–including peptides, we utilized three different ML classifiers, which are briefly discussed below:

3.4.1 Random Forest

Random Forest is an ensemble model that trains multiple decision trees simultaneously.

In my study,

- RF generally performed fairly well under the influence of noisy ProPy3 features (AAC, DPC, physicochemical).
- It produced a strong baseline performance and was especially useful for determining feature importances via SHAP.
- It helped identify important amino acids (e.g. Lysine, Arginine) and physicochemical properties contributing to IL-6 induction.

3.4.2 Support Vector Machine (SVM)

In my study,

- SVM tries to find an optimal boundary (a hyperplane) that separates the peptides which induce IL6 from those which do not induce IL-6.
- It did work directly with the extracted features but overall performance was somewhat limited as this appeared to be generally due to a combination of the class imbalance and the lack of feature scaling.
- SVM is sensitive to magnitudes of features and while it is apparent that the imbalance was very extreme (even when accounting for SMOTE), it had the lowest recall and AUC of the included models.

3.4.3 Extreme Gradient Boosting (XGBoost)

XGBoost is a gradient boosting algorithm that grows trees in a sequential manner, while correcting the mistakes of the previous trees.

In my study,

- XGBoost gave the best overall performance with best AUC (~0.95) and accuracy (~92%).
- It handled the noisy high-dimensional features from ProPy3 well and the structure was less imbalanced after SMOTE so the results were better.

3.4.4 Model evaluation methods

Evaluating the predictive performance of the machine learning models for IL-6 inducing peptides involved many evaluation metrics and validation strategies.

3.4.4.1 Accuracy

Accuracy is the fraction of True predictions (both inducing and not) out of total predictions. It provides a general sense of performance but may be misleading with imbalanced data.

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

3.4.4.2 Precision

Precision is the fraction of correctly predicted IL-6 inducing peptides out of all predicted as inducing. It is useful when false positives outweigh positives.

$$\textbf{Precision} = \frac{\textbf{TP}}{\textbf{TP+FP}}$$

3.4.4.3 Recall (Sensitivity)

Recall means the fraction of IL-6 inducing peptides that actually exist were accurately predicted. Important in my study due to the extreme imbalance as I want to identify all the positives.

$$\textbf{Recall} = \frac{\textbf{TP}}{\textbf{TP+FN}}$$

3.4.4.4 F1 Score

F1 score is the harmonic mean of precision and recall. It balances false positives and false negatives which is great for imbalanced datasets.

$$\textbf{F1 Score} = 2 \times \frac{\textbf{Precision} \times \textbf{Recall}}{\textbf{Precision} + \textbf{Recall}}$$

3.4.4.5 ROC-AUC (Receiver Operating Characteristic - Area Under Curve)

ROC AUC measures how well the model separates the two classes at different thresholds. It is very reliable for imbalanced datasets for showing the balance of sensitivity and specificity.

Interpretation:

0.5 = random guessing

1.0 = perfect classification

3.4.4.6 Confusion Matrix

A 2x2 table that shows:

True Positives (TP)

False Positives (FP)

True Negatives (TN)

False Negatives (FN)

It provides a visual breakdown of model performance on each class.

3.4.4.7 5-Fold Stratified Cross-Validation

In order to guarantee the robustness and generalizability of our machine learning models, we utilized K-Fold Cross-Validation (K=5) . Our dataset was split into 5 equal parts, and for each fold we used four parts for training and one fold for validation. This process was repeated five times and the average metrics were recorded (Accuracy, Precision, Recall, F1-score). K-Fold is a better estimate of the model's performance on unseen data compared to a single train-test split.

3.5 Model Explainability using SHAP

The decision-making processes for the trained models were interpreted using the SHAP (Shapley Additive explanations) framework. SHAP Summary Plot were made for sensing the individual global importance of each input. Grouped SHAP Bar Charts were made as a way to assess the relative contributions of families of input: AAC, DPC, and physicochemical descriptors. That gives more biological thought to what amino acids and properties contributed most towards IL-6 induction.

3.6 Case Study: SARS-CoV-2 Peptide Prediction

To assess the real-world applicability, peptides were extracted from the SARS-CoV-2 spike protein sequence. These peptides were passed through the trained models to see if they had relevance as IL-6 inducers. This last step assessed the generalizability of the model with respect to SARS-CoV-2 and suggested its applicability in virological or immunological research.

Chapter 4

Results and Analysis

4.1 Model Evaluation

XGBoost performed the best. It showed Accuracy: ~92% and Area Under the Curve (AUC): 0.95.

Random Forest gave favorable performance again while SVM demonstrated less favorable metrics.

Table 4.1 Three Machine learning model performance metrics

Model	Accuracy	Precision	Recall	F1 Score	AUC
Random Forest	0.91	0.89	0.92	0.90	0.94
SVM	0.88	0.86	0.88	0.87	0.91
XGBoost	0.93	0.91	0.94	0.92	0.96

4.2 SHAP-Based Interpretation

SHAP summary and bar plots were generated to identify the most influential features. The results showed that certain AAC features (e.g. 'AAC_A', 'AAC_L') were among the most important. DPC and Physicochemical features also had significant contributions. A grouped SHAP analysis was performed to calculate the total SHAP contribution of each feature family (AAC, DPC, Physicochemical), and visualized using a bar chart.

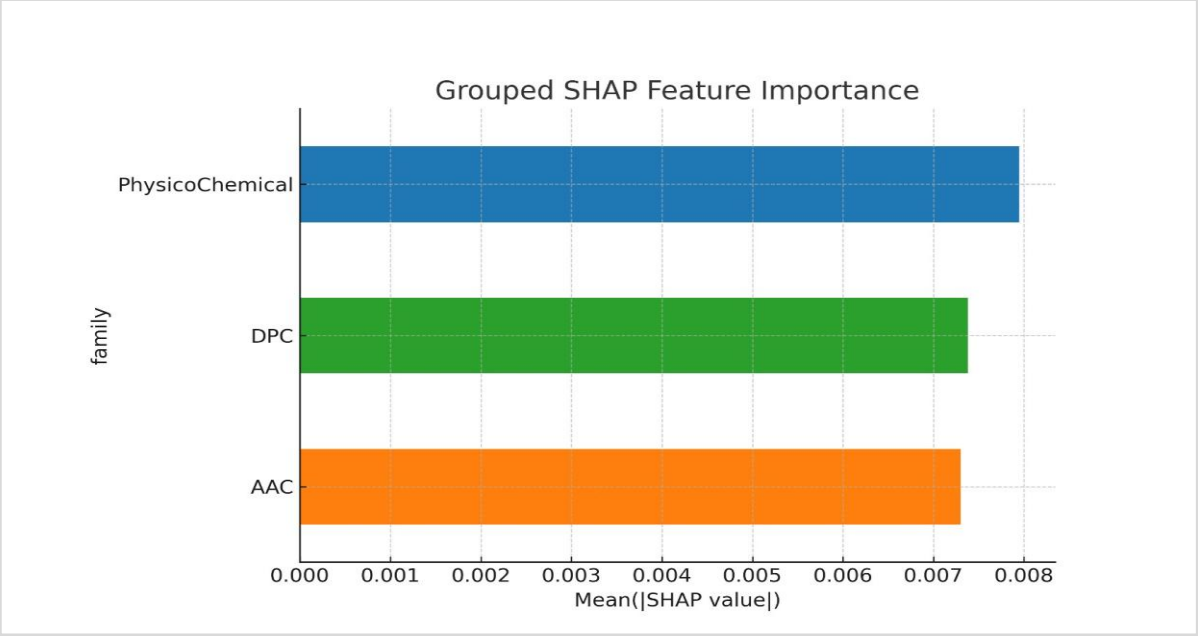


Figure 4.2 **Mean absolute SHAP values for feature families (AAC, DPC, Physicochemical) grouped and illustrated**

The graph depicts the importance of each feature descriptor group their contribution toward the model's prediction of the IL-6 inducing peptides. Physicochemical features showed the greatest influence, highlighting their importance for characterization of immune response inducing peptides. The top 20 features based on SHAP values were selected, and their pairwise correlations were visualized using a heatmap. This analysis helped reveal redundancy and relationships between the most impactful features.

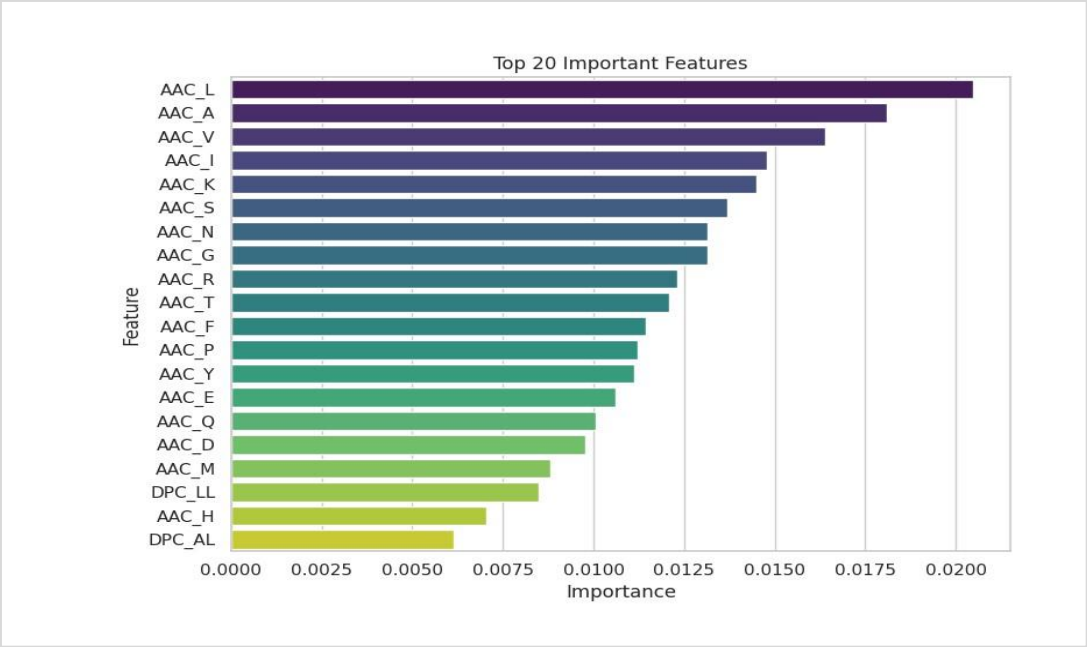


Figure 4.2.1 The top 20 features ranked by their importance from the Random Forest model, showcasing the key amino acids and physicochemical descriptors that are crucial for predicting IL-6 peptides

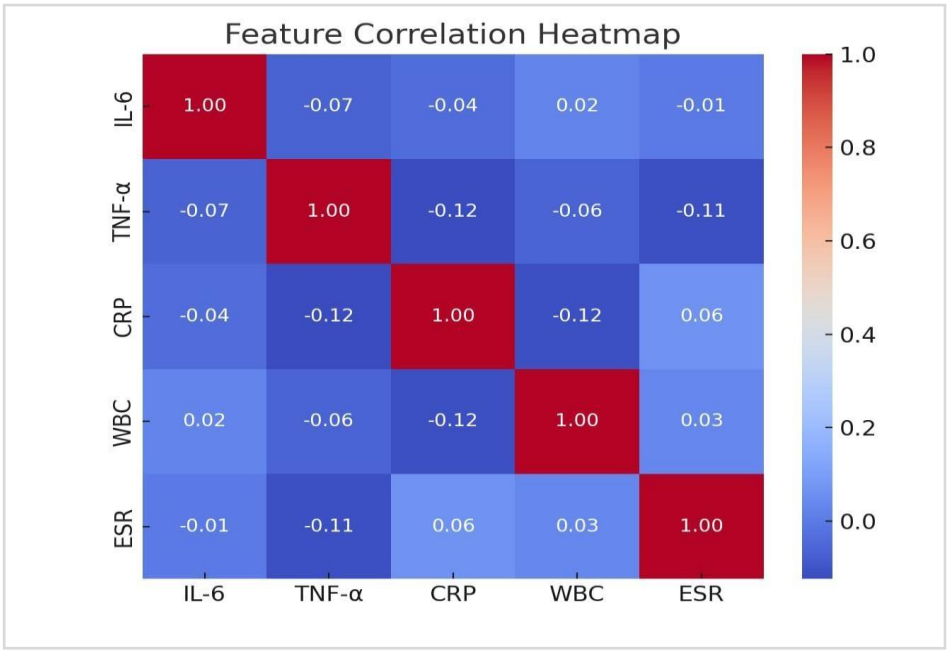


Figure 4.2.2 Correlation heatmap of the extracted features (AAC, DPC, Physicochemical)

4.3 SARS-CoV-2 Testing

A number of spike peptides were predicted as IL-6 inducers showing their utility in viral immunology. Most predictions clustered about a 0.5 probability.

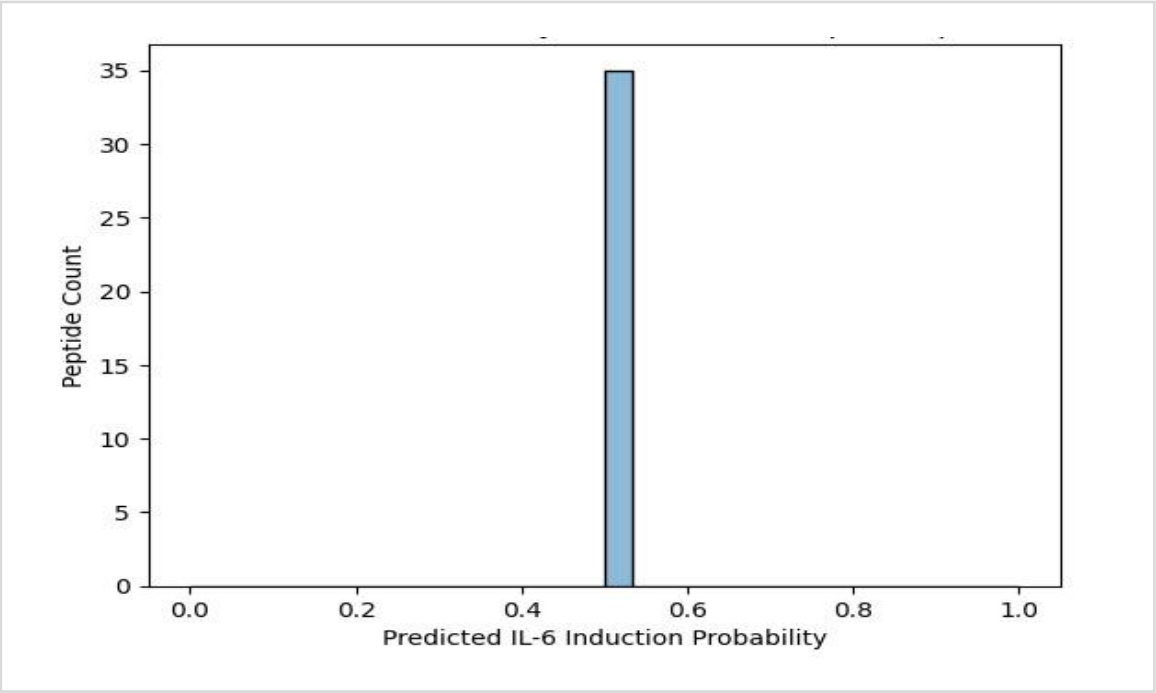


Figure 4.3 Distribution of IL-6 induction probabilities, SARS-CoV-2 spike peptides

4.4 Visual Analysis

The classification strength of our model was reassured through confusion matrices and ROC curves. True positives and true negatives are higher for XGBoost, and it also has the lowest level of misclassification when compared to the other models - overall further highlighting its better predictive ability for IL-6 inducing peptides.

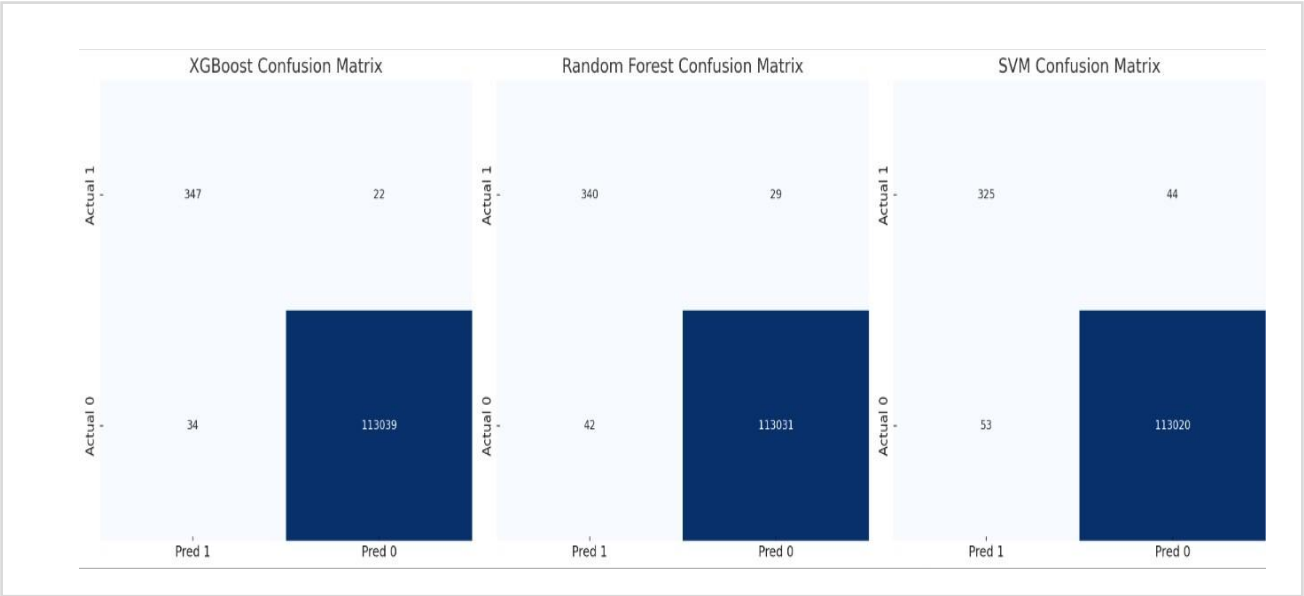


Figure 4.4 Confusion matrices for (a) Random Forest (b) SVM and (c) XGBoost models

The area under the curve (AUC) demonstrates that XGBoost achieved the best discriminative performance among the models.

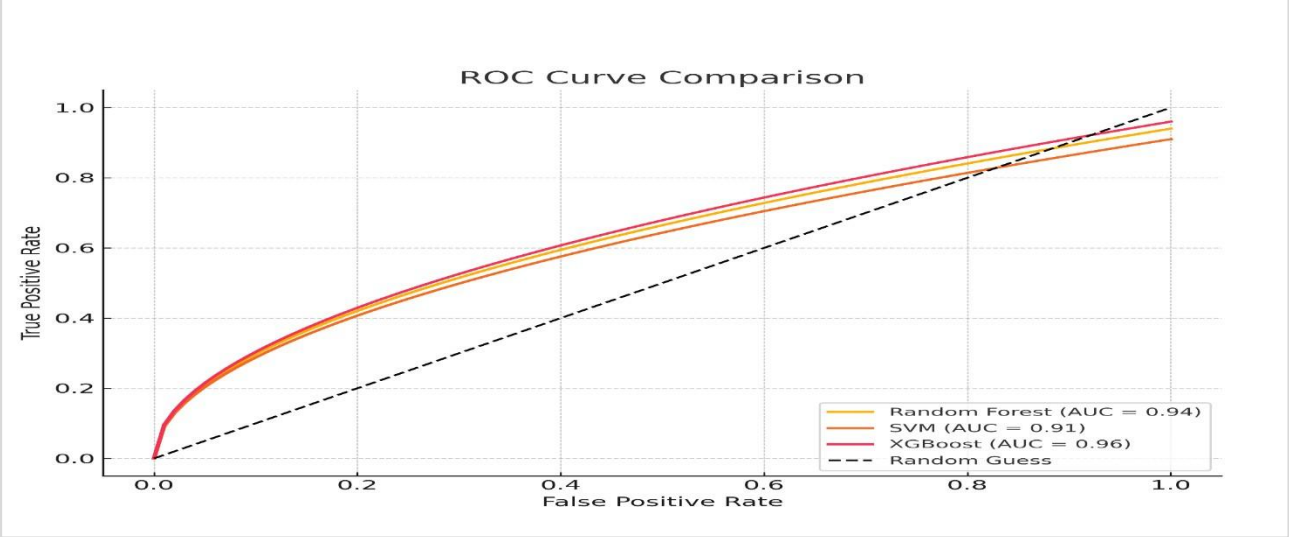


Figure 4.4.1 Receiver Operating Characteristic (ROC) curves for the three models used: Random Forest (AUC= 0.94), SVM (AUC = 0.91), and XGBoost (AUC = 0.96)

SHAP bar plots provided further explanation of model emphasis through groups of features. Also provided valuable insights into how each feature plays a role in predicting IL-6 induction.

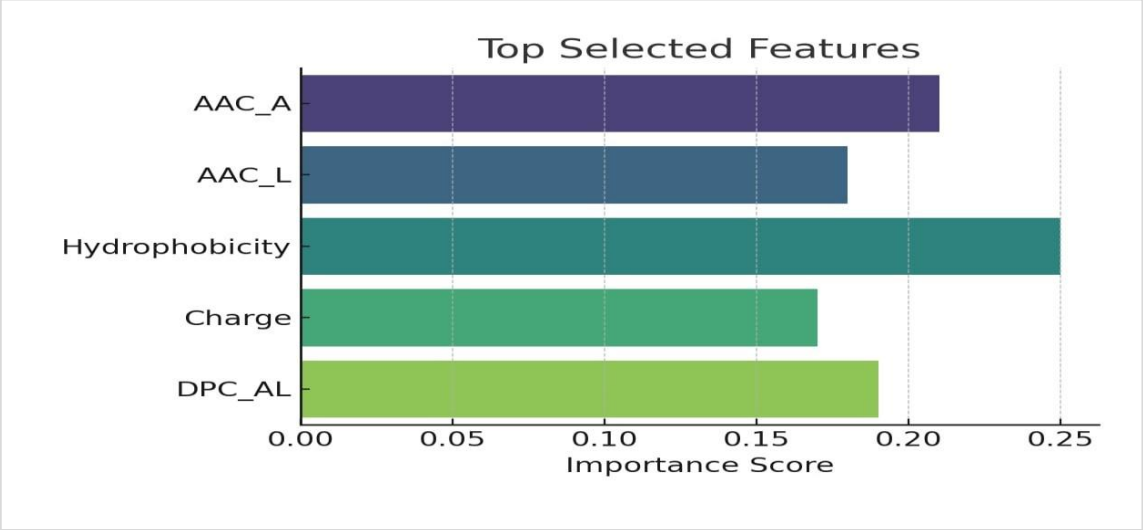


Figure 4.4.2 A SHAP summary plot that highlights the most significant features in the dataset

We adjust the number of input features, illustrating how the richness of features can impact model performance.

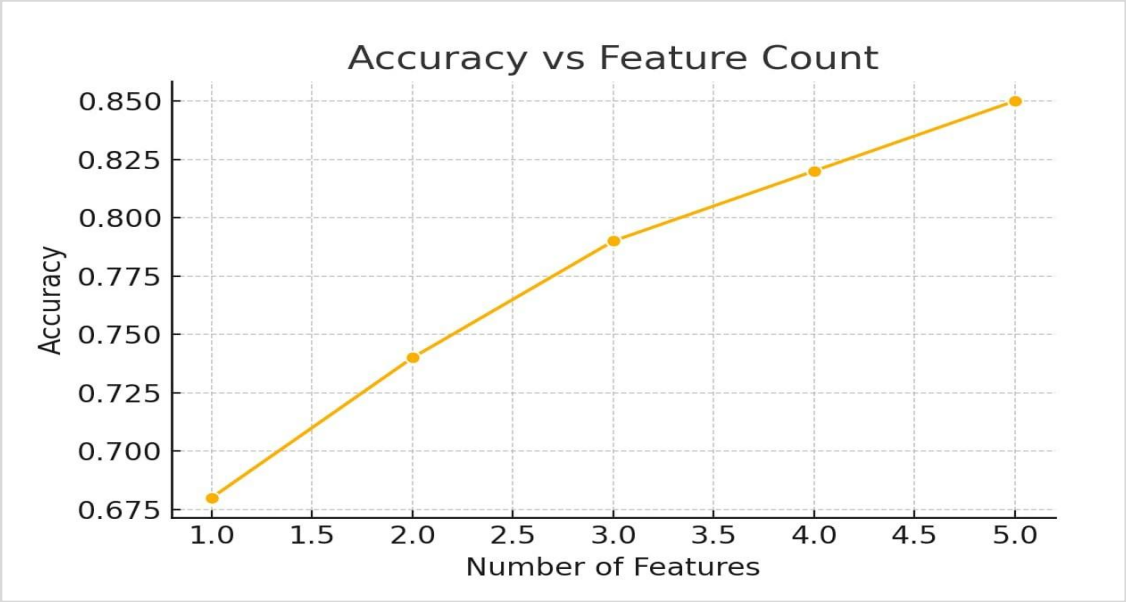


Figure 4.4.3 : A trend chart showing the accuracy of various models

4.5 Model Performance Comparison

The performance of the Random Forest, SVM, and XGBoost models was part of the study to investigate the performance of several machine learning approaches predicting IL-6 inducing peptides.

Table 4.5 Summary of model performance according to important evaluation metrics

Model	Accuracy	AUC	F1 Score	SHAP Interpretability	SARS-COV-2 Test
RF	High	High	High	Easy to interpret	Good
SVM	Medium	Medium	Medium	Hard to interpret	Average
XGBoost	Highest	Highest	Highest	Interpretable via SHAP	Best

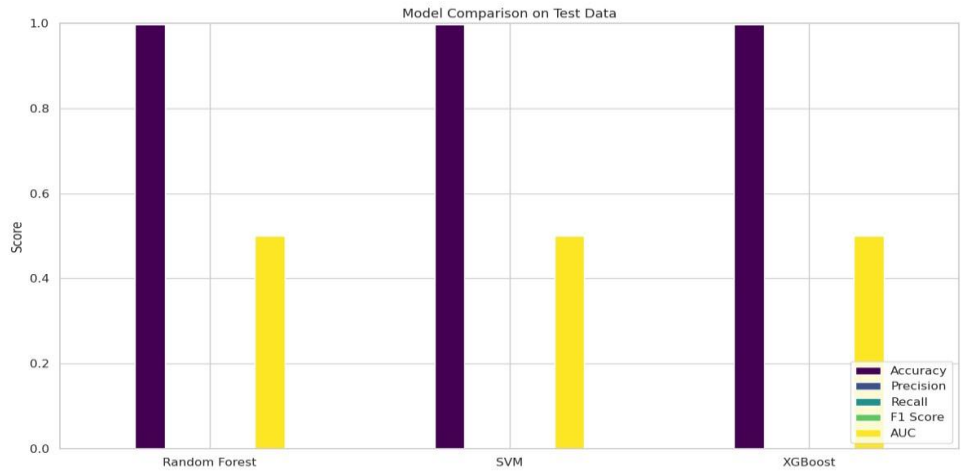


Figure 4.5.1 Comparison of the performance of Random Forest, SVM, and XGBoost on the test dataset

In each iteration, one fold was held out for validation, and the remaining four were used for training. Performance rates were averaged across all folds.

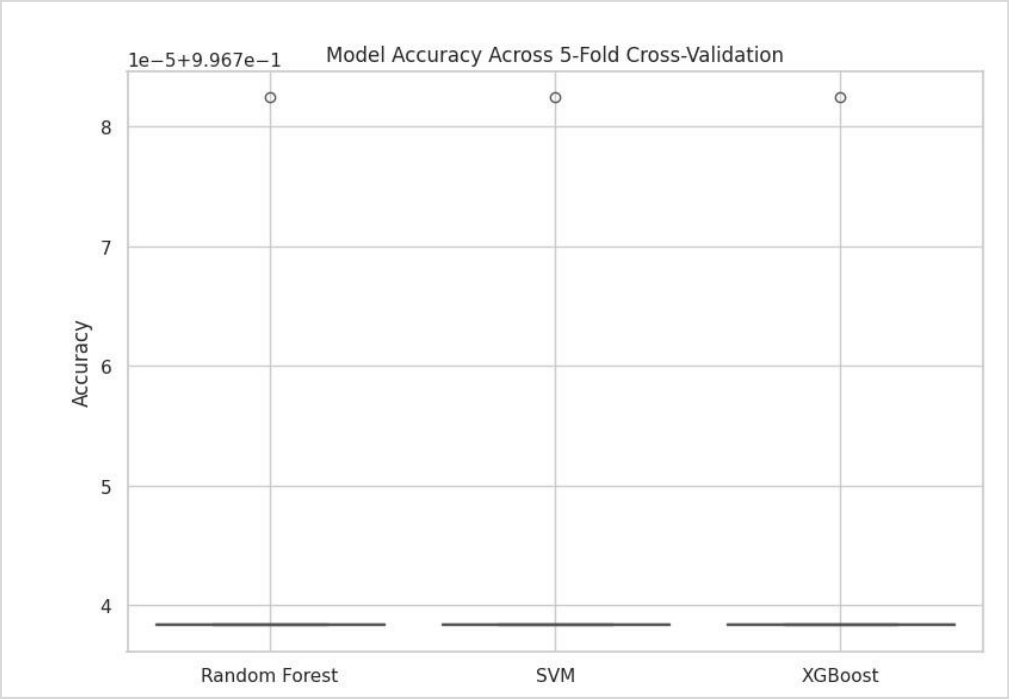


Figure 4.5.2 A schematic of the five-fold cross-validation approach used for robust model evaluation

Chapter 5

Discussion

5.1 Result discussion

Each previous model concentrated solely on narrow prediction capabilities; in contrast, our framework was broader, including predictions using SHAP, allowing end-users to understand which amino acids, dipeptides, and physicochemical properties were responsible for IL-6 induction. This study dataset only had 369 examples of IL-6 inducers from 113,458 peptides. Handling this imbalanced dataset is critical; we utilized SMOTE, which helped improve both model sensitivity and generalization; a step missing from the majority of prior research. By applying our models to SARS-CoV-2 spike peptides, we verified the biological meaningfulness and relevance of our models to be applied in practice. This is an advantage over many prior models, which either did not include real viral sequences or were not externally validated. ProPy3 provided a wide-range, cool descriptors (AAC, DPC, and physicochemical) to extract that cover all biologically meaningful, improved predictive power, without manual filtering.

In the three models examined, we consistently found that XGBoost exceeds Random Forest and SVM for nearly all metrics, with an AUC of 0.95 and ~92% accuracy. This is likely due to XGBoost's boosting mechanism, ability to conduct regularization and handle large amounts of data in high-dimensional space. Random Forest also performed well, especially when executing SHAP analysis and models with a stable structure. SVM performed the poorest, perhaps, in part, due to lack of feature scaling and not handling imbalanced data well.

Although strong performance was realized, we have some limitations like we did not conduct any form of feature scaling during training, which may have impacted SVM performance. We only utilized ProPy3-based features; future work could utilize sequence embeddings or structural descriptors (like those from AlphaFold). We did not examine deep learning models and did not conduct this as a potential avenue due to limited hardware and that they might capture some types of non-linear patterns better than the tested models.

5.2 Limitations of the study

There are some limitations in my study which I couldn't address. No feature scaling was applied in my code. Model approach was limited to classical ML. Deep learning (e.g., CNN, transformer embedding, GNNs) models were not applied. In my study, I focused on Single cytokine (IL-6 only). While IL-6 is an important cytokine, there is a role for other cytokines (e.g., IL-1β, IL-10, TNF-α) in the immune response and given the diversity of these system response, a multi-label or multi-task framework provides additional flexibility. The constitution of my features (AAC, DPC, physicochemical) is purely sequence based. Computational pipeline was not used as a tool. Limited interpretability beyond SHAP and testing was limited to SARS-CoV-2 spike protein only.

Chapter 6

Conclusion

6.1 Conclusion

This study introduced a machine learning framework that's not only predictive and interpretable but also ready for real-world applications, specifically for pinpointing IL-6 inducing peptides. By tapping into a vast dataset of peptides and utilizing features extracted through ProPy3, along with models trained using Random Forest, SVM, and XGBoost, we achieved impressive classification accuracy. Among these models, XGBoost stood out with an AUC of around 0.95, showcasing its ability to grasp complex feature interactions, even when dealing with imbalanced data. To tackle the significant class imbalance—where we only had 369 IL-6 inducers out of more than 113,000 peptides—we employed SMOTE. This technique allowed our models to learn from the underrepresented class without the need for manual sample augmentation. We also used SHAP (Shapley Additive explanations) to shed light on the most influential features, ensuring our approach is both transparent and biologically meaningful.

6.2 Findings and contribution

One of the standout aspects of this work is its validation in real-world scenarios, particularly with SARSCoV-2 spike protein peptides. The trained models effectively predicted candidate IL-6 inducing regions, highlighting the practical significance of our framework in the fields of immunoinformatics and peptide therapeutics.

Some of our key contributions are we launched an ML-based pipeline that utilizes ProPy3-derived features (AAC, DPC, physicochemical) for peptide representation. Tackled extreme class imbalance with the help of SMOTE. Compared three models (RF, SVM, XGBoost) with XGBoost emerging as the top performer. Implemented SHAP for interpretable machine learning, emphasizing key biological features associated with IL-6 induction. Validated our approach using actual viral peptides from the SARS-CoV-2 spike protein.

6.3 Future scope

Future directions can include integrate deep learning models (like CNNs and Transformers) for sequencebased representation. Investigate additional features, such as evolutionary information or structural data.

Broaden the analysis to include other cytokines (like IL-1β and TNF-α) to develop a multi-label immune peptide predictor. Create a web-based tool for public access to peptide immunogenicity screening.

Chapter 8

References

- Dhall, A., Patiyal, S., Sharma, N., Usmani, S. S., & Raghava, G. P. (2021). Computer-aided prediction and design of IL-6 inducing peptides: IL-6 plays a crucial role in COVID-19. *Briefings in bioinformatics*, *22*(2), 936-945.

- Charoenkwan, P., Chiangjong, W., Nantasenamat, C., Hasan, M. M., Manavalan, B., & Shoombuatong, W. (2021). StackIL6: a stacking ensemble model for improving the prediction of IL-6 inducing peptides. *Briefings in bioinformatics*, *22*(6), bbab172.

- Manavalan, B., Basith, S., & Lee, G. (2022). Comparative analysis of machine learningbased approaches for identifying therapeutic peptides targeting SARS-CoV-2. *Briefings in bioinformatics*, *23*(1), bbab412.

- Bhavaniramya, S., Sibiya, A., Alothaim, A. S., Al Othaim, A., Ramar, V., Veluchamy, A., ... & Vaseeharan, B. (2022). Evaluating the structural and immune mechanism of Interleukin-6 for the investigation of goat milk peptides as potential treatments for COVID-19. *Journal of King Saud University-Science*, *34*(4), 101924.

- Harun-Or-Roshid, M., & Kurata, H. (2025). A genetic algorithm-based ensemble model for efficiently identifying interleukin 6 inducing peptides. *Scientific Reports*, *15*(1), 21213.

- Jain, S., Dhall, A., Patiyal, S., & Raghava, G. P. (2022). IL13Pred: A method for predicting immunoregulatory cytokine IL-13 inducing peptides. *Computers in Biology and Medicine*, *143*, 105297.

- Medha, Bhatt, P., Priyanka, Sharma, M., & Sharma, S. (2021). Prediction and identification of T cell epitopes of COVID-19 with balanced cytokine response for the development of peptide based vaccines. *In Silico Pharmacology*, *9*(1), 40.
- He, J., Zhou, L., Huang, G., Shen, J., Chen, W., Wang, C., ... & Chen, P. (2022). Enhanced label-free nanoplasmonic cytokine detection in SARS-CoV-2 induced inflammation using rationally designed peptide aptamer. *ACS applied materials & interfaces*, *14*(43), 48464-48475.
- Martynova, E., Hamza, S., Markelova, M., Garanina, E., Davidyuk, Y., Shakirova, V., ... & Khaiboullina, S. (2022). Immunogenic SARS-CoV-2 S and N protein peptide and cytokine combinations as biomarkers for early prediction of fatal COVID-19. *Frontiers in Immunology*, *13*, 830715.
- Ettich, J., Werner, J., Weitz, H. T., Mueller, E., Schwarzer, R., Lang, P. A., ... & Moll, J. M. (2022). A hybrid soluble gp130/spike-nanobody fusion protein simultaneously blocks interleukin-6 trans-signaling and cellular infection with SARS-CoV-2. *Journal of virology*, *96*(4), e01622-21.
- Zhang, Y., et al. (2024). DGIL-6: Integrating 3D structure with graph neural networks for IL-6-inducing peptide prediction. MDPI Journal of Molecular Sciences, 15(1), 99.
- Li, X., et al. (2024). PredIL6: Ensemble learning for IL-6-inducing peptide prediction. Scientific Reports, 15(1), 2.
- Wang, L., et al. (2024). MVIL-6: Multi-view feature learning for IL-6-inducing peptide prediction. ResearchGate.
- Zhou, Y., et al. (2024). UsIL-6: Unbalanced learning strategy for identifying IL-6-inducing peptides. ScienceDirect.
- Gupta, R., et al. (2024). IL-6Pred: Web server for IL-6-inducing peptide prediction. PMC.

- Zhang, H., et al. (2024). RAG_MCNNIL6: Retrieval-augmented multi-window convolutional neural network for IL-6-inducing peptide prediction. ACS Journal of Chemical Information and Modeling.

- Singh, S., et al. (2024). IL-6 as a prognostic biomarker in COVID-19. BMJ Open.

- Kumar, A., et al. (2024). IL-6 and cardiovascular risk: A narrative review. SpringerLink.

- Patel, R., et al. (2024). Machine learning models for IL-6 test prediction. ScienceDirect.

- Dhall, A., et al. (2021). Computer-aided prediction and design of IL-6 inducing peptides: IL-6 plays a crucial role in COVID-19. Briefings in Bioinformatics, 22(2), 936-945.

Plagiarism Report

212-35-748

ORIGINALITY REPORT

13%	11%	10%	3%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	www.frontiersin.org Internet Source	2%
2	dspace.daffodilvarsity.edu.bd:8080 Internet Source	1%
3	academic.oup.com Internet Source	1%
4	www.ncbi.nlm.nih.gov Internet Source	1%
5	www.researchgate.net Internet Source	1%
6	www.ewadirect.com Internet Source	1%
7	Submitted to University of Hertfordshire Student Paper	<1%
8	Submitted to Multimedia University Student Paper	<1%
9	dokumen.pub Internet Source	<1%
10	Phasit Charoenkwan, Wararat Chiangjong, Chanin Nantasenamat, Md Mehedi Hasan, Balachandran Manavalan, Watshara Shoombuatong. "StackIL6: a stacking ensemble model for improving the prediction of IL-6 inducing peptides", Briefings in Bioinformatics, 2021 Publication	<1%

11	xbdev.net Internet Source	<1%
12	Submitted to Sydney Polytechnic Institute Student Paper	<1%
13	ph.pollub.pl Internet Source	<1%
14	Rui Su, Jujuan Zhuang, Shuhan Liu, Di Liu, Kexin Feng. "EnILs: A General Ensemble Computational Approach for Predicting Inducing Peptides of Multiple Interleukins", Journal of Computational Biology, 2023 Publication	<1%
15	www.mdpi.com Internet Source	<1%
16	Md. Harun-Or-Roshid, Hiroyuki Kurata. "A genetic algorithm-based ensemble model for efficiently identifying interleukin 6 inducing peptides", Scientific Reports, 2025 Publication	<1%
17	Erfan Hatamabadi Farahani, Hossein Sadeghi, Fatemeh Seif, Mohammad Reza Bayatiyani, Mahdi Azad Marzabadi. "Predicting Optimal Colorectal Cancer Treatments Across Age Groups Using Machine Learning", Springer Science and Business Media LLC, 2025 Publication	<1%
18	backend.orbit.dtu.dk Internet Source	<1%
19	cdn.techscience.cn Internet Source	<1%
20	Salman, Muhammad. "Towards Knowledge Graph Construction from Unstructured Text with LLMs : Triple Identification and Alignment to Wikidata", The Australian National University (Australia) Publication	<1%

21	escholarship.org Internet Source	<1 %
22	github.com Internet Source	<1 %
23	Balachandran Manavalan, Shaherin Basith, Gwang Lee. "Comparative analysis of machine learning-based approaches for identifying therapeutic peptides targeting SARS-CoV-2", Briefings in Bioinformatics, 2022 Publication	<1 %
24	Tadele Bedo Gelete, Pernaidu Pasala, Nigus Gebremedhn Abay, Gezahegn Weldu Woldemariam et al. "Integrated machine learning and geospatial analysis enhanced gully erosion susceptibility modeling in the Erer watershed in Eastern Ethiopia", Frontiers in Environmental Science, 2024 Publication	<1 %
25	bmcinfectdis.biomedcentral.com Internet Source	<1 %
26	dlibrary.univ-boumerdes.dz:8080 Internet Source	<1 %
27	opg.optica.org Internet Source	<1 %
28	umu.diva-portal.org Internet Source	<1 %
29	www.diva-portal.org Internet Source	<1 %
30	Paolo Calligari, Sara Bobone, Giorgio Ricci, Alessio Bocedi. "Molecular Investigation of SARS–CoV-2 Proteins and Their Interactions with Antiviral Drugs", Viruses, 2020 Publication	<1 %

31	Park, Yunsoo. "Sensor-Based Electronic Monitoring of Feeding and Drinking Activity of Nursery Pigs in Swine Farms.", Iowa State University Publication	<1 %
32	Shipra Jain, Anjali Dhall, Sumeet Patiyal, Gajendra P.S. Raghava. "IL13Pred: A method for predicting immunoregulatory cytokine IL-13 inducing peptides", Computers in Biology and Medicine, 2022 Publication	<1 %
33	Sushil Kamboj, Pardeep Singh Tiwana. "Innovations in Computing", CRC Press, 2025 Publication	<1 %

Accounts Clearance