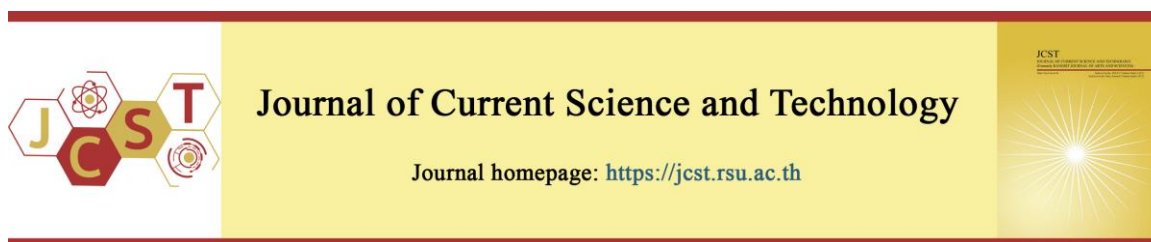


Cite this article: Nikom, J., Shoombuatong, W., Charoenkwan, P., & Musikasuwan, S. (2026). C400KNN: Predictive model for enhancing anti-HIV peptide classification. *Journal of Current Science and Technology*, 16(3), Article 191. <https://doi.org/10.59796/jcst.V16N2.2026.191>



C400KNN: Predictive Model for Enhancing Anti-HIV Peptide Classification

Jaru Nikom¹, ORCID: 0009-0002-8147-6782
Watshara Shoombuatong², ORCID: 0000-0002-3394-8709
Phasit Charoenkwan³, ORCID: 0000-0002-5656-2796
Salang Musikasuwan^{1,*}, ORCID: 0000-0002-8733-9626

¹Department of Mathematics and Computer Science, Faculty of Science and Technology, Prince of Songkla University, Pattani Campus, Pattani 94000, Thailand

²Center for Research Innovation and Biomedical Informatics, Faculty of Medical Technology, Mahidol University, Bangkok 10700, Thailand

³Modern Management and Information Technology, College of Arts, Media and Technology, Chiang Mai University, Chiang Mai 50200, Thailand

*Corresponding author; E-mail: salang.m@psu.ac.th

Received 30 September 2025; Revised 10 November 2025; Accepted 18 March 2026; Published online 21 July 2026

Abstract

The design of peptide sequences to perform anti-HIV activities is a particularly time-consuming stage in the production of AIDS medications. One way to solve this issue is by computer modeling. Predictive models for anti-HIV peptides help decrease the time and cost of producing anti-HIV peptide drugs. This paper introduces the C400KNN model, developed by a machine learning approach that uses non-antimicrobial peptides as the negative dataset and incorporates a feature selection procedure of 400 variables using the chi-square test, followed by training with the KNN algorithm. The amino acid sequences were gathered from databases containing anti-HIV, antiviral, and antimicrobial information. Next, all features were extracted from 12 descriptors, encompassing both the chemical structure and physicochemical properties. Ten classifiers were used to create the models. The models were evaluated to discover descriptors using AUC and ACC. Then, the features of those descriptors were combined, and feature selection methods were used to choose the most important features. The best model was selected based on its performance. The findings indicate that the descriptor groups that improve model efficiency are AAC, DPC, PAAC, and APAAC. The Chi-square method was used in the feature selection step. Additionally, the accuracy of the model was found to be 0.83. The C400KNN model is thought to be effective and might help researchers in creating anti-HIV peptides for use in pharmaceutical manufacturing.

Keywords: *predictive model, anti-HIV peptide, machine learning, chi-square*

1. Introduction

HIV is a pathogen that weakens the body's immune system, leading to AIDS. It inhibits the body's ability to defend against diseases, making it more vulnerable to various ailments. A highly effective approach for eliminating HIV is peptide therapy (Lama & Planelles, 2007). Most anti-HIV peptide drugs consist of structural proteins that bind to the glycoprotein 120 (gp120) of the HIV spike and the glycoprotein 41 (gp41) of the CD4 cell membrane

surface (Kwong et al., 1998; Prabakaran et al., 2007; Sharma et al., 2021). This attachment prevents HIV from binding to the membrane. In the early stages of HIV drug development, anti-HIV peptides are usually selected randomly from different sources such as plants, animals, or even the virus itself. (Greenberg & Cammack, 2004; Mori et al., 2005; Wang et al., 2004). The data have been collected and stored in a database from several studies conducted several decades ago. These datasets were generated in several laboratories

by empirical study (Steffler et al., 2021). Many databases used for modeling are focused on antiviral peptides and include data on their effectiveness against HIV. Examples of such databases include AVPPred (Thakur et al., 2012), AntiVPP 1.0 (Lissabet et al., 2019), PEPred-Suite (Wei et al., 2019), and Meta-iAVP (Schaduangrat et al., 2019). However, there are only a few databases dedicated to anti-HIV peptides.

Medical applications are currently using these datasets and several machine learning algorithms to build prediction models of peptide bioactivity (Deng et al., 2023; Liu et al., 2020; Xu et al., 2021). These approaches have been shown to increase model performance by improving the modeling process, creating a validated database of models, categorizing peptides based on features and learning data, maximizing model efficiency, analyzing significant characteristics, and developing a user-friendly web server for categorization that does not require complex statistics or mathematics (Charoenkwan et al., 2020). Antiviral peptide models commonly include prediction models for peptides that are effective against HIV. Hence, the accuracy of the model relies on the predictive accuracy of the antiviral peptide rather than the anti-HIV peptide. Currently, there exists only one published database specifically dedicated to anti-HIV peptides. The laboratory has confirmed the HIPdb database as one of the mentioned databases (Qureshi et al., 2013). Previous studies that used this database investigated numerous methods for developing prediction models for anti-HIV peptides. They applied the pseudo-amino acid composition (PseAAC) concept of Chou et al. (2012) (Esmaceli et al., 2010) and a multilayer perceptron with the support vector machine (SVM) method (Poorinmohammad & Mohabatkar, 2015). A study of prediction models employing rough-set-based methods was later published. The goal was to increase the accuracy of predicting anti-HIV type 1 peptides to 80% by using feature selection and an improved

Harmony search algorithm (Mathiyazhagan et al., 2022). This study found that feature selection strategies were used to increase the model's prediction performance. Furthermore, the learning method chosen is significant since it has a direct impact on the model's architecture, accuracy, scalability, and ability to handle complex data. Different methods are appropriate for various prediction tasks (Hatta et al., 2024). Table 1 summarizes examples of model features and strategies for classifying anti-HIV peptides using machine learning.

Figure 1 describes the process for constructing the models, which classify peptides with anti-HIV bioactivity into two major groups. The first step is to evaluate the baseline model's performance on two study datasets: the main dataset and the alternate dataset. We collected these datasets from five internet sources and classified them into three categories: anti-HIV peptides represent positive information, while the other two negative groups are non-antiviral and non-antimicrobial. The sequence of each group of amino acids was then confirmed for redundancy manually using the CD-HIT program (Huang et al., 2010; Li & Godzik, 2006). We then extracted the features of all 12 descriptors using both the main data (positive + non-antiviral) and the alternate data (positive + non-antimicrobial). Furthermore, the information in each descriptor was used to build models for all ten algorithms. We then assessed the model's performance and selected it for the next step. In the second step, the selected data sets from the previous step were extracted, along with the effective descriptor groups, to create new data groups. Next, the data collected were applied to select features through the process of chi-square, ANOVA, and mutual information with varying numbers of features: 20, 50, 100, 200, and 400. Next, we applied these features to construct models using various classifiers and evaluated the performance of each model. Our goal is to find the best model for classifying anti-HIV peptides.

Table 1 Summary of research findings on machine-learning models used for anti-HIV peptide classification

Model	Feature selection	Classifier	Model's performance
Poorinmohammad & Mohabatkar, 2015	-	Multilayer perceptron	ACC = 84%
			ACC= 85%
Poorinmohammad et al., 2015	-	Support vector machine	SEN=90%
			SEP=80%
			MCC=70%
Mathiyazhagan et al., 2022	Rough-set-based & Harmony search methodology	Rough-set-based & Harmony search methodology	ACC = 80%

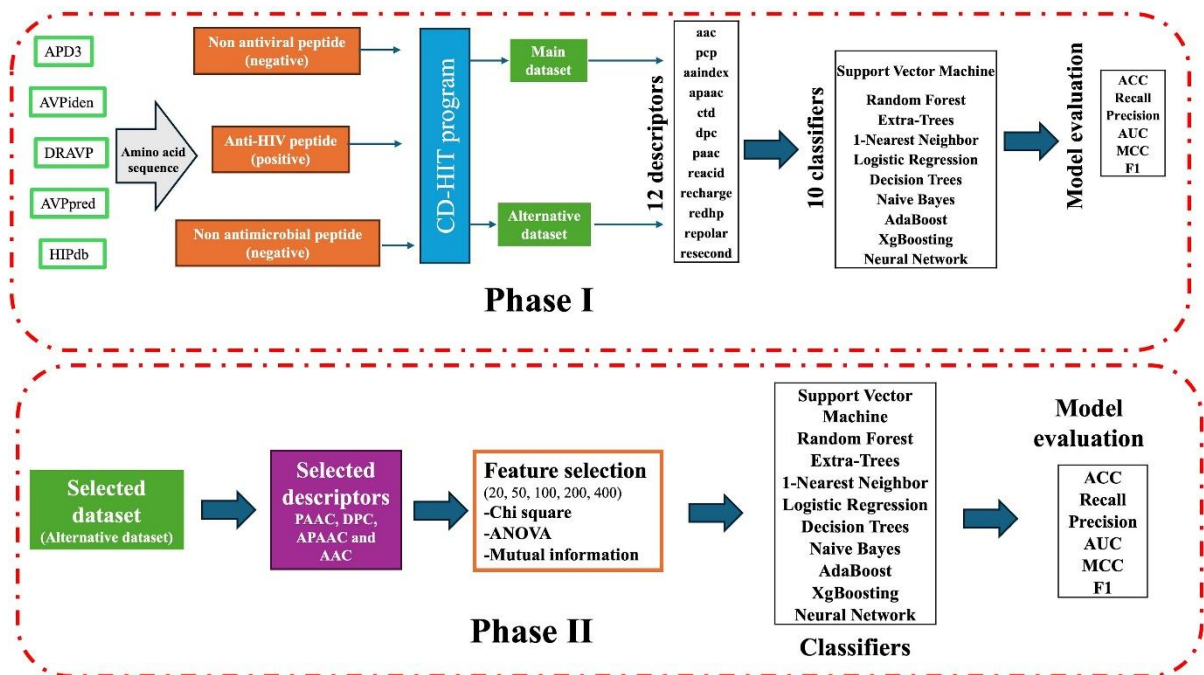


Figure 1 The model construction process for anti-HIV peptide classification

2. Objectives

The major objective of this research was to create a model for classifying peptides with anti-HIV bioactivity by collecting data from the anti-HIV peptide database. A peptide prediction model was developed by selecting the appropriate features and leveraging feature information from the amino acid sequence. Finally, an approach that enables the machine to learn efficiently was selected, and the model was evaluated to determine its performance. According to this predictive model, scientists will require a reduced amount of time to develop highly effective drugs against HIV. Furthermore, we investigated the impact of changes in data structures (part of the negative dataset) on model performance using the biometric distance concept.

3. Materials and Methods

3.1 Dataset Collection and Preparation

In this study, datasets were collected from several sources. Amino acid sequences were collected from five online databases: APD3 (Wang et al., 2016),

AVPiden (Pang et al., 2021), DRAVP (Liu et al., 2023), AVPpred (Thakur et al., 2012), and HIPdb (Qureshi et al., 2013). There were 6,032 peptides that were anti-HIV, non-antiviral, and non-antimicrobial. Furthermore, the dataset was separated into two classes: anti-HIV peptides and non-anti-HIV peptides. These data were used in a supervised machine learning system to categorize peptides that have anti-HIV activity.

This study identified differences in species distances in a negative dataset that influenced the machine learning model. Thus, in the negative data, two distinct groups of peptides with neither antiviral nor antibacterial activity were identified. The non-antiviral dataset was labeled as the main dataset. Additionally, datasets that do not include non-antiviral datasets were classified as alternatives. Then, we employed manual screening and CD-HIT software to decrease biological and amino acid sequence redundancy by computing all data (main and alternate datasets).

Table 2 Sources of data used in this study and the number of amino acid sequences with both anti-HIV and non-anti-HIV activity (accessed on Jan 13, 2025)

Source	Anti-HIV	Non-Anti-HIV	URL	Citation
AVPpred	241	1,056	http://crdd.osdd.net/servers/avppred/	Thakur et al., 2012
HIPdb	981	179	http://crdd.osdd.net/servers/hipdb/	Qureshi et al., 2013
APD3	105	-	https://aps.unmc.edu	Wang et al., 2016
AVPIDen	867	1,884	https://awi.cuhk.edu.cn/AVPIDen/#/	Pang et al., 2021
DRAVP	719	-	http://dravp.cpu-bioinfor.org	Liu et al., 2023
Total	2,913	3,119		

Table 3 The number of extracted features from 12 descriptors

Descriptor	Properties	Description	Number of features
AAC	Chemical Structure	Amino acids composition	20
APAAC		Amphiphilic pseudo-amino acid composition	22
CTD		Composition, transition, and distribution	147
DPC		Dipeptides composition	400
PAAC		Pseudo-amino acid composition	50
AAindex	Physicochemical Properties	Amino acid index database	531
PCP		Biochemical and biophysical properties	21
reAcid		Reduced amino acid sequences according to the acidity	32
reCharge		Reduced amino acid sequences according to the charge	50
reDHP		Reduced amino acid sequences according to the DHP	32
rePolar		Reduced amino acid sequences according to polarity	32
reSecond		Reduced amino acid sequences according to the secondary structure	40

3.2 Selection of the Dataset for Predictive Model Construction

The main and alternative datasets were transformed from the amino acid sequences, which were originally represented by letter symbols, into numerical data that reflected the chemical structure and physicochemical properties: therefore, a specific approach was necessary. This study used 12 categories of descriptors: Amino Acid Composition (AAC), DiPeptide Composition (DPC), Composition/Transition/Distribution (CTD), Pseudo-Amino Acid Composition (PAAC), Amphiphilic Pseudo-Amino Acid Composition (APAAC), Amino Acid index (AAindex), Physicochemical property (PCP), Reduced Acid (reAcid), Reduced Polar (rePolar), Reduced Charge (reCharge), Reduced Second (reSecond), and Reduced DHP (reDHP) (Charoenkwan et al., 2023). The calculated peptide sequence descriptors will be used to build a model. Finally, the model was assessed, and the insights derived from all models were compared. Next, we selected the dataset with the greatest performance model (ACC and AUC) and use it to construct a baseline dataset for future study in the following phase.

3.3 Features Selection for Model Construction

In this process, an algorithm to identify suitable characteristics for the data was employed, with the goal of creating a high-performance model. In the first step, all features from the candidate descriptors were incorporated into the baseline dataset, and the most effective features were selected using the feature selection algorithms. Three feature selection techniques consisting of mutual information, ANOVA, and chi-square testing have been used in this study. Feature selection was carried out by keeping the top k ranked features, with k set to 20, 50, 100, 200, and 400, to examine the impact of feature dimensionality on model performance. These values, which range from compact to more comprehensive feature representations, were selected to reflect gradually rising feature subset sizes. This approach makes it possible to systematically assess how model complexity, information retention, and prediction performance are traded off. The mathematical formulas of the feature selection algorithms used in this study are described as follows.

1) Mutual Information

Mutual information measures the statistical dependency between a feature variable X and the

class label Y , quantifying how much information one variable provides about the other. It is defined as:

$$MI(X;Y) = \sum_{x \in X} \sum_{y \in Y} p(x,y) \log \frac{p(x,y)}{p(x)p(y)} \quad (1)$$

where $p(x,y)$ denotes the joint probability distribution of X and Y , and $p(x)$ and $p(y)$ represent their marginal probability distributions. Features with higher MI values are considered more informative for classification.

2) Analysis of Variance (ANOVA)

ANOVA assesses whether the mean values of a continuous feature differ significantly among multiple classes by comparing between-group and within-group variances. The F-statistic is defined as:

$$F = \frac{\sum_{k=1}^K n_k (\bar{x}_k - \bar{x})^2 / (K-1)}{\sum_{k=1}^K \sum_{i=1}^{n_k} (x_{ik} - \bar{x}_k)^2 / (N-K)} \quad (2)$$

where K is the number of classes, n_k is the number of samples in class k , $\bar{x}_k$ is the mean of class k , $\bar{x}$ is the overall mean, and N is the total number of samples.

3) Chi-square Test

The chi-square test evaluates the degree of association between discrete feature variables and class labels by comparing observed and expected frequencies. The chi-square statistic is calculated as:

$$\chi^2 = \sum_{i=1}^n \frac{(O_i - E_i)^2}{E_i} \quad (3)$$

where O_i and E_i are the observed and expected frequencies, respectively. A larger χ^2 value indicates a stronger dependency between the feature and the target class.

3.4 Construction of Models using Multiple Machine Learning Algorithms

Supervised machine learning refers to the process of obtaining knowledge via the application of various algorithms. In the first stage, the data were separated into two sets: 80% for training and 20% for testing. Subsequently, all features were extracted from both datasets, and the data was normalized using the MinMaxScaler method. In the model training process, a technique known as 10-fold cross-validation was employed. The peptide classifiers were constructed using 10 alternative model algorithms, including Support Vector Machine (SVM), Random Forest (RF), Extra Tree (ET), K-Nearest Neighbor (KNN), Logistic Regression (LR), Decision Tree (DT), Naïve Bayes (NB), AdaBoost (AB), XGBoost (XGB), and Neural Network (NN). Table 4 shows the algorithm parameters that were adjusted to obtain optimal tuning performance using the grid search method. Furthermore, testing data is defined as data that the model had never encountered or been trained on previously.

3.5 Evaluation of the Models

In this study, to evaluate the models' performance for predicting the peptide anti-HIV peptide properties, six evaluation methods were applied which consist of precision, F1 score (F1), area under the curve (AUC), Recall, accuracy (ACC), and Matthews' correlation coefficient (MCC). For this work, Editplus (version 5.3.0, Korea) was used to manage the amino acid sequence. The iFeature package (Chen et al., 2018) was used to extract the necessary features. Scikit-learn version 1.4.1 (Pedregosa et al., 2011) was compatible with the system and was used to create Python code for the Kaggle (Quaranta et al., 2021) online platform.

Table 4 Grid search-tuned hyperparameters for model construction

Algorithm	Parameters
SVM	{'C': [0.1, 1, 10], 'gamma': [0.1, 1, 10], 'kernel': ['linear', 'rbf']}
RF	{'n_estimators': [50, 100, 200], 'max_depth': [None, 5, 10]}
ET	{'n_estimators': [50, 100, 200], 'max_depth': [None, 5, 10]}
KNN	{'n_neighbors': [1, 3, 5]}
LR	{'C': [0.1, 1, 10]}
DT	{'max_depth': [None, 5, 10, 20], 'min_samples_split': [2, 5, 10]}
NB	()
AB	{'n_estimators': [50, 100, 200], 'learning_rate': [0.01, 0.1, 1]}
XGB	{'n_estimators': [100, 200]}
NN	{'hidden_layer_sizes': [(100, (50, 100, 50)), 'alpha': [0.0001, 0.001, 0.01], 'max_iter': [50, 100]}

$$ACC = \frac{TP + TN}{(TP + TN + FP + FN)} \quad (4)$$

$$recall = \frac{TP}{TP + FN} \quad (5)$$

$$precision = \frac{TP}{TP + FP} \quad (6)$$

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \quad (7)$$

$$F1 \text{ score} = \frac{TP}{TP + \frac{1}{2}(FP + FN)} \quad (8)$$

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

The area under the ROC curve (AUC) was also used to evaluate the model's performance by plotting the True Positive Rate (TPR) against the False Positive Rate and calculating the area under the graph.

4. Results

4.1 Descriptive Analysis

High-quality data are crucial for enhancing the efficiency of classification modeling, as the model learns from the data (Charoenkwan et al., 2021). A thorough literature review conducted at the start of January 2024 identified the HIPdb database as one valuable resource for classification models of anti-HIV peptides. Recently, the researchers used data from databases such as AVPpred and APD3 (Wang et al., 2016), which specifically focused on the antiviral and antibacterial properties. These databases contain data specifically related to anti-HIV peptides (Poorinmohammad & Mohabatkar, 2015). This study used data from various databases, including the anti-

HIV database (HIPdb), the antiviral databases (AVPpred, AVPI den, and DRAVP), and the antimicrobial database (APD3). This dataset will serve as a backup. Similar to data science applications, these data were applied to create a machine learning model that classifies peptides according to their anti-HIV properties. Our goal in this analysis was to reduce the data in this study by simplifying and eliminating redundant amino acid sequences. Table 5 shows the biomolecular redundancy identified using the CD-HIT program.

A total of 6,032 peptides were initially curated, followed by the removal of manual duplicates. Subsequent processing using the CD-HIT program yielded 5,328 non-redundant peptides, which were classified as follows: in the main group data, there were 2,584 positive or anti-HIV peptides that went through the CD-HIT program with a 100% threshold. There were also 565 cases of negative data or data without AVP properties and 2,179 cases of alternative group-negative data or data without AMP properties. The table displays the results. At the 90% threshold, 624 anti-HIV peptides remained. The negative data, which met these criteria for an antiviral peptide (AVP), had a value of 479. The alternative group-negative data for an antimicrobial peptide (AMP) is 1,488. The number of anti-HIV data points was 452. There were 405 negatives. Furthermore, setting the threshold at 80% resulted in 1,281 negative data. We conducted a detailed analysis of the amino acid sequences of peptides from anti-HIV peptides. The analysis showed that the positive-frequency dataset had higher average values for V, I, and L, which are all hydrophobic amino acids, while M had the lowest frequency, as shown in Figure 2.

Table 5 Amino acid sequence data in the process of data deduplication using the CD-HIT program at various threshold levels, where non-AVP is non-antiviral and non-AMP is non-antimicrobial

Class	main			alternative		
	total	anti-HIV	non-AVP	total	anti-HIV	non-AMP
Original	3,149	2,584	565	4,763	2,584	2,179
90%	1,103	624	479	2,112	624	1,488
80%	857	452	405	1,733	452	1,281

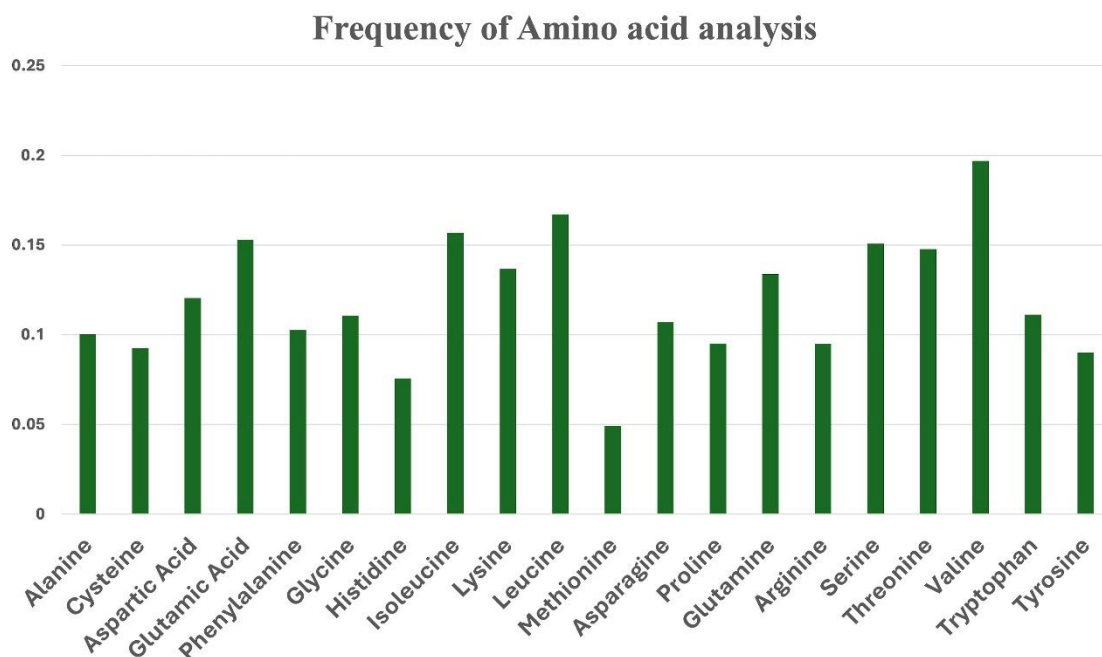


Figure 2 Frequency of Amino acid analysis in the dataset, where V=Valine, L=Leucine, I=Isoleucine, E=Glutamic Acid, S=Serine, T=Threonine, K=Lysine, Q=Glutamine, D=Aspartic Acid, W=Tryptophan, G=Glycine, N=Asparagine, F=Phenylalanine, A=Alanine, P=Proline, R=Arginine, C=Cysteine, Y =Tyrosine, H=Histidine, M=Methionine

4.2 Evaluating the Predictive Models

The machine learning approach partitioned the data into two subsets: 80% for training and 20% for testing. The training dataset was used for constructing and training predictive models. The testing dataset was used to evaluate the effectiveness of the models when they encountered new and unknown data, this process called independent testing. Each feature was used to construct a model. Subsequently, the training data were employed to extract the characteristics of the peptide. This study involved extracting the characteristics of 12 descriptors: AAC, DPC, PAAC, APAAC, CTD, reduced acid, reduced polar, reduced charge, reduced second, reduced DHP, PCP, and AAINDEX. In this study, ten machine learning methods, consisting of SVM, RF, ET, KNN, LR, DT, NB, AB, XGB, and NN were applied. We then incorporated the model into the model assessment process to test it using metrics such as ACC, Recall, Precision, AUC, MCC, and F1 score. Subsequently, the remaining 20% of the data underwent testing.

The performance of the baseline model was examined by comparing its ACC and AUC values on both the training and independent testing datasets.

Figure 3 displays the performance of the baseline model created using the main dataset. The AUC value for the cross validation (CV) dataset was 0.706, while the independent testing (IN) dataset had a higher value of 0.735. The CV and IN datasets had the highest values of ACC values, measuring 0.707 and 0.737, respectively.

The AUC and ACC values of the CV and IN datasets for the models constructed from the alternative datasets with 12 descriptors were analyzed and presented in Figure 4. The AUC value for the CV dataset was 0.707, whereas the IN dataset had the highest AUC value of 0.719. The ACC value reached its peak at 0.811 and 0.812 for the CV and IN datasets, respectively. Both the CV and IN groups *showed* that the alternative *dataset* had higher values than the main dataset. In addition, when examining the features that impact the modeling of anti-HIV peptide classification, it can be observed that AAC, APAAC, DPC, and PAAC exhibit ACC values of 0.800 or higher. These descriptors will be used to extract information from the alternative dataset. This was used to generate a new dataset, aiming to construct a more efficient model.

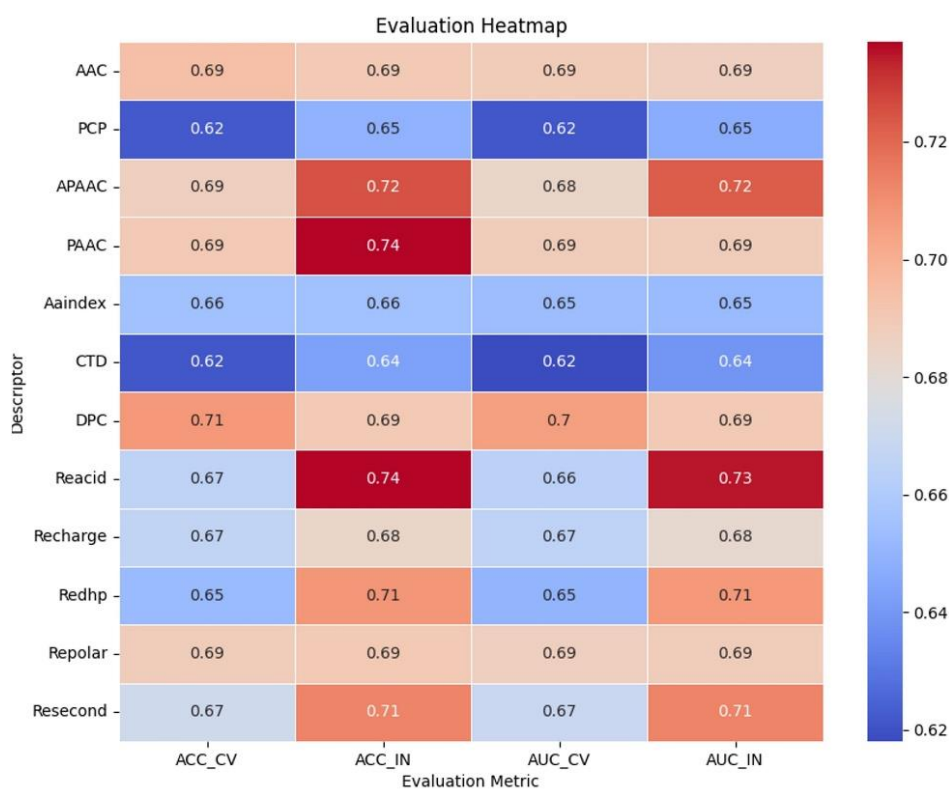


Figure 3 The ACC and AUC values of the model from the CV and IN datasets when comparing the 12 descriptors using the main dataset

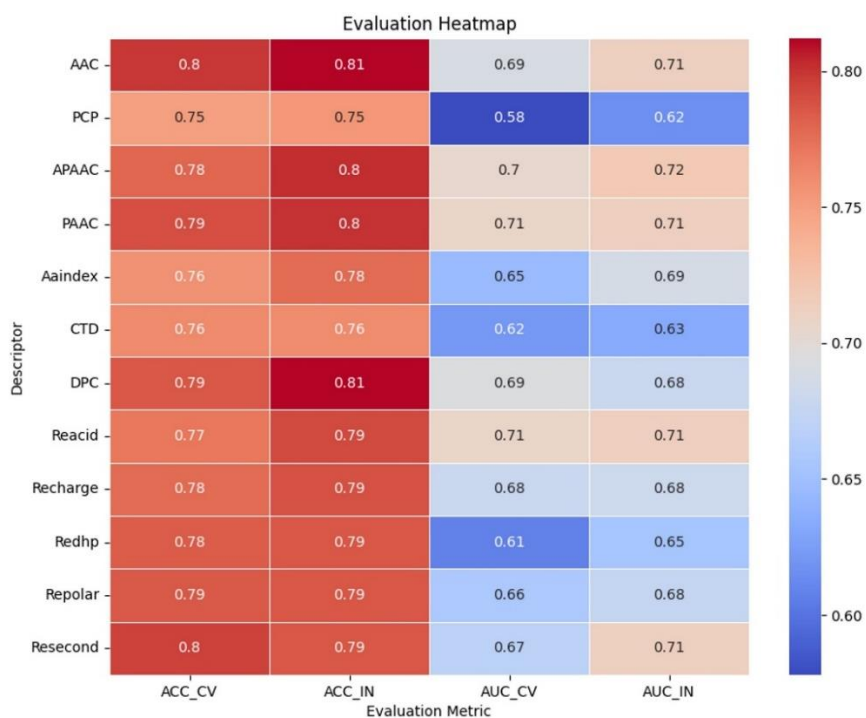


Figure 4 The AUC and the ACC values of the model from the CV and IN datasets when comparing the top 12 descriptors using the alternative dataset

4.3 Appropriateness of Feature Analysis

In this study, t-SNE diagram analysis was used to assess the suitability of feature analysis for the anti-HIV peptide prediction model. Figure 5 shows the distribution of anti-HIV peptide (blue dots) and non-anti-HIV peptide data points (red dots). Figures 5(A), 5(B), and 5(C) correspond to features derived from the main dataset, the alternative dataset, and the 462 selected features from the alternative dataset, respectively. The anti-HIV peptide data points are most closely clustered in the alternative dataset with 462 features. This data sample is part of a new dataset that includes the APAAC, DPC, AAC, and PAAC descriptors, according to the results presented in Section 4.2.

For the second step, a new dataset was generated by using an alternate dataset. This new dataset included features from four descriptor sets,

including APAAC, DPC, AAC, and PAAC. A total of 462 features were used for creating classification models. The new dataset was imported into the system. Various statistical tests and techniques, including the chi-square test, ANOVA, and mutual information, were used in the feature selection process to evaluate and select potentially suitable features for implementation. In this study, top-feature thresholds of 20, 50, 100, 200, and 400 were considered, were considered.

Figure 6 shows a graph illustrating the top 12 ACC values. It is noteworthy that the C400KNN model had the highest ACC value in the IN dataset (0.832), while both of M400KNN and F400KNN models demonstrated impressive ACC values in the CV dataset (0.827 and 0.829, respectively). Additionally, models with 400 features showed high ACC values.

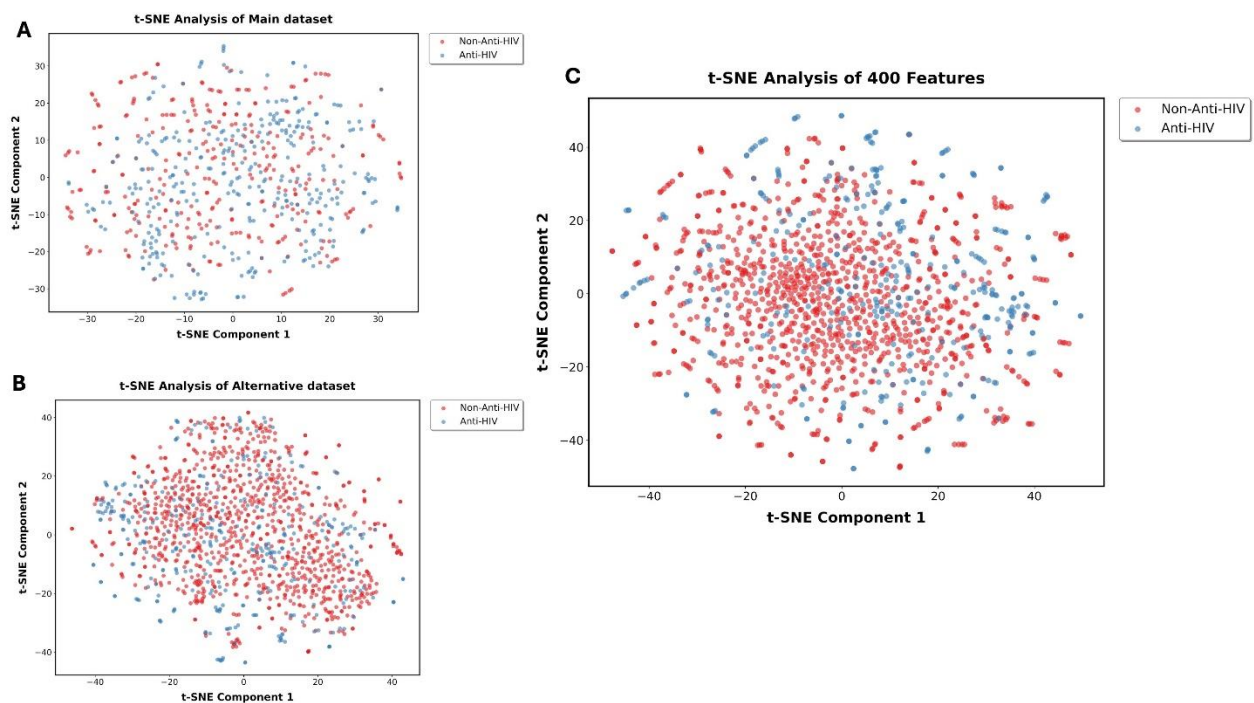


Figure 5 The t-SNE visualization of the feature space used for anti-HIV peptide classification, where red and blue points denote non-anti-HIV and anti-HIV peptides, respectively (random state = 42)

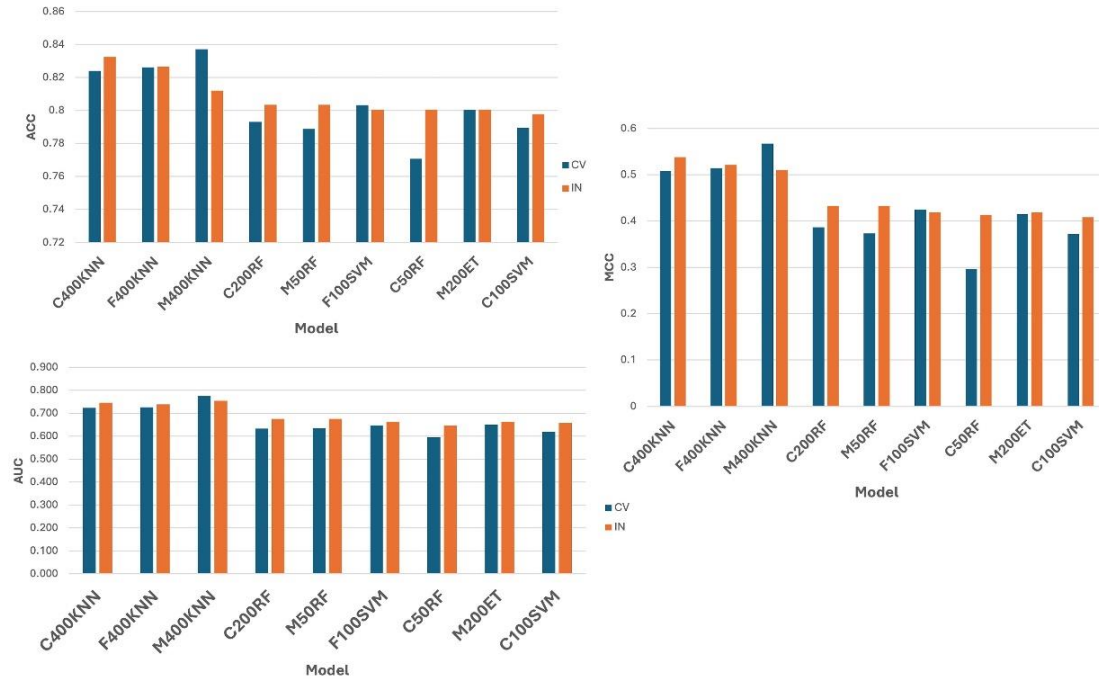


Figure 6 Performance of the proposed model evaluated by ACC, MCC, and AUC across different feature selection methods and numbers of selected features (C: chi-square, F: ANOVA, M: mutual information)

Table 6 Performance evaluation of three candidate models for creating models for anti-HIV peptide classification

Model	ACC		Recall		Precision		MCC		AUC		F1	
	CV	IN	CV	IN	CV	IN	CV	IN	CV	IN	CV	IN
C400KNN	0.824	0.832	0.514	0.567	0.732	0.729	0.508	0.538	0.724	0.746	0.604	0.638
F400KNN	0.826	0.827	0.517	0.556	0.739	0.714	0.514	0.521	0.726	0.739	0.608	0.625
M400KNN	0.828	0.829	0.503	0.578	0.755	0.712	0.516	0.533	0.723	0.748	0.604	0.638

CV = cross-validation, and IN = independent testing

When evaluating the model's performance using the Matthews Correlation Coefficient (MCC), it can be observed that the C400KNN model outperformed the other 119 models with the highest values. In Table 6, the values in the CV and IN datasets are 0.508 and 0.538, respectively. The M400KNN model ranks second with MCC values of 0.516 (CV) and 0.533 (IN). The F400KNN model is in third place, with MCC values of 0.514 and 0.521 in the CV and IN datasets, respectively. The graph shows the top 12 values obtained from model evaluation using AUC. The M400KNN, C400KNN, and F400KNN models have the highest values. The CV set provides consistent values of 0.722, 0.724, and 0.726 for these models, respectively. Within the IN set, the values are 0.748, 0.746, and 0.739, respectively.

4.4 Selection of Suitable Computational Models

The evaluation of model performance using ACC, MCC, and AUC values consistently places C400KNN, M400KNN, and F400KNN models in the top three. It can be observed that all three models use the same classifier, KNN, and have the same number of features, which is 400, with Chi-square (C), mutual information (M), and ANOVA (F) were used as the feature selection methods. Table 6 presents the performance evaluation values for all six categories: ACC, Recall, Precision, MCC, AUC, and F1 score. The evaluation results of the C400KNN model are shown in the table. This model performs best in terms of accuracy (ACC), precision, Matthews correlation coefficient (MCC), and F1 score. The M400KNN model has the highest recall and AUC values. In Figure 7, the confusion matrix for the C400KNN model presents

239 true negatives, 48 true positives, 17 false positives, and 39 false negatives, indicating that non-anti-HIV peptides are identified more accurately than HIV-resistant peptides. Figure 8 presents the SHAP value analysis, which revealed the significance of the features in the C400KNN model. Based on the SHAP analysis, the features AAC19 and PAAC2 demonstrated the highest correlation with the model's

predictive output, indicating a potential association between these specific descriptors and the classification direction. While other descriptors exhibited less influence, the AAC group (specifically AAC19 and AAC9) appeared to play a more prominent role in distinguishing anti-HIV peptides within the analyzed dataset.

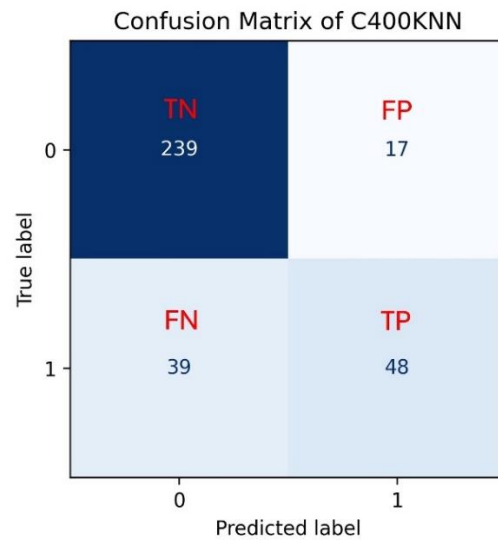


Figure 7 Confusion matrix of the C400KNN model for anti-HIV peptide classification

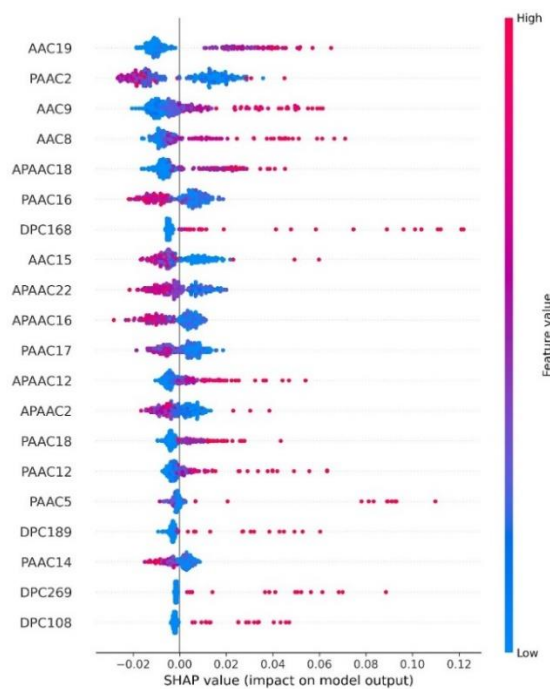


Figure 8 Positive and negative SHAP values in the C400KNN feature importance indicating a high likelihood of anti-HIV and non-anti-HIV prediction outputs

Table 7 Accuracy comparison between the C400KNN model and previously reported machine learning models

Model	Database	Method	ACC
Poorinmohammad & Mohabatkar, 2015	HIPdb, AVPpred	ML	0.840
Poorinmohammad et al., 2015	HIPdb, AVPpred, APD2	ML	0.850
Mathiyazhagan et al., 2022	HIPdb	ML	0.800
C400KNN	HIPdb, AVPpred, APD3, AVPI den, DRAVP	ML	0.832

5. Discussion

This study demonstrated that the selection of negative dataset components significantly influences the performance of anti-HIV peptide prediction models. Training on a negative dataset composed of non-antimicrobial peptides yielded higher ACC and AUC scores across both cross-validation and independent testing phases. These findings suggest that a greater divergence between positive and negative training datasets enhances model learning efficacy (Pang et al., 2021). Previous analyses of amino acid composition revealed that anti-HIV peptides are characterized by high contents of hydrophobic and non-ionic amino acids, particularly isoleucine and leucine (Qureshi et al., 2013). The C400KNN model was initialized as a baseline model utilizing 12 descriptors. Comparative performance assessment indicated that only four descriptors contributed significantly: APAAC, DPC, AAC, and PAAC. Subsequently, feature selection was conducted employing Chi-square, ANOVA, and Mutual Information approaches. A subset comprising 400 features, selected via the Chi-square test, notably improved the performance of the evaluated models. Furthermore, the integration of these features with the K-Nearest Neighbors (KNN) classifier yielded superior overall performance in terms of ACC, MCC, F1-score, and precision (Chicco & Jurman, 2023; Kumar & Dogra, 2022; Roberts et al., 2017; Vamathevan et al., 2019; Zhou et al., 2018). The top features identified by the SHAP analysis are AAC and PAAC, which represent sequence order and physicochemical properties, suggesting a compatibility between the KNN algorithm and these descriptors. This observation aligns with the theoretical foundation of KNN, which relies on proximity-based data similarity. As AAC and PAAC generate structured, continuous datasets, the algorithm can achieve robust performance without requiring intricate modeling techniques (Zhang et al., 2024). These results underscore the effectiveness of descriptor-based feature selection and the strategic design of negative datasets for the robust classification of anti-HIV peptides.

A comparative analysis of the C400KNN model's performance in predicting anti-HIV peptides against previously validated machine learning models revealed accuracy (ACC) scores ranging from 0.80 to 0.85. Notably, the C400KNN model achieved an ACC of 0.83, as detailed in Table 7. However, the results highlight several limitations attributed to the size and nature of the dataset employed during the learning process. Specifically, the study utilized an imbalanced dataset in which positive instances were significantly outnumbered by negative ones. This imbalance may bias the learning process toward non-anti-HIV peptides rather than anti-HIV peptides, thereby compromising predictive performance (Lertampaiporn et al., 2025). While previous studies suggest that data balancing techniques can enhance model performance, such modifications risk the loss of critical information, potentially preventing the model from learning the full data spectrum (Van den Goorbergh et al., 2022).

During the training phase, a 10-fold cross-validation technique was employed to evaluate the model, a standard procedure in machine learning for peptide prediction. A further limitation concerns the disconnect between mathematical and biological feature interpretation, stemming from the use of diverse descriptors and high-dimensional features. Consequently, the model may fail to fully elucidate the underlying biological mechanisms associated with amino acids or their properties. To address this, SHAP analysis was utilized to aid interpretation; however, this method provides only partial insights. Thus, these findings should be regarded as supplementary explanations rather than primary biological interpretations (Samek et al., 2019). Finally, given the significant genetic diversity and rapid mutation rates of HIV, the model's utility may be constrained when applied to novel strains (Gen & Lin, 2023). Therefore, further validation using independent datasets that reflect these viral variations is essential to assess the model's robustness and applicability in real-world scenarios (Hemelaar, 2012; Rambaut et al., 2004).

6. Conclusion

The C400KNN model has been designed for predicting anti-HIV peptides, which are a significant component of the drug design process for future AIDS treatments. This model was developed using machine learning techniques. The data used in this study were extracted from amino acid sequence with a significant gap between positive and negative data points. The data were then transformed into four descriptors: APAAC, DPC, AAC, and PAAC, based on chemical structure and physicochemical properties. A chi-square test was then performed to select the top 400 features for learning with the KNN algorithm. When the model was evaluated, the test results revealed a model accuracy of more than 0.80, which is comparable to models constructed using descriptor principles and trained by machine learning methods. Although the C400KNN model is promising, it has limitations due to data imbalances and a feature selection method that has yet to identify features that reflect biological mechanisms resulting from amino acids or their inherent qualities influencing the model. As a result, future research should prioritize enhancing data collection and selecting data features that may connect mathematical and biological principles, representing biological mechanisms influencing the models. This will improve the model's performance and help generate future medication designs for treating AIDS.

7. Acknowledgements

The authors would like to acknowledge the Faculty of Science and Technology, Prince of Songkla University, Pattani Campus and the Faculty of Medical Technology at Mahidol University for providing financial support and assistance in conducting various sections of this research.

8. CRediT Statement

Jaru Nikom: Data Curation, Formal Analysis, Investigation, Methodology, Software, Validation, Writing – Original Draft and Writing – Review & Editing

Watshara Shoombuatong: Conceptualization and Writing – Review & Editing

Phasit Charoenkwan: Methodology, Software and Writing – Review & Editing

Salang Musikasuwan: Conceptualization, Methodology, Writing – Review & Editing, Resources, Supervision and Project Administration.

9. Abbreviations

Abbreviations	Full Term
CV	Cross-Validation
IN	Independent Test
ACC	Accuracy
MCC	Matthews correlation coefficient
AUC	Area under the ROC Curve
F1	F1 score
SVC	Support Vector Machine
RF	Random Forest
ET	Extra Tree
KNN	K-Nearest Neighbor
LR	Logistic Regression
DT	Decision Tree
NB	Naïve Bayes
AB	Adaboost
XGB	XGBoost
NN	Neural Network
AAC	Amino Acid Composition
DPC	DiPeptide Composition
CTD	Composition/Transition/Distribution
PAAC	Pseudo-Amino Acid Composition
APAAC	Amphiphilic Pseudo-Amino Acid Composition
AAindex	Amino Acid index
PCP	Physicochemical property
reAcid	Reduced Acid
rePolar	Reduced Polar
reCharge	Reduced Charge
reSecond	Reduced Second
reDHP	Reduced DHP
C400KNN	The model uses the KNN classifier and has the number of features, which is 400, with Chi-square.
F400KNN	The model uses the KNN classifier and has the number of features, which is 400, with ANOVA.
M400KNN	The model uses the KNN classifier and has the number of features, which is 400, with mutual information.

10. References

- Charoenkwan, P., Anuwongcharoen, N., Nantasenamat, C., Hasan, M. M., & Shoombuatong, W. (2021). *In silico* approaches for the prediction and analysis of antiviral peptides: A review. *Current Pharmaceutical Design*, 27(18), 2180-2188. <https://doi.org/10.2174/1381612826666201102105827>

- Charoenkwan, P., Nantasenamat, C., Hasan, M. M., & Shoombuatong, W. (2020). Meta-iPVP: A sequence-based meta-predictor for improving the prediction of phage virion proteins using effective feature representation. *Journal of Computer-Aided Molecular Design*, 34(10), 1105-1116. <https://doi.org/10.1007/s10822-020-00323-z>
- Charoenkwan, P., Schaduengrat, N., & Shoombuatong, W. (2023). StackTTCA: A stacking ensemble learning-based framework for accurate and high-throughput identification of tumor T cell antigens. *BMC Bioinformatics*, 24(1), Article 301. <https://doi.org/10.1186/s12859-023-05421-x>
- Chen, Z., Zhao, P., Li, F., Leier, A., Marquez-Lago, T. T., Wang, Y., ... & Song, J. (2018). *iFeature*: A python package and web server for features extraction and selection from protein and peptide sequences. *Bioinformatics*, 34(14), 2499-2502. <https://doi.org/10.1093/bioinformatics/bty140>
- Chicco, D., & Jurman, G. (2023). The Matthews correlation coefficient (MCC) should replace the ROC AUC as the standard metric for assessing binary classification. *BioData Mining*, 16(1), Article 4. <https://doi.org/10.1186/s13040-023-00322-4>
- Chou, K. C., Wu, Z. C., & Xiao, X. (2012). iLoc-Hum: Using the accumulation-label scale to predict subcellular locations of human proteins with both single and multiple sites. *Molecular Biosystems*, 8(2), 629-641. <https://doi.org/10.1039/C1MB05420A>
- Deng, H., Ding, M., Wang, Y., Li, W., Liu, G., & Tang, Y. (2023). ACP-MLC: A two-level prediction engine for identification of anticancer peptides and multi-label classification of their functional types. *Computers in Biology and Medicine*, 158, Article 106844. <https://doi.org/10.1016/j.combiomed.2023.106844>
- Esmacili, M., Mohabatkari, H., & Mohsenzadeh, S. (2010). Using the concept of Chou's pseudo amino acid composition for risk type prediction of human papillomaviruses. *Journal of Theoretical Biology*, 263(2), 203-209. <https://doi.org/10.1016/j.jtbi.2009.11.016>
- Gen, M., & Lin, L. (2023). *Genetic algorithms and their applications*. Springer handbook of engineering statistics. London, UK: Springer London. https://doi.org/10.1007/978-1-4471-7503-2_33
- Greenberg, M. L., & Cammack, N. (2004). Resistance to enfuvirtide, the first HIV fusion inhibitor. *Journal of Antimicrobial Chemotherapy*, 54(2), 333-340. <https://doi.org/10.1093/jac/dkh330>
- Hatta, M., Wahid, W. N., Yusuf, F., Hidayat, F., Santoso, N. A., & Aini, Q. (2024). Enhancing predictive models in system development using machine learning algorithms. *International Journal of Cyber and IT Service Management*, 4(2), 80-87. <https://doi.org/10.34306/ijcitsm.v4i2.159>
- Hemelaar, J. (2012). The origin and diversity of the HIV-1 pandemic. *Trends in Molecular Medicine*, 18(3), 182-192. <https://doi.org/10.1016/j.molmed.2011.12.001>
- Huang, Y., Niu, B., Gao, Y., Fu, L., & Li, W. (2010). CD-HIT Suite: A web server for clustering and comparing biological sequences. *Bioinformatics*, 26(5), 680-682. <https://doi.org/10.1093/bioinformatics/btq003>
- Kumar, V., & Dogra, N. (2022). A comprehensive review on deep synergistic drug prediction techniques for cancer. *Archives of Computational Methods in Engineering*, 29(3), 1443-1461. <https://doi.org/10.1007/s11831-021-09617-3>
- Kwong, P. D., Wyatt, R., Robinson, J., Sweet, R. W., Sodroski, J., & Hendrickson, W. A. (1998). Structure of an HIV gp120 envelope glycoprotein in complex with the CD4 receptor and a neutralizing human antibody. *Nature*, 393(6686), 648-659. <https://doi.org/10.1038/31405>
- Lama, J., & Planelles, V. (2007). Host factors influencing susceptibility to HIV infection and AIDS progression. *Retrovirology*, 4(1), Article 52. <https://doi.org/10.1186/1742-4690-4-52>
- Lertampaiporn, S., Wattanapornprom, W., Thammarongtham, C., & Hongsthong, A. (2025). EnsembleNPPred: A robust approach to neuropeptide prediction and recognition using ensemble machine learning and deep learning methods. *Life*, 15(7), Article 1010. <https://doi.org/10.3390/life15071010>
- Li, W., & Godzik, A. (2006). Cd-hit: A fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics*, 22(13), 1658-1659. <https://doi.org/10.1093/bioinformatics/btl158>

- Lissabet, J. F. B., Belén, L. H., & Farias, J. G. (2019). AntiVPP 1.0: A portable tool for prediction of antiviral peptides. *Computers in Biology and Medicine*, 107, 127-130. <https://doi.org/10.1016/j.combiomed.2019.02.011>
- Liu, Y., Ouyang, X. H., Xiao, Z. X., Zhang, L., & Cao, Y. (2020). A review on the methods of peptide-MHC binding prediction. *Current Bioinformatics*, 15(8), 878-888. <https://doi.org/10.2174/1574893615999200429122801>
- Liu, Y., Zhu, Y., Sun, X., Ma, T., Lao, X., & Zheng, H. (2023). DRAVP: A comprehensive database of antiviral peptides and proteins. *Viruses*, 15(4), Article 820. <https://doi.org/10.3390/v15040820>
- Mathiyazhagan, B., Liyaskar, J., Azar, A. T., Inbarani, H. H., Javed, Y., Kamal, N. A., & Fouad, K. M. (2022). Rough set based classification and feature selection using improved harmony search for peptide analysis and prediction of anti-HIV-1 activities. *Applied Sciences*, 12(4), Article 2020. <https://doi.org/10.3390/app12042020>
- Mori, T., O'Keefe, B. R., Sowder, R. C., Bringans, S., Gardella, R., Berg, S., ... & Boyd, M. R. (2005). Isolation and characterization of griffithsin, a novel HIV-inactivating protein, from the red alga *Griffithsia* sp. *Journal of Biological Chemistry*, 280(10), 9345-9353. <https://doi.org/10.1074/jbc.M411122200>
- Pang, Y., Yao, L., Jhong, J. H., Wang, Z., & Lee, T. Y. (2021). AVPIDen: A new scheme for identification and functional prediction of antiviral peptides based on machine learning approaches. *Briefings in Bioinformatics*, 22(6), Article bbab263. <https://doi.org/10.1093/bib/bbab263>
- Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... & Duchesnay, É. (2011). Scikit-learn: Machine learning in Python. *The Journal of Machine Learning Research*, 12, 2825-2830.
- Poorinmohammad, N., & Mohabatkar, H. (2015). A comparison of different machine learning algorithms for the prediction of anti-HIV-1 peptides based on their sequence-related properties. *International Journal of Peptide Research and Therapeutics*, 21(1), 57-62. <https://doi.org/10.1007/s10989-014-9432-x>
- Poorinmohammad, N., Mohabatkar, H., Behbahani, M., & Biria, D. (2015). Computational prediction of anti HIV-1 peptides and in vitro evaluation of anti HIV-1 activity of HIV-1 P24-derived peptides. *Journal of Peptide Science*, 21(1), 10-16. <https://doi.org/10.1002/psc.2712>
- Prabakaran, P., Dimitrov, A. S., Fouts, T. R., & Dimitrov, D. S. (2007). Structure and function of the HIV envelope glycoprotein as entry mediator, vaccine immunogen, and target for inhibitors. *Advances in Pharmacology*, 55, 33-97. [https://doi.org/10.1016/S1054-3589\(07\)55002-7](https://doi.org/10.1016/S1054-3589(07)55002-7)
- Quaranta, L., Calefato, F., & Lanubile, F. (2021). *Kgtorrent: A dataset of python jupyter notebooks from kaggle* [Conference presentation]. 2021 IEEE/ACM 18th International Conference on Mining Software Repositories (MSR). IEEE, Madrid, Spain. <https://doi.org/10.1109/MSR52588.2021.00072>
- Qureshi, A., Thakur, N., & Kumar, M. (2013). HIPdb: A database of experimentally validated HIV inhibiting peptides. *PloS One*, 8(1), Article e54908. <https://doi.org/10.1371/journal.pone.0054908>
- Rambaut, A., Posada, D., Crandall, K. A., & Holmes, E. C. (2004). The causes and consequences of HIV evolution. *Nature Reviews Genetics*, 5(1), 52-61. <https://doi.org/10.1038/nrg1246>
- Roberts, D. R., Bahn, V., Ciuti, S., Boyce, M. S., Elith, J., Guillerá-Arroita, G., ... & Dormann, C. F. (2017). Cross-validation strategies for data with temporal, spatial, hierarchical, or phylogenetic structure. *Ecography*, 40(8), 913-929. <https://doi.org/10.1111/ecog.02881>
- Samek, W., Montavon, G., Vedaldi, A., Hansen, L. K., & Müller, K. R. (Eds.). (2019). *Explainable AI: Interpreting, explaining and visualizing deep learning*. Switzerland: Springer Nature. <https://doi.org/10.1007/978-3-030-28954-6>
- Schaduangrat, N., Nantasenamat, C., Prachayasittikul, V., & Shoombuatong, W. (2019). Meta-iAVP: A sequence-based meta-predictor for improving the prediction of antiviral peptides using effective feature representation. *International Journal of Molecular Sciences*, 20(22), Article 5743. <https://doi.org/10.3390/ijms20225743>
- Sharma, R., Shrivastava, S., Singh, S. K., Kumar, A., Singh, A. K., & Saxena, S. (2021). Deep-AVPpred: Artificial intelligence driven discovery of peptide drugs for viral infections. *IEEE Journal of Biomedical and Health*

- Informatics*, 26(10), 5067-5074.
<https://doi.org/10.1109/JBHI.2021.3130825>
- Steffler, M., Li, Y., Weir, S., Shaikh, S., Murtada, F., Wright, J. G., & Kantarevic, J. (2021). Trends in prevalence of chronic disease and multimorbidity in Ontario, Canada. *Canadian Medical Association Journal*, 193(8), E270-E277. <https://doi.org/10.1503/cmaj.201473>
- Thakur, N., Qureshi, A., & Kumar, M. (2012). AVPpred: Collection and prediction of highly effective antiviral peptides. *Nucleic Acids Research*, 40(W1), W199-W204.
<https://doi.org/10.1093/nar/gks450>
- Vamathevan, J., Clark, D., Czodrowski, P., Dunham, I., Ferran, E., Lee, G., ... & Zhao, S. (2019). Applications of machine learning in drug discovery and development. *Nature Reviews Drug Discovery*, 18(6), 463-477.
<https://doi.org/10.1038/s41573-019-0024-5>
- Van den Goorbergh, R., Van Smeden, M., Timmerman, D., & Van Calster, B. (2022). The harm of class imbalance corrections for risk prediction models: Illustration and simulation using logistic regression. *Journal of the American Medical Informatics Association*, 29(9), 1525-1534.
<https://doi.org/10.1093/jamia/ocac093>
- Wang, G., Li, X., & Wang, Z. (2016). APD3: The antimicrobial peptide database as a tool for research and education. *Nucleic Acids Research*, 44(D1), D1087-D1093.
<https://doi.org/10.1093/nar/gkv1278>
- Wang, W., Owen, S. M., Rudolph, D. L., Cole, A. M., Hong, T., Waring, A. J., ... & Lehrer, R. I. (2004). Activity of α - and θ -defensins against primary isolates of HIV-1. *The Journal of Immunology*, 173(1), 515-520.
<https://doi.org/10.4049/jimmunol.173.1.515>
- Wei, L., Zhou, C., Su, R., & Zou, Q. (2019). PEPred-Suite: Improved and robust prediction of therapeutic peptides using adaptive feature representation learning. *Bioinformatics*, 35(21), 4272-4280.
<https://doi.org/10.1093/bioinformatics/btz246>
- Xu, J., Li, F., Leier, A., Xiang, D., Shen, H. H., Marquez Lago, T. T., ... & Song, J. (2021). Comprehensive assessment of machine learning-based methods for predicting antimicrobial peptides. *Briefings in Bioinformatics*, 22(5), Article bbab083.
<https://doi.org/10.1093/bib/bbab083>
- Zhang, W., Ding, Y., Wei, L., Guo, X., & Ni, F. (2024). Therapeutic peptides identification via kernel risk sensitive loss-based k-nearest neighbor model and multi-Laplacian regularization. *Briefings in Bioinformatics*, 25(6), Article bbae534.
<https://doi.org/10.1093/bib/bbae534>
- Zhou, Y., Fujikura, K., Mkrtchian, S., & Lauschke, V. M. (2018). Computational methods for the pharmacogenetic interpretation of next generation sequencing data. *Frontiers in Pharmacology*, 9, Article 1437.
<https://doi.org/10.3389/fphar.2018.01437>