TRENDS IN SCIENCES 2026; 23(3): 11717
https://doi.org/10.48048/tis.2026.11717 RESEARCH ARTICLE

GAXGB: A Two-Stage Ensemble Framework Integrating Genetic Algorithms and XGBoost for Anti-HIV Peptide Prediction

Jaru Nikom[1], Watshara Shoombuatong[2], Phasit Charoenkwan[3] and Salang Musikasuwan[1,*]

[1]*Department of Mathematics and Computer Science, Faculty of Science and Technology, Prince of Songkla University, Pattani Campus, Pattani 94000, Thailand*
[2]*Center for Research Innovation and Biomedical Informatics, Faculty of Medical Technology, Mahidol University, Nakhon Pathom 10700, Thailand*
[3]*Modern Management and Information Technology, College of Arts, Media and Technology, Chiang Mai University, Chiang Mai 50200, Thailand*

([*]Corresponding author's e-mail: salang.m@psu.ac.th)

Received: 28 August 2025, Revised: 17 September 2025, Accepted: 10 October 2025, Published: 15 December 2025

Abstract

This study applied a computational approach to derive amino acid sequence features relevant to AIDS treatment. A predictive model, GAXGB, was developed to classify anti-HIV peptides based on their amino acid sequence characteristics. The model was built using a 2-stage learning procedure. In the first stage, features were extracted from amino acid sequences using 12 descriptors, and 120 baseline models were constructed with 10 different classifiers. In the second stage, these baseline models generated 120 predictive probability scores, which were used as input features. Various feature selection methods, including chi-square, ANOVA, mutual information, and a genetic algorithm, were employed to identify the most significant features for the final model. Subsequently, 10 classifiers were trained and evaluated. Performance evaluation showed that the GAXGB model, which combines genetic algorithm-based feature selection with an XGBoost classifier, achieved superior predictive accuracy. The model reached an accuracy of 90%, significantly outperforming other models that achieved approximately 80% accuracy. This approach offers a promising tool to accelerate the design and discovery of novel anti-HIV peptides for AIDS treatment.

Keywords: Classification model, Anti-HIV peptides, Genetic algorithms, Ensemble learning methods

Introduction

AIDS has been a worldwide epidemic for nearly 50 years. HIV infection produces this disease, which affects the immune system by killing white blood cells. Despite ongoing drug development and the introduction of drug resistance genes in this virus, drug research continues [1]. Furthermore, researchers formed these drug groupings to treat the coronavirus, which first emerged a few years ago [2]. The advancement of this medicine class will considerably help global society. A common approach to investigating the development of peptides as drugs involves selecting random amino acid sequences from different organisms and analyzing them in a laboratory for biological activity [3]. This approach wastes a significant amount of time and research costs [4,5]. Over the last decade, researchers have used computational models to predict or classify peptides with various biological properties, including antimicrobial, antiviral, and anticancer peptides [2,6-12].

Methods based on machine learning are widely used in the development of models to predict or classify peptides having biological activity. Effective data collection is an essential step in creating this type of model. Nowadays, most anti-HIV peptide databases are part of antimicrobial and antiviral peptide databases, such as AVPpred [13], dbAMP 2.0 [14], AVPdb [15], SATPdb [16], DRAMP [17] and DBAASP [18] Researchers consider the HIPdb [19] database to be the only 1 dedicated to anti-HIV peptides.

Poorinmohammad and Mohabatkar [20]; Poorinmohammad *et al.* [21] used the feature stages of pseudo-amino acid composition [22] and Chou's pseudo-amino acid composition [23] to develop a prediction model for the anti-HIV-1 peptide using multilayer perceptron [24] and support vector machine (SVM) [25]. In a modeling study, [10] predicted an anti-HIV peptide utilizing chemical composition, physical-chemistry properties, and the Random Forest algorithm [10]. In 2022, the research are going to apply rough set-based categorization and feature selection to create a model that assesses and predicts peptide anti-HIV activity [26]. When comparing models used to predict peptides with bioactivities, it is apparent that there are only a small number of models that predict or classify anti-HIV peptides. Most of these models have estimations that do not go over 90% accuracy. The remaining models are rather modest in comparison to other models.

Based on the highlighted points, GAXGB is introduced, a machine learning model that learns from the anti-HIV peptide positive dataset. The antimicrobial prediction model's negative data sets were used. Reports indicate that this will improve the model, resulting in a higher model evaluation value [10]. The feature extraction was calculated from all the data in the 2 steps. The first step is to create a feature baseline consisting of 12 descriptors representing chemical composition and physicochemical properties. The second step is to combine the baseline feature values with the predictive probability of a model for identifying anti-HIV peptides using 10 classifiers, creating a new set of 120 features. The chi-square test, ANOVA, multiple information, and genetic algorithm were used for selecting appropriate features for modeling. Use the features to build a model with the 10-fold cross-validation training method. Next, test the model with the evaluation values until it achieves an accuracy of around 90% in the group of independent data sets it can handle. This is the greatest point of the constructed model.

Material and method

Overview framework: The GAXGB model

Figure 1 represents the progress of the GAXGB model development study. It was widely divided into 4 steps: Data collection, baseline creation, feature selection, and model development, as described below: The first step was to group together the data used in this study into amino acid sequences totaling 1,719 peptides collected from an online database. The second step is the feature calculation process, extracting features from 12 descriptors and then valuing the predictive probabilities score with 10 classifiers to create 120 NF of new features. The third step is determining the appropriate features to build a model for classifying anti-HIV peptides using 4 methods: Chi square, ANOVA, mutual information, and genetic algorithm. And the last step involves developing the model by constructing and evaluating its performance to achieve the aim of the GAXGB model.

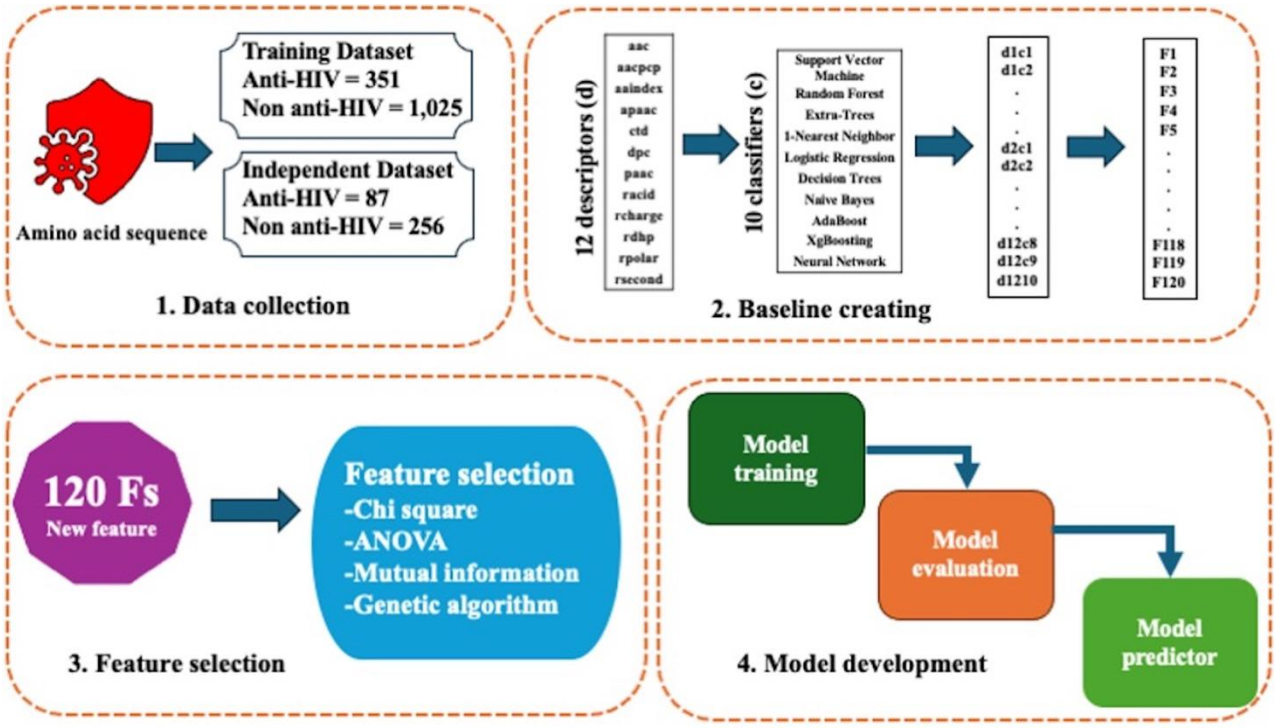

Figure 1 Workflow for GAXGB model development. The method begins with the collecting of peptide sequence data, which is followed by the establishment of a baseline performance using standard descriptive terms. The most informative variables are then identified using a genetic algorithm-based feature selection method. Finally, the optimized feature set is used to create a model with XGBoost, yielding the complete prediction framework.

Data collection and cleaning

This classification model uses amino acid sequence data from a collection of 1,719 peptides from 5 online database [10,13,19,27,28]. The peptides are further divided into positive and negative dataset. Each dataset is then split into 80% for training and 20% for independent testing. The mentioned data set was examined for any duplication across different groups. And with in the positive and negative groups, there were examined for biomolecular redundancy using the CD-hit program at a threshold level of 80% [29]. Ultimately, this study will construct the model by utilizing a training set consisting of 351 positive peptides, 1,025 negative peptides, and an independent set containing 87 negative peptides and 256 negative peptides.

Feature extracting and computing

The goal of this stage is to translate amino acid sequence data into numerical data that machine learning models can compute using chemical structure and physicochemical properties. This study employed 12 descriptors: aac, pcp, aaindex, apaac, ctd, dpc, paac, racid, rcharge, rdhp, rpolar, and rsecond, for a total of 1,338 sub-features, as shown in **Table 1**. This was based on the StackTTCA [30] model for calculating training and independent test data. This study is using the iFeature package [31], identifying the whole set of data as F. This study has used the data to calculate in step 2 to create new data. Furthermore, in step 2, the data was used to create new data. This data is used by each descriptor to calculate the predictive probabilities of the classifier's anti-HIV peptide classification model. The technique uses 10 different algorithms, including Support Vector Machine (svc), Random Forest (rf), Extra-Trees (et), K-Nearest Neighbor (knn), Logistic Regression (lr), Decision Trees (dt), Naive Bayes (nb), AdaBoost (ada), XGBoosting (xgb), and Neural Network (nn), to generate a total of 120 new features (NF).

Table 1 Details about the 12 descriptors, as well as the number of features used in this study.

Descriptor	Description	Number of features
AAC	Frequency of 20 amino acids	20
AAindex	Different biochemical and biophysical properties extracted from the AAindex database	531
APAAC	Amphiphilic pseudo-amino acid composition	22
CTD	Composition, transition and distribution	147
DPC	Frequency of 400 dipeptides	400
PAAC	Pseudo amino acid composition	11
PCP	Different biochemical and biophysical properties extracted from the AAindex database	21
rAcid	Reduced amino acid sequences according to acidity	32
rCharge	Reduced amino acid sequences according to charge	50
rDHP	Reduced amino acid sequences according to DHP	32
rPolar	Reduced amino acid sequences according to polarity	32
rSecond	Reduced amino acid sequences according to secondary structure	40

Feature selection

Next, the data from the feature value calculation were feed into the feature selection technique. In this study, 4 methods were used: The chi square method, ANOVA, mutual information, and the genetic algorithm. These methods will use the F data to determine the best features for developing a model that accurately classifies anti-HIV peptides. The scikit learn package [32] uses the chi square, ANOVA, and mutual information approaches, with feature counts set to 12, 30, 60, and 90, respectively. The DEAP library [33] uses the genetic algorithm method in this parameter (population = 50, generation = 20, crossover probability = 0.5 and mutation probability = 0.2). In each process investigated, a model must be developed to simulate the learning of all 10 classifier algorithms (svm, rf, et, knn, lr, dt, nb, ada, xgb, and nn) using the 10-fold cross-validation use with training data. An independent testing dataset was used to evaluate the performance of the generated models.

Model performance evaluation

The model's performance value indicates its abilities in several areas. The developed model in this study tested the following performance values: accuracy (ACC), recall, precision, and Matthews' correlation coefficient (MCC). The formula can calculate the measurements. As follows:

$$\mathrm{ACC} = \frac{\mathrm{TP} + \mathrm{TN}}{(\mathrm{TP} + \mathrm{TN} + \mathrm{FP} + \mathrm{FN})} \quad (1)$$

$$\mathrm{recall} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}} \quad (2)$$

$$\mathrm{precision} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}} \quad (3)$$

$$\mathrm{MCC} = \frac{\mathrm{TP} \times \mathrm{TN} - \mathrm{FP} \times \mathrm{FN}}{\sqrt{(\mathrm{TP} + \mathrm{FP})(\mathrm{TP} + \mathrm{FN})(\mathrm{TN} + \mathrm{FP})(\mathrm{TN} + \mathrm{FN})}} \quad (4)$$

$$\mathrm{F1\ score} = \frac{\mathrm{TP}}{\mathrm{TP} + \frac{1}{2}(\mathrm{FP} + \mathrm{FN})} \quad (5)$$

The variables TN and TP represent the number of negative and positive samples shown to be negative or positive, respectively. Meanwhile, FN and FP represent the number of positive and negative samples predicted to be negative or positive, respectively. In addition, the area under the receiver operating characteristic (ROC) curve (AUC) was used to evaluate the model's robustness.

Results and discussion

Results

Data collection and analysis

The GAXGB model is constructed using a dataset consisting of 438 peptides that have anti-HIV properties and another dataset containing 1,281 peptides that do not have anti-HIV properties. In order to eliminate redundancy, manual data cleaning was performed and used the CD-hit tool with a threshold level set at 0.8. Upon analyzing the density features of the amino acids in the anti-HIV and non-HIV datasets, it was found that both datasets had a specific quantity of valine (V), which is a type of amino acid. **Figure 2** displays the presence of isoleucine (I), glutamic acid (E), lysine (K), glutamine (Q), tryptophan (W), glycine (G), and cysteine (C) in the denser anti-HIV data set compared to the non anti-HIV data set. During the analysis phase, the physicochemical characteristics were investigated. Upon comparing the amino acid sequences of the anti-HIV and non-HIV datasets, the anti-HIV dataset exhibited more prominent characteristics of being tiny and neutral compared to the non-HIV dataset, as depicted in **Figure 3**.

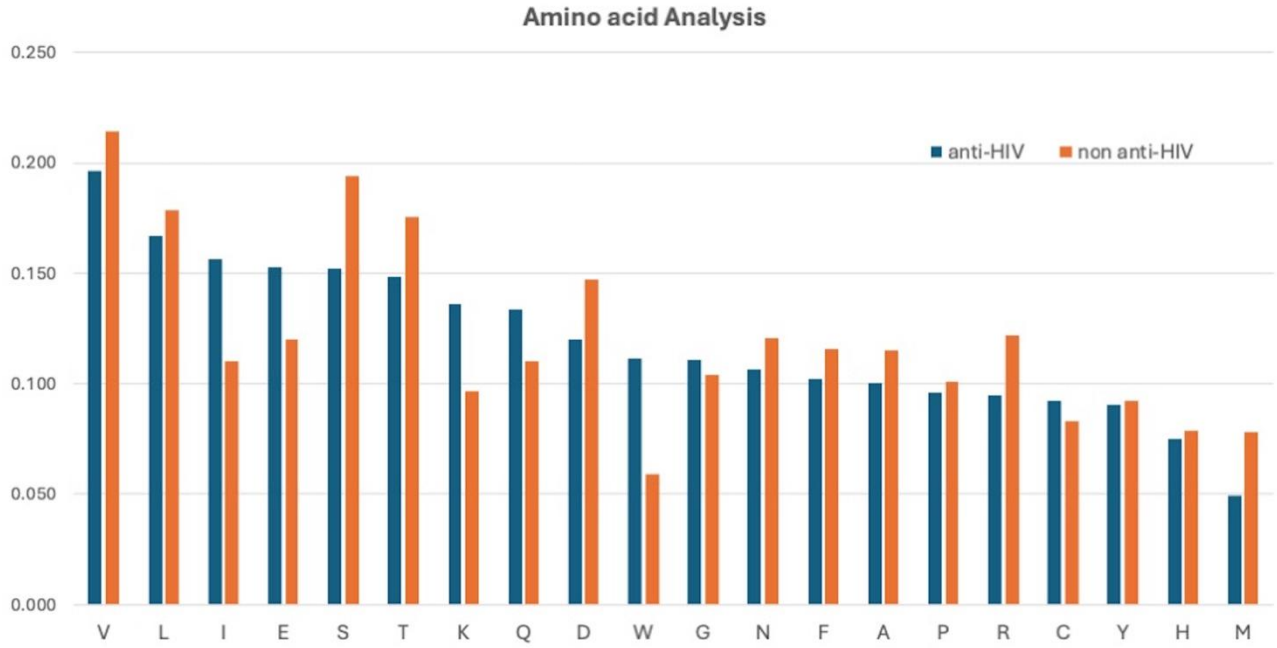

Figure 2 Density distribution of amino acid frequencies in anti-HIV peptides versus non-anti-HIV peptides. Each curve depicts the normalized occurrence of amino acids in the 2 data sets. Peaks show a higher frequency of certain residues in a dataset. For reference, amino acid abbreviations are as follows: V = Valine, L = Leucine, I = Isoleucine, E = Glutamic Acid, S = Serine, T = Threonine, K = Lysine, Q = Glutamine, D = Aspartic Acid, W = Trytophan, G = Glycine, N = Asparagine, F = Phenylalanine, A = Alanine, P = Proline, R = Arginine, C = Cysteine, Y = Tyrosine, H = Histidine and M = Methionine.

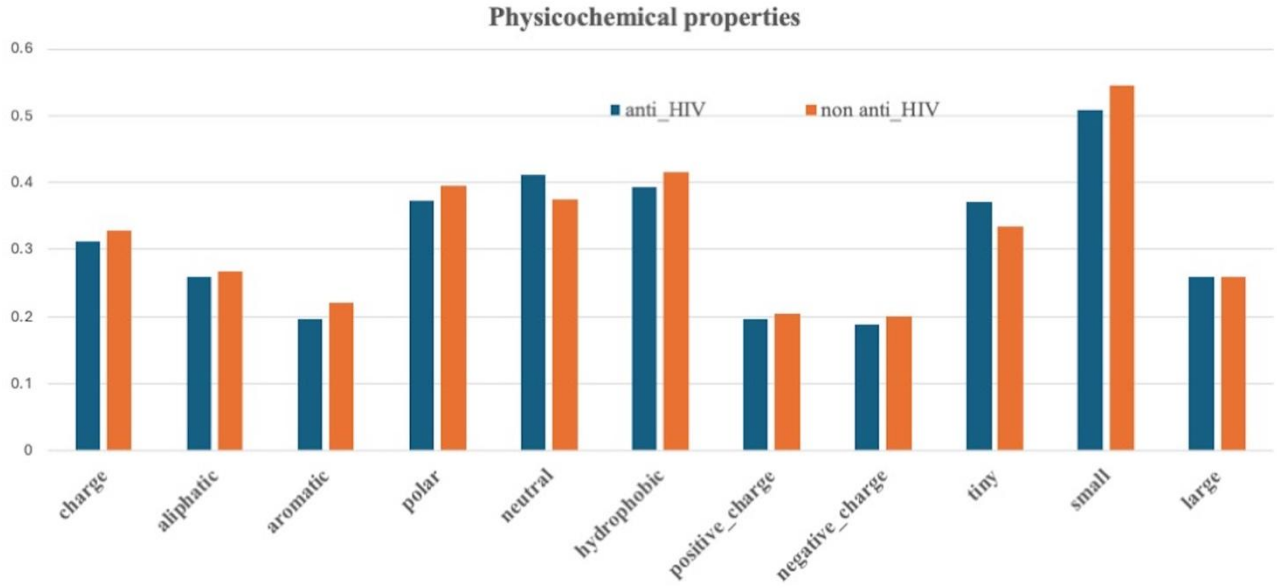

Figure 3 Comparative distribution of physicochemical features based on amino acid content of anti-HIV peptides and non-anti-HIV peptides. The graph depicts discrepancies in amino acid counts that could lead to functional differentiation between the 2 datasets.

Development and evaluation of a baseline model

Following this step, the data to extract features using 12 descriptors were utilized, including aac, pcp, aaindex, apaac, ctd, dpc, paac, racid, rcharge, rdhp, rpolar, and rsecond. Next, a model is constructed using 10 different algorithms, including svm, rf, et, knn, lr, dt, nb, ada, xgb, and nn, resulting in an overall total of 120 baseline models. Afterwards, the model's performance was evaluated using ACC, recall, precision, MCC, and AUC values in both the cross-validation set and the independent testing set. These values, together with the corresponding parameters, are presented in **Tables 2** and **3**. When considering the 120 baseline values shown in **Figures 4** and **5**, which show the ACC values of CV and IN sets that are in the top 12, it is found that the dpc_et model in CV set has the highest value of 0.824, followed by aac_xgb, dpc_rf and aac_et, which have values of 0.818, 0.812, and 0.810, respectively, are shown in **Figure 4**. **Figure 5** displays the evaluation of ACC performance for the top 12 baseline models using an IN set. It indicates that the aac_xgb model and dpc_rf model have the highest ACC value, which is 0.816. Following them are the dpc_et model and rcharge_xgb model, with ACC values of 0.813 and 0.802, respectively. The Matthews' correlation coefficient is a statistic used to assess the performance of classification models. This is especially accurate when the data class is imbalanced [34]. **Figures 6** and **7** represent the numerical values of the initial 12 nodes in the baseline model. **Figure 6** presents an informative summary of the CV set. The dpc_et model has the highest value of 0.490, followed by the aac_xgb, dpc_rf, and racid_xgb models, with values of 0.474, 0.443, and 0.439, respectively. **Figure 7** shows the Matthews Correlation Coefficient (MCC) values for the 12 highest-ranking models in the IN set. The aac_xgb model has the greatest value, with a value of 0.486. It is followed by dpc_rf, dpc_et, and dpc_knn, which have values of 0.459, 0.448, and 0.432, respectively.

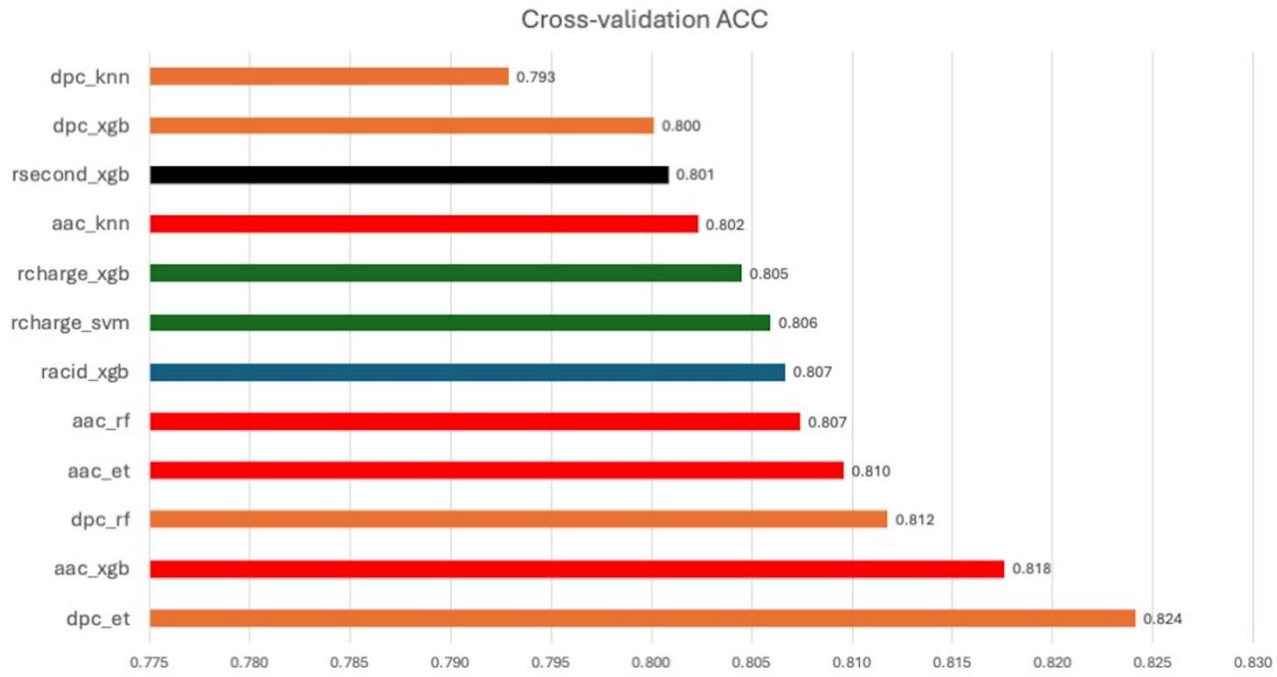

Figure 4 Support Vector Machine (SVM), Random Forest (RF), XGBoost (XGB), Logistic Regression (LR), and other baseline machine learning models are ranked in descending order of performance, with bars representing mean ACC values across CV folds.

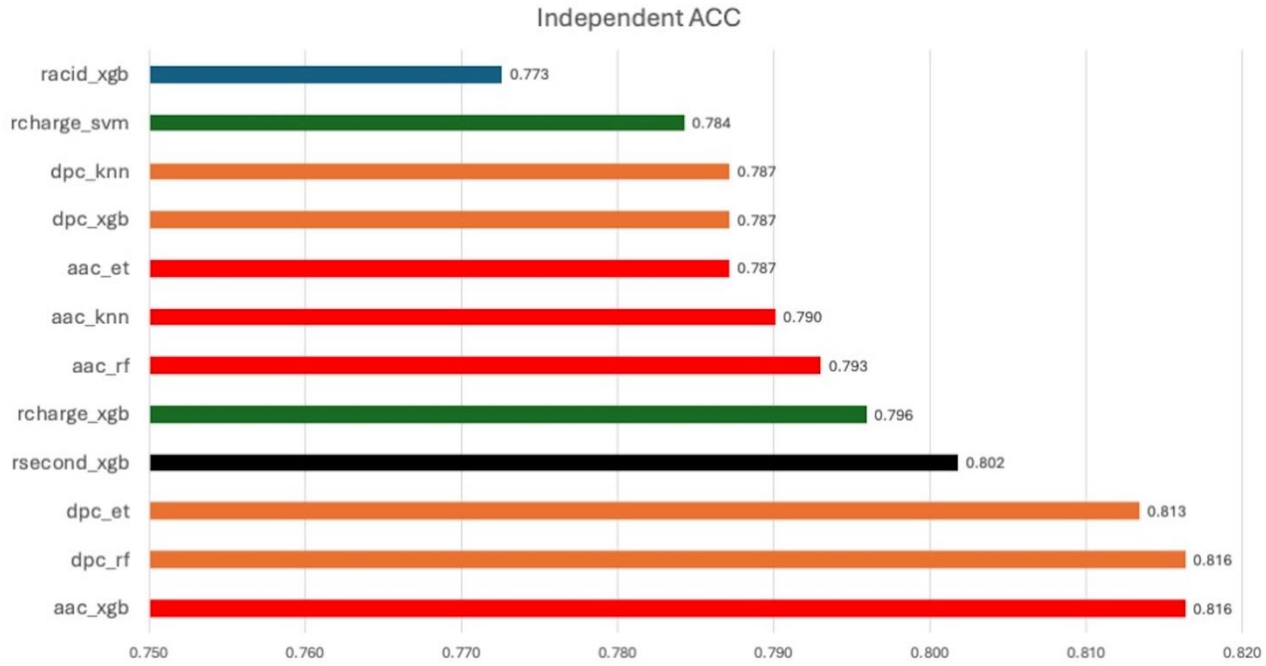

Figure 5 Accuracy (ACC) performance estimates for the 12 baseline classification models tested on an independent test set (IN). Each bar indicates the average accuracy across cross-validation folds. Models are sorted from highest to lowest accuracy to highlight performance disparities.

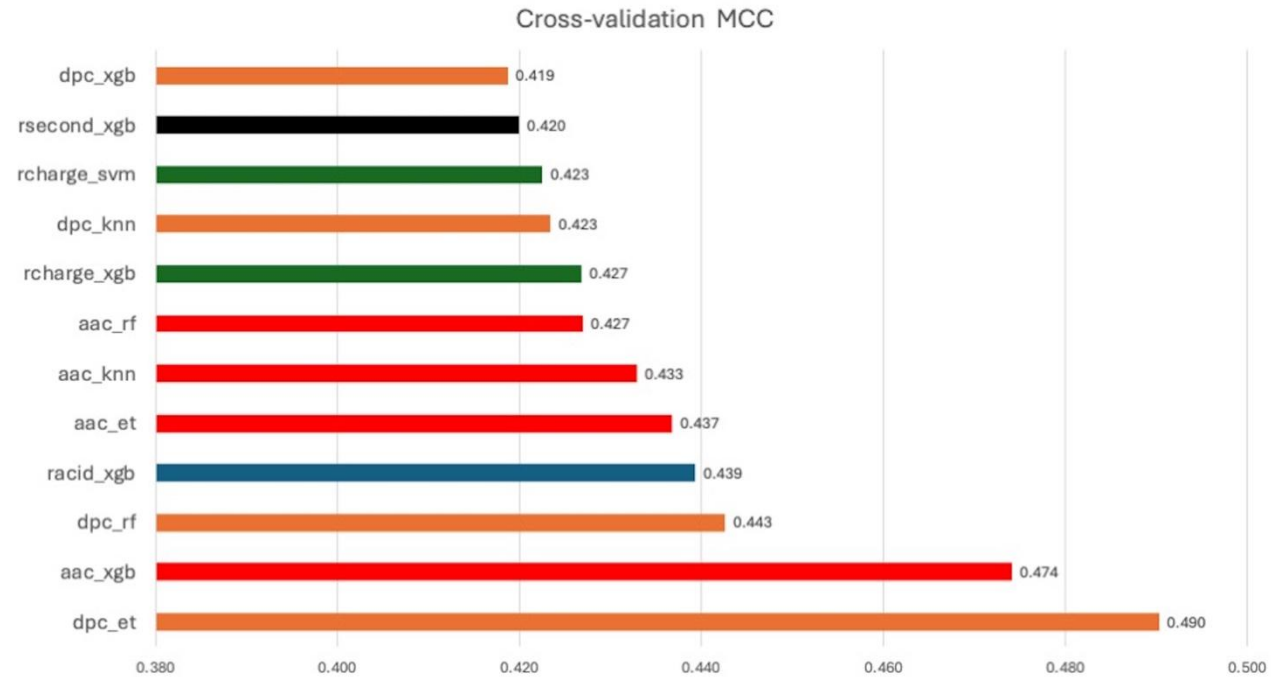

Figure 6 Matthews Correlation Coefficient (MCC) performance estimates for the top 12 baseline machine learning models were calculated using 10-fold cross-validation on the training dataset. Each bar shows the average MCC across folds. Higher MCC values indicate improved overall classification performance, which balances sensitivity and specificity.

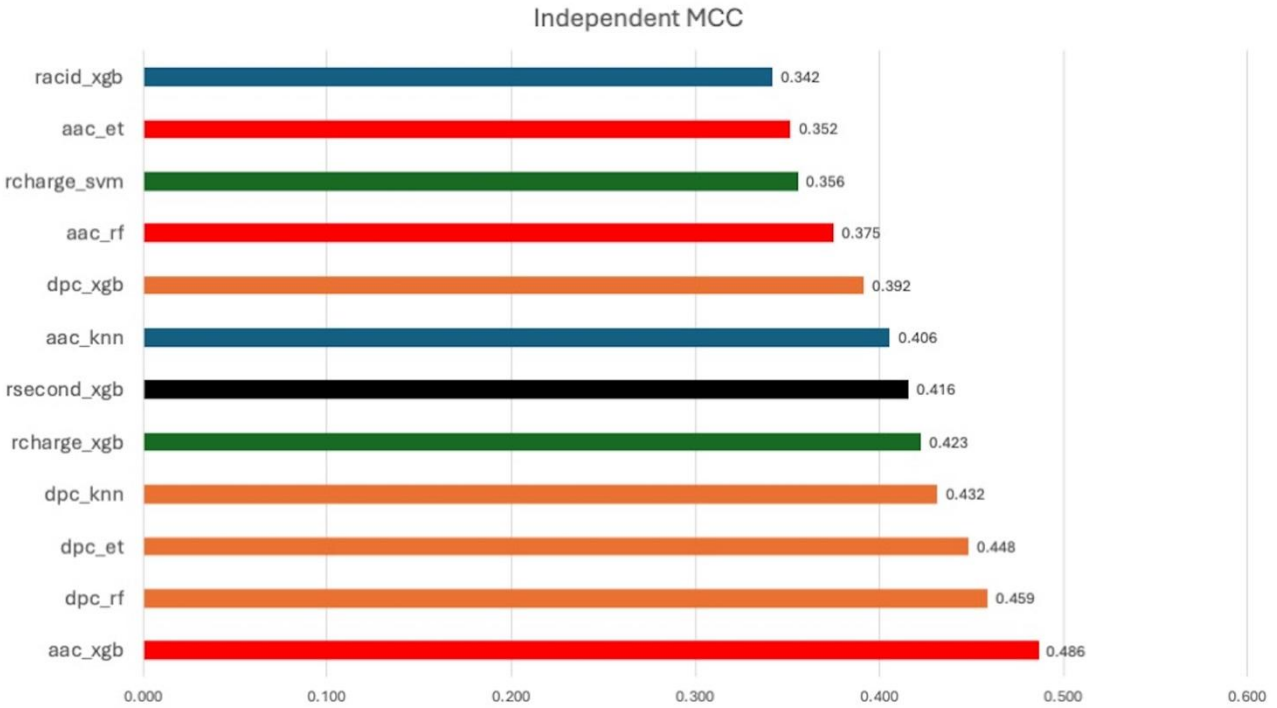

Figure 7 Matthews Correlation Coefficient (MCC) performance estimates of the top 12 baseline machine learning models evaluated on the independent (IN) dataset. Higher MCC values indicate better predictive performance.

Table 2 Cross-validation results of 120 baseline models as developed with 10 ML algorithms and 12 feature encoding schemes.

Descriptor	Classifier	ACC	Recall	Precision	MCC	AUC	Parameter
ACC	SVM	0.803	0.296	0.813	0.410	0.636	{'C': 1, 'gamma': 0.1, 'kernel': 'rbf'}
	RF	0.807	0.302	0.841	0.427	0.641	{'max_depth': None, 'n_estimators': 200}
	ET	0.810	0.296	0.874	0.437	0.641	{'max_depth': None, 'n_estimators': 200}
	KNN	0.802	0.459	0.663	0.433	0.689	{'n_neighbors': 3}
	LR	0.759	0.168	0.596	0.218	0.565	{'C': 0.1}
	DT	0.775	0.282	0.631	0.309	0.613	{'max_depth': 5, 'min_samples_split': 5}
	NB	0.730	0.356	0.461	0.234	0.607	{}
	ADA	0.775	0.419	0.583	0.357	0.658	{'learning_rate': 1, 'n_estimators': 200}
	XGB	0.818	0.464	0.721	0.474	0.701	{'n_estimators': 200}
	NN	0.786	0.376	0.638	0.369	0.651	{'alpha': 0.0001, 'hidden_layer_sizes': (100,), 'max_iter': 100}
AAindex	SVM	0.753	0.034	0.923	0.150	0.517	{'C': 10, 'gamma': 0.1, 'kernel': 'rbf'}
	RF	0.786	0.274	0.706	0.342	0.617	{'max_depth': None, 'n_estimators': 200}
	ET	0.783	0.234	0.732	0.326	0.602	{'max_depth': None, 'n_estimators': 200}
	KNN	0.767	0.293	0.585	0.290	0.611	{'n_neighbors': 5}
	LR	0.758	0.174	0.587	0.217	0.566	{'C': 0.1}
	DT	0.742	0.316	0.491	0.240	0.602	{'max_depth': 5, 'min_samples_split': 10}
	NB	0.647	0.524	0.366	0.193	0.606	{}
	ADA	0.765	0.185	0.637	0.248	0.575	{'learning_rate': 0.1, 'n_estimators': 200}
	XGB	0.777	0.348	0.610	0.336	0.636	{'n_estimators': 200}
	NN	0.768	0.373	0.570	0.323	0.638	{'alpha': 0.001, 'hidden_layer_sizes': (100,), 'max_iter': 50}
APAAC	SVM	0.793	0.268	0.770	0.369	0.620	{'C': 1, 'gamma': 0.1, 'kernel': 'rbf'}
	RF	0.790	0.234	0.804	0.356	0.607	{'max_depth': None, 'n_estimators': 100}
	ET	0.796	0.248	0.837	0.381	0.616	{'max_depth': None, 'n_estimators': 100}
	KNN	0.794	0.467	0.631	0.416	0.687	{'n_neighbors': 5}
	LR	0.749	0.177	0.525	0.190	0.561	{'C': 0.1}

Descriptor	Classifier	ACC	Recall	Precision	MCC	AUC	Parameter
	DT	0.736	0.205	0.462	0.169	0.562	{'max_depth': 5, 'min_samples_split': 10}
	NB	0.736	0.276	0.471	0.208	0.585	{}
	ADA	0.759	0.339	0.546	0.289	0.621	{'learning_rate': 1, 'n_estimators': 100}
	XGB	0.783	0.370	0.625	0.358	0.647	{'n_estimators': 100}
	NN	0.788	0.387	0.638	0.376	0.656	{'alpha': 0.001, 'hidden_layer_sizes': (100,), 'max_iter': 100}
CTD	SVM	0.745	0.000	0.000	0.000	0.500	{'C': 0.1, 'gamma': 0.1, 'kernel': 'rbf'}
	RF	0.773	0.145	0.797	0.275	0.566	{'max_depth': None, 'n_estimators': 200}
	ET	0.773	0.140	0.831	0.279	0.565	{'max_depth': None, 'n_estimators': 200}
	KNN	0.760	0.293	0.557	0.273	0.607	{'n_neighbors': 5}
	LR	0.744	0.299	0.498	0.237	0.598	{'C': 0.1}
	DT	0.701	0.268	0.378	0.132	0.558	{'max_depth': 5, 'min_samples_split': 10}
	NB	0.624	0.564	0.352	0.184	0.604	{}
	ADA	0.761	0.140	0.645	0.216	0.557	{'learning_rate': 0.1, 'n_estimators': 200}
	XGB	0.773	0.274	0.627	0.302	0.609	{'n_estimators': 100}
	NN	0.756	0.413	0.527	0.312	0.643	{'alpha': 0.01, 'hidden_layer_sizes': (100,), 'max_iter': 50}
DPC	SVM	0.745	0.000	0.000	0.000	0.500	{'C': 0.1, 'gamma': 0.1, 'kernel': 'rbf'}
	RF	0.812	0.348	0.803	0.443	0.659	{'max_depth': None, 'n_estimators': 100}
	ET	0.824	0.350	0.898	0.490	0.668	{'max_depth': None, 'n_estimators': 200}
	KNN	0.793	0.501	0.615	0.423	0.697	{'n_neighbors': 5}
	LR	0.723	0.407	0.453	0.247	0.619	{'C': 0.1}
	DT	0.765	0.276	0.581	0.278	0.604	{'max_depth': 10, 'min_samples_split': 10}
	NB	0.495	0.798	0.310	0.173	0.594	{}
	ADA	0.766	0.162	0.671	0.245	0.568	{'learning_rate': 0.1, 'n_estimators': 200}
	XGB	0.800	0.425	0.671	0.419	0.677	{'n_estimators': 100}
	NN	0.772	0.479	0.562	0.371	0.675	{'alpha': 0.0001, 'hidden_layer_sizes': (100,), 'max_iter': 100}

Descriptor	Classifier	ACC	Recall	Precision	MCC	AUC	Parameter
PAAC	SVM	0.793	0.276	0.758	0.369	0.623	{‘C’: 1, ‘gamma’: 0.1, ‘kernel’: ‘rbf’}
	RF	0.776	0.205	0.713	0.296	0.588	{‘max_depth’: 10, ‘n_estimators’: 50}
	ET	0.798	0.256	0.841	0.390	0.620	{‘max_depth’: None, ‘n_estimators’: 200}
	KNN	0.787	0.541	0.590	0.425	0.706	{‘n_neighbors’: 3}
	LR	0.756	0.197	0.561	0.220	0.572	{‘C’: 1}
	DT	0.741	0.222	0.481	0.190	0.570	{‘max_depth’: 5, ‘min_samples_split’: 10}
	NB	0.738	0.293	0.479	0.221	0.592	{}
	ADA	0.760	0.165	0.611	0.222	0.565	{‘learning_rate’: 0.1, ‘n_estimators’: 200}
	XGB	0.779	0.385	0.605	0.353	0.649	{‘n_estimators’: 200}
	NN	0.788	0.385	0.640	0.376	0.655	{‘alpha’: 0.01, ‘hidden_layer_sizes’: (100,), ‘max_iter’: 100}
PCP	SVM	0.758	0.091	0.696	0.188	0.539	{‘C’: 1, ‘gamma’: 1, ‘kernel’: ‘rbf’}
	RF	0.763	0.171	0.632	0.235	0.568	{‘max_depth’: 10, ‘n_estimators’: 50}
	ET	0.764	0.179	0.630	0.241	0.572	{‘max_depth’: None, ‘n_estimators’: 100}
	KNN	0.731	0.231	0.448	0.172	0.567	{‘n_neighbors’: 5}
	LR	0.743	0.000	0.000	-0.022	0.499	{‘C’: 0.1}
	DT	0.734	0.165	0.443	0.140	0.547	{‘max_depth’: 5, ‘min_samples_split’: 2}
	NB	0.715	0.154	0.362	0.086	0.531	{}
	ADA	0.746	0.066	0.511	0.108	0.522	{‘learning_rate’: 0.1, ‘n_estimators’: 200}
	XGB	0.751	0.342	0.517	0.271	0.616	{‘n_estimators’: 200}
	NN	0.741	0.026	0.391	0.041	0.506	{‘alpha’: 0.001, ‘hidden_layer_sizes’: (100,), ‘max_iter’: 50}
rAcid	SVM	0.798	0.325	0.735	0.393	0.642	{‘C’: 10, ‘gamma’: 0.1, ‘kernel’: ‘rbf’}
	RF	0.794	0.251	0.807	0.372	0.615	{‘max_depth’: None, ‘n_estimators’: 100}
	ET	0.797	0.262	0.821	0.387	0.621	{‘max_depth’: None, ‘n_estimators’: 200}
	KNN	0.789	0.456	0.615	0.399	0.679	{‘n_neighbors’: 3}
	LR	0.762	0.242	0.582	0.259	0.591	{‘C’: 10}
	DT	0.750	0.242	0.521	0.224	0.583	{‘max_depth’: 5, ‘min_samples_split’: 2}

Descriptor	Classifier	ACC	Recall	Precision	MCC	AUC	Parameter
	NB	0.722	0.407	0.451	0.246	0.619	{}
	ADA	0.766	0.179	0.649	0.249	0.573	{‘learning_rate’: 0.1, ‘n_estimators’: 200}
	XGB	0.807	0.439	0.691	0.439	0.686	{‘n_estimators’: 200}
	NN	0.777	0.373	0.601	0.344	0.644	{‘alpha’: 0.001, ‘hidden_layer_sizes’: (100,), ‘max_iter’: 100}
rCharge	SVM	0.806	0.350	0.759	0.423	0.656	{‘C’: 10, ‘gamma’: 0.1, ‘kernel’: ‘rbf’}
	RF	0.791	0.248	0.784	0.359	0.612	{‘max_depth’: None, ‘n_estimators’: 200}
	ET	0.799	0.276	0.808	0.392	0.627	{‘max_depth’: None, ‘n_estimators’: 50}
	KNN	0.800	0.359	0.716	0.405	0.655	{‘n_neighbors’: 5}
	LR	0.754	0.222	0.542	0.225	0.579	{‘C’: 1}
	DT	0.765	0.248	0.592	0.267	0.595	{‘max_depth’: 5, ‘min_samples_split’: 5}
	NB	0.723	0.407	0.453	0.247	0.619	{}
	ADA	0.763	0.125	0.698	0.223	0.553	{‘learning_rate’: 0.1, ‘n_estimators’: 100}
	XGB	0.805	0.407	0.701	0.427	0.674	{‘n_estimators’: 100}
	NN	0.757	0.436	0.529	0.324	0.652	{‘alpha’: 0.001, ‘hidden_layer_sizes’: (50, 100, 50), ‘max_iter’: 50}
rDHP	SVM	0.786	0.191	0.859	0.340	0.590	{‘C’: 10, ‘gamma’: 0.1, ‘kernel’: ‘rbf’}
	RF	0.782	0.208	0.768	0.321	0.593	{‘max_depth’: None, ‘n_estimators’: 100}
	ET	0.791	0.236	0.806	0.359	0.608	{‘max_depth’: None, ‘n_estimators’: 200}
	KNN	0.784	0.305	0.669	0.344	0.627	{‘n_neighbors’: 5}
	LR	0.759	0.251	0.561	0.251	0.592	{‘C’: 10}
	DT	0.728	0.231	0.438	0.165	0.565	{‘max_depth’: 5, ‘min_samples_split’: 5}
	NB	0.719	0.413	0.445	0.242	0.618	{}
	ADA	0.754	0.097	0.607	0.166	0.538	{‘learning_rate’: 0.1, ‘n_estimators’: 100}
	XGB	0.789	0.376	0.650	0.377	0.653	{‘n_estimators’: 100}
	NN	0.776	0.390	0.593	0.348	0.649	{‘alpha’: 0.0001, ‘hidden_layer_sizes’: (100,), ‘max_iter’: 100}
rPolar	SVM	0.761	0.077	0.844	0.208	0.536	{‘C’: 10, ‘gamma’: 0.1, ‘kernel’: ‘rbf’}

Descriptor	Classifier	ACC	Recall	Precision	MCC	AUC	Parameter
	RF	0.784	0.205	0.800	0.331	0.594	{'max_depth': None, 'n_estimators': 100}
	ET	0.798	0.256	0.841	0.390	0.620	{'max_depth': None, 'n_estimators': 200}
	KNN	0.795	0.339	0.704	0.385	0.645	{'n_neighbors': 5}
	LR	0.756	0.214	0.556	0.227	0.578	{'C': 0.1}
	DT	0.741	0.222	0.484	0.192	0.571	{'max_depth': 5, 'min_samples_split': 10}
	NB	0.710	0.399	0.427	0.220	0.608	{}
	ADA	0.767	0.165	0.682	0.252	0.569	{'learning_rate': 0.1, 'n_estimators': 200}
	XGB	0.797	0.393	0.673	0.401	0.664	{'n_estimators': 200}
	NN	0.770	0.322	0.589	0.308	0.622	{'alpha': 0.01, 'hidden_layer_sizes': (100,), 'max_iter': 50}
rSecond	SVM	0.797	0.305	0.754	0.388	0.635	{'C': 10, 'gamma': 0.1, 'kernel': 'rbf'}
	RF	0.789	0.222	0.813	0.350	0.602	{'max_depth': None, 'n_estimators': 200}
	ET	0.790	0.219	0.837	0.357	0.602	{'max_depth': None, 'n_estimators': 100}
	KNN	0.787	0.345	0.658	0.363	0.642	{'n_neighbors': 5}
	LR	0.757	0.188	0.574	0.221	0.570	{'C': 0.1}
	DT	0.762	0.256	0.577	0.264	0.596	{'max_depth': 5, 'min_samples_split': 10}
	NB	0.735	0.390	0.476	0.260	0.621	{}
	ADA	0.768	0.179	0.670	0.258	0.575	{'learning_rate': 0.1, 'n_estimators': 200}
	XGB	0.801	0.422	0.676	0.420	0.676	{'n_estimators': 200}
	NN	0.777	0.336	0.615	0.332	0.632	{'alpha': 0.001, 'hidden_layer_sizes': (100,), 'max_iter': 50}

Table 3 Independent test results of 120 baseline models as developed with 10 ML algorithms and 12 feature encoding schemes.

Descriptor	Classifier	ACC	Recall	Precision	MCC	AUC	Parameter
ACC	SVM	0.802	0.322	0.757	0.402	0.643	{'C': 1, 'gamma': 0.1, 'kernel': 'rbf'}
	RF	0.793	0.333	0.690	0.375	0.641	{'max_depth': None, 'n_estimators': 200}
	ET	0.787	0.310	0.675	0.352	0.630	{'max_depth': None, 'n_estimators': 200}
	KNN	0.790	0.471	0.612	0.406	0.685	{'n_neighbors': 3}

Descriptor	Classifier	ACC	Recall	Precision	MCC	AUC	Parameter
	LR	0.755	0.161	0.560	0.197	0.559	{'C': 0.1}
	DT	0.778	0.276	0.649	0.316	0.613	{'max_depth': 5, 'min_samples_split': 5}
	NB	0.749	0.391	0.507	0.287	0.631	{}
	ADA	0.778	0.402	0.593	0.356	0.654	{'learning_rate': 1, 'n_estimators': 200}
	XGB	0.816	0.540	0.671	0.486	0.725	{'n_estimators': 200}
	NN	0.764	0.356	0.554	0.304	0.629	{'alpha': 0.0001, 'hidden_layer_sizes': (100,), 'max_iter': 100}
AAindex	SVM	0.735	0.034	0.300	0.018	0.504	{'C': 10, 'gamma': 0.1, 'kernel': 'rbf'}
	RF	0.761	0.241	0.568	0.251	0.589	{'max_depth': None, 'n_estimators': 200}
	ET	0.758	0.207	0.563	0.228	0.576	{'max_depth': None, 'n_estimators': 200}
	KNN	0.741	0.287	0.481	0.221	0.591	{'n_neighbors': 5}
	LR	0.755	0.161	0.560	0.197	0.559	{'C': 0.1}
	DT	0.735	0.264	0.460	0.196	0.579	{'max_depth': 5, 'min_samples_split': 10}
	NB	0.668	0.575	0.394	0.247	0.637	{}
	ADA	0.752	0.161	0.538	0.187	0.557	{'learning_rate': 0.1, 'n_estimators': 200}
	XGB	0.781	0.391	0.607	0.359	0.652	{'n_estimators': 200}
	NN	0.781	0.425	0.597	0.370	0.664	{'alpha': 0.001, 'hidden_layer_sizes': (100,), 'max_iter': 50}
APAAC	SVM	0.778	0.276	0.649	0.316	0.613	{'C': 1, 'gamma': 0.1, 'kernel': 'rbf'}
	RF	0.776	0.253	0.647	0.300	0.603	{'max_depth': None, 'n_estimators': 100}
	ET	0.799	0.345	0.714	0.395	0.649	{'max_depth': None, 'n_estimators': 100}
	KNN	0.810	0.552	0.649	0.476	0.725	{'n_neighbors': 5}
	LR	0.746	0.138	0.500	0.155	0.546	{'C': 0.1}
	DT	0.749	0.172	0.517	0.184	0.559	{'max_depth': 5, 'min_samples_split': 10}
	NB	0.711	0.230	0.385	0.127	0.552	{}
	ADA	0.770	0.333	0.580	0.310	0.626	{'learning_rate': 1, 'n_estimators': 100}
	XGB	0.773	0.391	0.576	0.338	0.647	{'n_estimators': 100}
	NN	0.781	0.425	0.597	0.370	0.664	{'alpha': 0.001, 'hidden_layer_sizes': (100,), 'max_iter': 100}

Descriptor	Classifier	ACC	Recall	Precision	MCC	AUC	Parameter
CTD	SVM	0.746	0.000	0.000	0.000	0.500	{'C': 0.1, 'gamma': 0.1, 'kernel': 'rbf'}
	RF	0.758	0.161	0.583	0.208	0.561	{'max_depth': None, 'n_estimators': 200}
	ET	0.761	0.161	0.609	0.219	0.563	{'max_depth': None, 'n_estimators': 200}
	KNN	0.732	0.287	0.455	0.202	0.585	{'n_neighbors': 5}
	LR	0.764	0.322	0.560	0.291	0.618	{'C': 0.1}
	DT	0.694	0.299	0.371	0.137	0.563	{'max_depth': 5, 'min_samples_split': 10}
	NB	0.606	0.575	0.338	0.169	0.596	{}
	ADA	0.770	0.161	0.700	0.255	0.569	{'learning_rate': 0.1, 'n_estimators': 200}
	XGB	0.764	0.310	0.563	0.286	0.614	{'n_estimators': 100}
	NN	0.743	0.425	0.493	0.291	0.638	{'alpha': 0.01, 'hidden_layer_sizes': (100,), 'max_iter': 50}
DPC	SVM	0.746	0.000	0.000	0.000	0.500	{'C': 0.1, 'gamma': 0.1, 'kernel': 'rbf'}
	RF	0.816	0.402	0.761	0.459	0.680	{'max_depth': None, 'n_estimators': 100}
	ET	0.813	0.391	0.756	0.448	0.674	{'max_depth': None, 'n_estimators': 200}
	KNN	0.787	0.563	0.583	0.432	0.713	{'n_neighbors': 5}
	LR	0.738	0.448	0.481	0.291	0.642	{'C': 0.1}
	DT	0.735	0.276	0.462	0.202	0.583	{'max_depth': 10, 'min_samples_split': 10}
	NB	0.504	0.851	0.320	0.220	0.619	{}
	ADA	0.793	0.230	0.833	0.365	0.607	{'learning_rate': 0.1, 'n_estimators': 200}
	XGB	0.787	0.448	0.609	0.392	0.675	{'n_estimators': 100}
	NN	0.767	0.437	0.551	0.343	0.658	{'alpha': 0.0001, 'hidden_layer_sizes': (100,), 'max_iter': 100}
PAAC	SVM	0.773	0.264	0.622	0.294	0.605	{'C': 1, 'gamma': 0.1, 'kernel': 'rbf'}
	RF	0.767	0.218	0.613	0.260	0.586	{'max_depth': 10, 'n_estimators': 50}
	ET	0.802	0.345	0.732	0.405	0.651	{'max_depth': None, 'n_estimators': 200}
	KNN	0.776	0.575	0.556	0.414	0.709	{'n_neighbors': 3}
	LR	0.758	0.195	0.567	0.223	0.572	{'C': 1}
	DT	0.735	0.264	0.460	0.196	0.579	{'max_depth': 5, 'min_samples_split': 10}

Descriptor	Classifier	ACC	Recall	Precision	MCC	AUC	Parameter
	NB	0.717	0.241	0.404	0.146	0.560	{}
	ADA	0.755	0.149	0.565	0.192	0.555	{‘learning_rate’: 0.1, ‘n_estimators’: 200}
	XGB	0.778	0.437	0.585	0.368	0.666	{‘n_estimators’: 200}
	NN	0.802	0.460	0.656	0.430	0.689	{‘alpha’: 0.01, ‘hidden_layer_sizes’: (100,), ‘max_iter’: 100}
PCP	SVM	0.752	0.092	0.571	0.151	0.534	{‘C’: 1, ‘gamma’: 1, ‘kernel’: ‘rbf’}
	RF	0.732	0.161	0.424	0.128	0.543	{‘max_depth’: 10, ‘n_estimators’: 50}
	ET	0.735	0.195	0.447	0.157	0.557	{‘max_depth’: None, ‘n_estimators’: 100}
	KNN	0.708	0.218	0.373	0.114	0.547	{‘n_neighbors’: 5}
	LR	0.746	0.000	0.000	0.000	0.500	{‘C’: 0.1}
	DT	0.720	0.184	0.390	0.116	0.543	{‘max_depth’: 5, ‘min_samples_split’: 2}
	NB	0.723	0.195	0.405	0.130	0.549	{}
	ADA	0.735	0.069	0.375	0.062	0.515	{‘learning_rate’: 0.1, ‘n_estimators’: 200}
	XGB	0.714	0.310	0.415	0.180	0.581	{‘n_estimators’: 200}
	NN	0.743	0.103	0.474	0.122	0.532	{‘alpha’: 0.001, ‘hidden_layer_sizes’: (100,), ‘max_iter’: 50}
rAcid	SVM	0.784	0.379	0.623	0.363	0.651	{‘C’: 10, ‘gamma’: 0.1, ‘kernel’: ‘rbf’}
	RF	0.755	0.241	0.538	0.234	0.586	{‘max_depth’: None, ‘n_estimators’: 100}
	ET	0.781	0.310	0.643	0.334	0.626	{‘max_depth’: None, ‘n_estimators’: 200}
	KNN	0.784	0.494	0.589	0.401	0.689	{‘n_neighbors’: 3}
	LR	0.790	0.276	0.727	0.355	0.620	{‘C’: 10}
	DT	0.778	0.299	0.634	0.322	0.620	{‘max_depth’: 5, ‘min_samples_split’: 2}
	NB	0.755	0.494	0.518	0.343	0.669	{}
	ADA	0.761	0.207	0.581	0.237	0.578	{‘learning_rate’: 0.1, ‘n_estimators’: 200}
	XGB	0.773	0.402	0.574	0.342	0.650	{‘n_estimators’: 200}
	NN	0.784	0.437	0.603	0.381	0.670	{‘alpha’: 0.001, ‘hidden_layer_sizes’: (100,), ‘max_iter’: 100}
rCharge	SVM	0.784	0.356	0.633	0.356	0.643	{‘C’: 10, ‘gamma’: 0.1, ‘kernel’: ‘rbf’}

Descriptor	Classifier	ACC	Recall	Precision	MCC	AUC	Parameter
	RF	0.793	0.333	0.690	0.375	0.641	{'max_depth': None, 'n_estimators': 200}
	ET	0.776	0.322	0.609	0.321	0.626	{'max_depth': None, 'n_estimators': 50}
	KNN	0.784	0.368	0.627	0.359	0.647	{'n_neighbors': 5}
	LR	0.764	0.207	0.600	0.246	0.580	{'C': 1}
	DT	0.761	0.230	0.571	0.246	0.586	{'max_depth': 5, 'min_samples_split': 5}
	NB	0.743	0.448	0.494	0.302	0.646	{}
	ADA	0.770	0.161	0.700	0.255	0.569	{'learning_rate': 0.1, 'n_estimators': 100}
	XGB	0.796	0.483	0.627	0.423	0.693	{'n_estimators': 100}
	NN	0.752	0.437	0.514	0.313	0.648	{'alpha': 0.001, 'hidden_layer_sizes': (50, 100, 50), 'max_iter': 50}
rDHP	SVM	0.784	0.241	0.724	0.329	0.605	{'C': 10, 'gamma': 0.1, 'kernel': 'rbf'}
	RF	0.770	0.241	0.618	0.278	0.595	{'max_depth': None, 'n_estimators': 100}
	ET	0.776	0.241	0.656	0.297	0.599	{'max_depth': None, 'n_estimators': 200}
	KNN	0.776	0.345	0.600	0.329	0.633	{'n_neighbors': 5}
	LR	0.776	0.264	0.639	0.303	0.607	{'C': 10}
	DT	0.778	0.253	0.667	0.310	0.605	{'max_depth': 5, 'min_samples_split': 5}
	NB	0.720	0.437	0.447	0.255	0.627	{}
	ADA	0.776	0.149	0.813	0.284	0.569	{'learning_rate': 0.1, 'n_estimators': 100}
	XGB	0.822	0.494	0.717	0.490	0.714	{'n_estimators': 100}
	NN	0.767	0.437	0.551	0.343	0.658	{'alpha': 0.0001, 'hidden_layer_sizes': (100,), 'max_iter': 100}
rPolar	SVM	0.758	0.126	0.611	0.193	0.550	{'C': 10, 'gamma': 0.1, 'kernel': 'rbf'}
	RF	0.778	0.241	0.677	0.307	0.601	{'max_depth': None, 'n_estimators': 100}
	ET	0.781	0.276	0.667	0.325	0.614	{'max_depth': None, 'n_estimators': 200}
	KNN	0.787	0.345	0.652	0.360	0.641	{'n_neighbors': 5}
	LR	0.776	0.241	0.656	0.297	0.599	{'C': 0.1}
	DT	0.767	0.287	0.581	0.285	0.609	{'max_depth': 5, 'min_samples_split': 10}
	NB	0.714	0.529	0.447	0.291	0.653	{}

Descriptor	Classifier	ACC	Recall	Precision	MCC	AUC	Parameter
	ADA	0.764	0.195	0.607	0.242	0.576	{'learning_rate': 0.1, 'n_estimators': 200}
	XGB	0.802	0.448	0.661	0.427	0.685	{'n_estimators': 200}
	NN	0.781	0.402	0.603	0.363	0.656	{'alpha': 0.01, 'hidden_layer_sizes': (100,), 'max_iter': 50}
rSecond	SVM	0.776	0.356	0.596	0.333	0.637	{'C': 10, 'gamma': 0.1, 'kernel': 'rbf'}
	RF	0.778	0.230	0.690	0.305	0.597	{'max_depth': None, 'n_estimators': 200}
	ET	0.787	0.253	0.733	0.341	0.611	{'max_depth': None, 'n_estimators': 100}
	KNN	0.778	0.345	0.612	0.336	0.635	{'n_neighbors': 5}
	LR	0.758	0.195	0.567	0.223	0.572	{'C': 0.1}
	DT	0.746	0.241	0.500	0.212	0.580	{'max_depth': 5, 'min_samples_split': 10}
	NB	0.708	0.391	0.420	0.212	0.604	{}
	ADA	0.764	0.138	0.667	0.223	0.557	{'learning_rate': 0.1, 'n_estimators': 200}
	XGB	0.802	0.402	0.686	0.416	0.670	{'n_estimators': 200}
	NN	0.778	0.368	0.604	0.344	0.643	{'alpha': 0.001, 'hidden_layer_sizes': (100,), 'max_iter': 50}

Feature engineering and examination

This work produced 120 new features (120 F) by computing the predictive probabilities of the anti-HIV peptide classification model using 12 descriptors (d) and 10 classifiers (c). Subsequently, these characteristics were utilized to provide new data, construct an alternative model with 10 classifiers, and evaluate the model's performance, as represented in **Table 4**. Upon analyzing the results of the model performance testing, that was discovered the ACC, MCC, AUC, and F1 score values were notably high when using LR or logistic regression. Specifically, the values achieved in the CV dataset were 0.84, 0.52, 0.73, and 0.61, respectively (**Figure 8**). The highest values achieved for the IN dataset (**Figure 9**) were 0.84 for accuracy (ACC), 0.62 for recall, 0.61 for Matthew's correlation coefficient (MCC), 0.78 for area under the curve (AUC), and 0.70 for F1 score. The scores demonstrated the efficacy of the model constructed using the Naive Bayes technique.

Table 4 Performance evaluation results of the model built from 120 NF with 10 classifiers, with cross-validation set (CV) and independent set (IN).

Classifier	ACC		Recall		MCC		AUC		F1	
	CV	IN	CV	IN	CV	IN	CV	IN	CV	IN
SVC	0.796	0.754	0.515	0.286	0.418	0.351	0.700	0.622	0.548	0.414
RF	0.821	0.797	0.455	0.381	0.461	0.492	0.696	0.680	0.550	0.533
ET	0.836	0.797	0.485	0.381	0.508	0.492	0.716	0.680	0.587	0.533
KNN	0.832	0.783	0.424	0.333	0.488	0.449	0.693	0.656	0.549	0.483

Classifier	ACC		Recall		MCC		AUC		F1	
	CV	IN	CV	IN	CV	IN	CV	IN	CV	IN
LR	0.836	0.826	0.530	0.429	0.517	0.586	0.731	0.714	0.609	0.600
DT	0.781	0.826	0.455	0.524	0.365	0.567	0.670	0.741	0.500	0.647
NB	0.774	0.841	0.591	0.619	0.407	0.607	0.711	0.778	0.557	0.703
ADA	0.807	0.797	0.394	0.381	0.406	0.492	0.666	0.680	0.495	0.533
XGB	0.810	0.797	0.470	0.429	0.436	0.486	0.694	0.693	0.544	0.563
NN	0.818	0.812	0.515	0.381	0.467	0.548	0.714	0.690	0.576	0.552

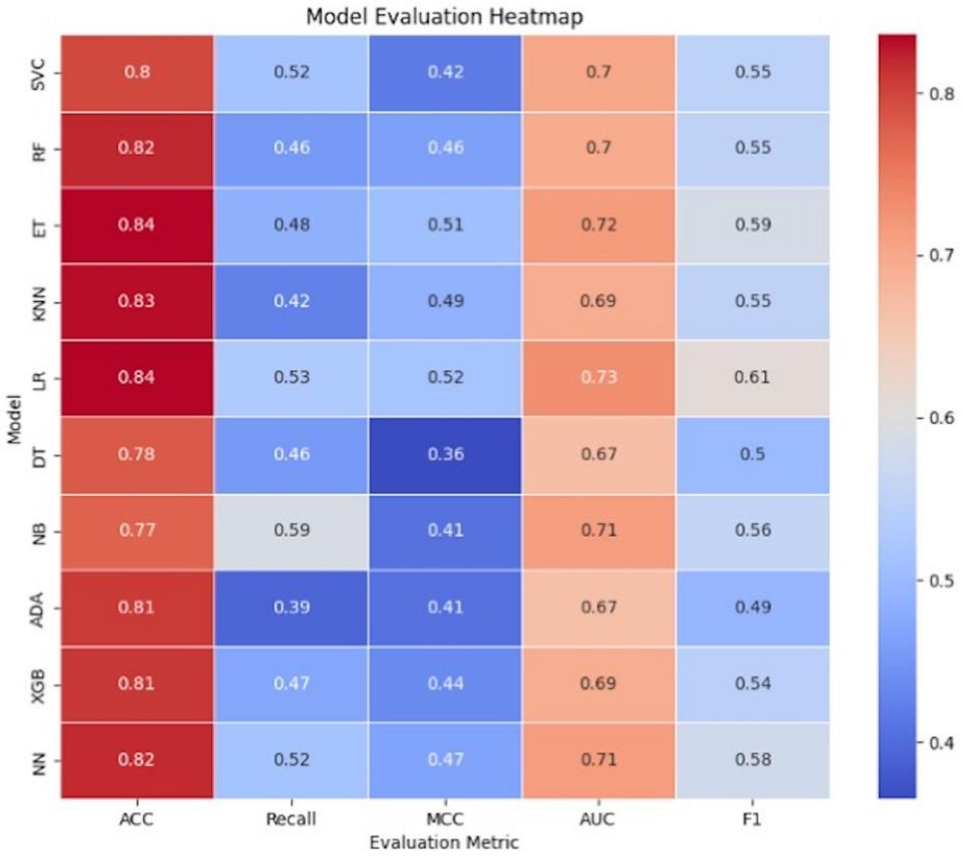

Figure 8 Heatmap showing the cross-validation performance of 10 machine learning classifiers (SVM, RF, ET, KNN, LR, DT, NB, ADA, XGB, and NN) using 120 selected features (120 F). Color intensity represents classification accuracy, with darker shades indicating higher performance. This visualization highlights which classifiers achieve the best predictive power under the given feature set.

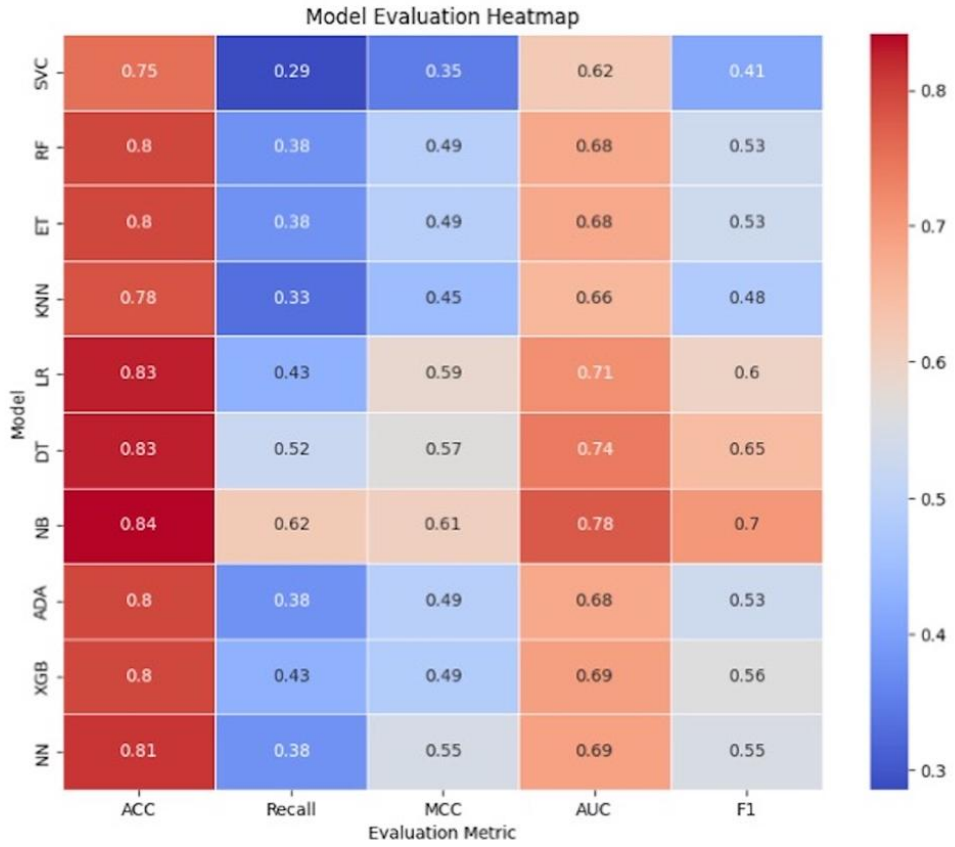

Figure 9 The heatmap depicts the predictive performance of 10 machine learning classifiers (columns) across 120 features (rows) from the IN dataset. Color intensity denotes balanced accuracy, with darker tones signifying better performance. This graph shows which classifiers perform best for feature subsets.

The procedure includes selecting features that improve efficiency and assessing the model performance

This study utilized 4 feature selection methods: Chi square, ANOVA, mutual information, and genetic algorithm. Next, the features selected are utilized to generate a model, and the performance evaluation results of the model are then compared. As indicated in **Tables 5** and **6**. The performance evaluation results of the models developed through the feature selection process are presented in **Table 5**. The models were evaluated using 3 different methods: chi square, ANOVA, and mutual information. The evaluation was done with varying numbers of features (k), including 12, 30, 60, and 90 features. **Table 6** presents the performance evaluation results of the models created using the feature selection process in the genetic algorithm (GA) with various methods. The effectiveness of 4 models: Chi2_12k_nn, anova_30k_lr, m_12k_nb, and GAXGB were assessed. The chi2_12k_nn model employs the chi square method with 12 features and a neural network classifier. The anova_30k_lr model utilizes the ANOVA method with 30 features and logistic regression. The m_12k_nb model applies a mutual information method with 12 features and naive bayes. Lastly, the GAXGB model employs a genetic algorithm to create a model with an XGBoost classifier that can summarize the set cross-validation. The graph in **Figure 10** shows that the model developed using the ANOVA method had the greatest values for ACC, MCC, and F1-score, which were 0.843, 0.545, and 0.639, respectively. The model that employed mutual information, chi-square, and genetic algorithms achieved the second rank. In addition, the independent testing set demonstrates that the feature-selective technique, which utilizes a genetic algorithm, produces results for assessing model performance based on ACC, MCC, and F1-score, with respective values of 0.913, 0.797, and 0.833.

Table 5 presenting the performance evaluation results of the models developed through the feature selection process in all 3 kinds: Chi square, ANOVA, and mutual information with varying numbers of features (k).

Method	k	Classifier	ACC		MCC		F1	
			CV	IN	CV	IN	CV	IN
Chi square	12	SVC	0.821	0.797	0.464	0.486	0.559	0.563
		RF	0.825	0.826	0.489	0.586	0.593	0.600
		ET	0.836	0.797	0.511	0.492	0.595	0.533
		KNN	0.810	0.797	0.424	0.492	0.519	0.533
		LR	0.832	0.855	0.498	0.658	0.582	0.688
		DT	0.774	0.754	0.349	0.361	0.492	0.485
		NB	0.814	0.841	0.499	0.606	0.622	0.686
		ADA	0.803	0.768	0.380	0.404	0.449	0.429
		XGB	0.752	0.812	0.286	0.548	0.443	0.552
		NN	0.814	0.870	0.450	0.684	0.557	0.743
	30	SVC	0.836	0.797	0.513	0.486	0.602	0.563
		RF	0.839	0.812	0.523	0.533	0.607	0.581
		ET	0.839	0.826	0.521	0.572	0.600	0.625
		KNN	0.825	0.783	0.471	0.449	0.556	0.483
		LR	0.818	0.841	0.459	0.611	0.561	0.667
		DT	0.788	0.812	0.380	0.527	0.508	0.606
		NB	0.796	0.826	0.453	0.568	0.588	0.667
		ADA	0.847	0.768	0.546	0.398	0.625	0.467
		XGB	0.836	0.783	0.517	0.467	0.609	0.444
		NN	0.821	0.812	0.472	0.528	0.574	0.629
	60	SVC	0.821	0.812	0.489	0.527	0.602	0.606
		RF	0.821	0.826	0.464	0.586	0.559	0.600
		ET	0.847	0.812	0.544	0.533	0.618	0.581
		KNN	0.832	0.739	0.492	0.314	0.566	0.437
		LR	0.810	0.841	0.428	0.611	0.527	0.667
		DT	0.796	0.826	0.418	0.567	0.548	0.647
		NB	0.781	0.826	0.420	0.572	0.565	0.684
		ADA	0.839	0.783	0.518	0.449	0.593	0.483
		XGB	0.828	0.783	0.491	0.449	0.584	0.483
		NN	0.810	0.812	0.441	0.533	0.552	0.581
	90	SVC	0.843	0.797	0.534	0.492	0.613	0.533
		RF	0.832	0.826	0.501	0.586	0.589	0.600
		ET	0.832	0.826	0.492	0.572	0.566	0.625
		KNN	0.832	0.768	0.501	0.400	0.589	0.500
		LR	0.814	0.812	0.446	0.548	0.549	0.552
		DT	0.796	0.826	0.407	0.567	0.533	0.647
		NB	0.774	0.826	0.407	0.572	0.557	0.684
		ADA	0.828	0.783	0.491	0.449	0.584	0.483
		XGB	0.821	0.797	0.476	0.492	0.581	0.533
		NN	0.818	0.841	0.463	0.622	0.569	0.645
ANOVA	12	SVC	0.839	0.841	0.529	0.611	0.621	0.667
		RF	0.839	0.826	0.526	0.572	0.614	0.625

Method	k	Classifier	ACC		MCC		F1	
			CV	IN	CV	IN	CV	IN
		ET	0.850	0.826	0.555	0.572	0.624	0.625
		KNN	0.839	0.783	0.536	0.444	0.633	0.545
		LR	0.836	0.841	0.508	0.622	0.587	0.645
		DT	0.807	0.797	0.428	0.486	0.539	0.563
		NB	0.821	0.826	0.514	0.567	0.632	0.647
		ADA	0.839	0.812	0.529	0.533	0.621	0.581
		XGB	0.810	0.812	0.436	0.527	0.544	0.606
		NN	0.825	0.826	0.489	0.586	0.593	0.600
		SVC	0.861	0.797	0.590	0.492	0.655	0.533
		RF	0.828	0.826	0.487	0.572	0.577	0.625
	30	ET	0.836	0.812	0.513	0.533	0.602	0.581
		KNN	0.836	0.768	0.508	0.404	0.587	0.429
		LR	0.843	0.855	0.545	0.658	0.639	0.688
		DT	0.781	0.826	0.359	0.567	0.492	0.647
		NB	0.788	0.841	0.439	0.607	0.580	0.703
		ADA	0.850	0.783	0.561	0.443	0.643	0.516
		XGB	0.839	0.812	0.533	0.548	0.627	0.552
		NN	0.832	0.812	0.507	0.548	0.603	0.552
		SVC	0.825	0.797	0.461	0.508	0.520	0.500
		RF	0.832	0.797	0.498	0.492	0.582	0.533
		ET	0.843	0.812	0.529	0.548	0.598	0.552
		KNN	0.843	0.783	0.529	0.449	0.598	0.483
		LR	0.825	0.826	0.478	0.572	0.571	0.625
	60	DT	0.770	0.812	0.329	0.527	0.471	0.606
		NB	0.788	0.826	0.433	0.572	0.574	0.684
		ADA	0.828	0.783	0.491	0.449	0.584	0.483
		XGB	0.825	0.797	0.481	0.492	0.579	0.533
		NN	0.825	0.812	0.474	0.533	0.564	0.581
		SVC	0.839	0.841	0.521	0.622	0.600	0.645
		RF	0.843	0.826	0.529	0.586	0.598	0.600
		ET	0.843	0.812	0.531	0.533	0.606	0.581
		KNN	0.843	0.783	0.534	0.449	0.613	0.483
		LR	0.836	0.826	0.517	0.586	0.609	0.600
	90	DT	0.785	0.826	0.367	0.567	0.496	0.647
		NB	0.763	0.826	0.382	0.572	0.539	0.684
		ADA	0.821	0.783	0.464	0.449	0.559	0.483
		XGB	0.821	0.797	0.472	0.492	0.574	0.533
		NN	0.832	0.826	0.501	0.586	0.589	0.600
		SVC	0.850	0.812	0.552	0.548	0.610	0.552
		RF	0.858	0.797	0.578	0.486	0.642	0.563
Mutual Information	12	ET	0.854	0.826	0.571	0.572	0.649	0.625
		KNN	0.843	0.826	0.539	0.567	0.626	0.647
		LR	0.854	0.841	0.567	0.622	0.636	0.645
		DT	0.799	0.797	0.432	0.486	0.560	0.563

Method	k	Classifier	ACC		MCC		F1	
			CV	IN	CV	IN	CV	IN
	30	NB	0.818	0.855	0.506	0.645	0.627	0.737
		ADA	0.843	0.812	0.534	0.548	0.613	0.552
		XGB	0.818	0.797	0.472	0.492	0.583	0.533
		NN	0.854	0.826	0.576	0.572	0.661	0.625
		SVC	0.847	0.783	0.542	0.449	0.611	0.483
		RF	0.836	0.797	0.513	0.486	0.602	0.563
		ET	0.854	0.812	0.569	0.533	0.643	0.581
		KNN	0.818	0.783	0.451	0.443	0.545	0.516
		LR	0.843	0.812	0.531	0.548	0.606	0.552
		DT	0.781	0.826	0.371	0.572	0.508	0.625
	60	NB	0.799	0.826	0.471	0.572	0.604	0.684
		ADA	0.821	0.797	0.494	0.487	0.608	0.588
		XGB	0.818	0.797	0.463	0.492	0.569	0.533
		NN	0.847	0.812	0.551	0.548	0.638	0.552
		SVC	0.847	0.841	0.551	0.611	0.638	0.667
		RF	0.836	0.826	0.500	0.586	0.554	0.600
		ET	0.847	0.826	0.542	0.572	0.611	0.625
		KNN	0.847	0.826	0.542	0.572	0.611	0.625
		LR	0.843	0.812	0.534	0.533	0.613	0.581
		DT	0.788	0.812	0.386	0.527	0.517	0.606
	90	NB	0.777	0.826	0.413	0.568	0.561	0.667
		ADA	0.814	0.768	0.441	0.404	0.541	0.429
		XGB	0.828	0.797	0.494	0.492	0.591	0.533
		NN	0.861	0.797	0.592	0.492	0.661	0.533
		SVC	0.847	0.812	0.539	0.548	0.580	0.552
		RF	0.814	0.826	0.437	0.586	0.532	0.600
		ET	0.854	0.841	0.565	0.622	0.623	0.645
		KNN	0.847	0.812	0.539	0.533	0.596	0.581
		LR	0.839	0.826	0.521	0.586	0.600	0.600
		DT	0.781	0.797	0.395	0.486	0.538	0.563
		NB	0.781	0.826	0.420	0.572	0.565	0.684
		ADA	0.818	0.797	0.455	0.492	0.554	0.533
		XGB	0.799	0.826	0.416	0.586	0.538	0.600
		NN	0.836	0.841	0.513	0.622	0.602	0.645

Table 6 Displaying the performance evaluation results of the models created using the feature selection process in genetic algorithm (GA) with different methods.

GA method	Classifier	ACC		MCC		F1 score	
		CV	IN	CV	IN	CV	IN
Support Vector Machine (svc)	SVC	0.843	0.826	0.531	0.586	0.606	0.600
	RF	0.850	0.797	0.552	0.508	0.610	0.500
	ET	0.850	0.855	0.552	0.658	0.602	0.688
	KNN	0.832	0.841	0.490	0.622	0.558	0.645

GA method	Classifier	ACC		MCC		F1 score	
		CV	IN	CV	IN	CV	IN
	LR	0.839	0.870	0.521	0.694	0.600	0.727
	DT	0.785	0.797	0.384	0.492	0.520	0.533
	NB	0.755	0.855	0.362	0.650	0.525	0.750
	ADA	0.810	0.797	0.424	0.492	0.519	0.533
	XGB	0.821	0.812	0.472	0.548	0.574	0.552
	NN	0.810	0.841	0.432	0.607	0.536	0.703
AdaBoost (ada)	SVC	0.847	0.826	0.546	0.586	0.625	0.600
	RF	0.825	0.812	0.474	0.548	0.564	0.552
	ET	0.828	0.826	0.475	0.586	0.535	0.600
	KNN	0.814	0.826	0.446	0.586	0.549	0.600
	LR	0.832	0.812	0.498	0.533	0.582	0.581
	DT	0.803	0.797	0.419	0.492	0.534	0.533
	NB	0.766	0.826	0.388	0.572	0.543	0.684
	ADA	0.832	0.870	0.498	0.694	0.582	0.727
	XGB	0.796	0.855	0.402	0.658	0.525	0.688
	NN	0.807	0.797	0.432	0.492	0.547	0.533
Decision Trees (dt)	SVC	0.825	0.768	0.471	0.398	0.556	0.467
	RF	0.799	0.812	0.381	0.548	0.476	0.552
	ET	0.807	0.812	0.395	0.548	0.465	0.552
	KNN	0.828	0.783	0.473	0.444	0.525	0.545
	LR	0.818	0.841	0.441	0.611	0.519	0.667
	DT	0.766	0.870	0.335	0.694	0.484	0.727
	NB	0.752	0.855	0.356	0.645	0.521	0.737
	ADA	0.781	0.812	0.311	0.548	0.412	0.552
	XGB	0.785	0.797	0.372	0.486	0.504	0.563
	NN	0.836	0.841	0.517	0.606	0.609	0.686
Extra-Trees (et)	SVC	0.843	0.783	0.534	0.444	0.613	0.545
	RF	0.836	0.841	0.511	0.622	0.595	0.645
	ET	0.821	0.826	0.447	0.586	0.505	0.600
	KNN	0.814	0.855	0.430	0.658	0.514	0.688
	LR	0.843	0.841	0.531	0.611	0.606	0.667
	DT	0.766	0.797	0.335	0.492	0.484	0.533
	NB	0.777	0.841	0.420	0.611	0.567	0.718
	ADA	0.810	0.841	0.428	0.622	0.527	0.645
	XGB	0.810	0.754	0.446	0.354	0.559	0.452
	NN	0.818	0.812	0.467	0.527	0.576	0.606
K-Nearest Neighbor (knn)	SVC	0.843	0.826	0.529	0.586	0.598	0.600
	RF	0.828	0.797	0.484	0.492	0.569	0.533
	ET	0.850	0.812	0.552	0.533	0.610	0.581

GA method	Classifier	ACC		MCC		F1 score	
		CV	IN	CV	IN	CV	IN
	KNN	0.818	0.899	0.438	0.763	0.510	0.800
	LR	0.843	0.826	0.534	0.586	0.613	0.600
	DT	0.839	0.783	0.529	0.443	0.621	0.516
	NB	0.774	0.812	0.407	0.532	0.557	0.649
	ADA	0.810	0.826	0.450	0.567	0.567	0.647
	XGB	0.810	0.812	0.432	0.533	0.536	0.581
	NN	0.810	0.855	0.450	0.658	0.567	0.688
Logistic Regression (lr)	SVC	0.799	0.884	0.410	0.728	0.530	0.765
	RF	0.810	0.812	0.436	0.548	0.544	0.552
	ET	0.814	0.841	0.437	0.622	0.532	0.645
	KNN	0.803	0.754	0.400	0.351	0.500	0.414
	LR	0.825	0.899	0.478	0.763	0.571	0.800
	DT	0.777	0.783	0.363	0.443	0.504	0.516
	NB	0.763	0.826	0.375	0.572	0.532	0.684
	ADA	0.810	0.826	0.450	0.567	0.567	0.647
	XGB	0.796	0.812	0.413	0.548	0.541	0.552
	NN	0.792	0.797	0.394	0.492	0.521	0.533
Naive Bayes (nb)	SVC	0.828	0.826	0.487	0.572	0.577	0.625
	RF	0.825	0.783	0.468	0.449	0.547	0.483
	ET	0.828	0.812	0.479	0.548	0.552	0.552
	KNN	0.828	0.797	0.477	0.508	0.544	0.500
	LR	0.828	0.826	0.487	0.586	0.577	0.600
	DT	0.752	0.754	0.356	0.383	0.521	0.541
	NB	0.759	0.870	0.375	0.683	0.535	0.769
	ADA	0.818	0.783	0.435	0.449	0.500	0.483
	XGB	0.814	0.768	0.446	0.406	0.549	0.529
	NN	0.821	0.841	0.480	0.622	0.588	0.645
Neural Network (nn)	SVC	0.847	0.855	0.544	0.658	0.618	0.688
	RF	0.814	0.812	0.434	0.533	0.523	0.581
	ET	0.850	0.812	0.552	0.533	0.610	0.581
	KNN	0.818	0.768	0.455	0.406	0.554	0.529
	LR	0.839	0.870	0.518	0.694	0.593	0.727
	DT	0.777	0.754	0.388	0.396	0.534	0.564
	NB	0.763	0.826	0.382	0.572	0.539	0.684
	ADA	0.825	0.783	0.489	0.467	0.593	0.444
	XGB	0.807	0.812	0.437	0.533	0.555	0.581
	NN	0.832	0.841	0.515	0.606	0.617	0.686
Random Forest (rf)	SVC	0.843	0.826	0.531	0.586	0.606	0.600
	RF	0.825	0.841	0.474	0.622	0.564	0.645

GA method	Classifier	ACC		MCC		F1 score	
		CV	IN	CV	IN	CV	IN
	ET	0.818	0.797	0.444	0.508	0.528	0.500
	KNN	0.818	0.826	0.441	0.586	0.519	0.600
	LR	0.832	0.855	0.501	0.658	0.589	0.688
	DT	0.763	0.754	0.341	0.354	0.496	0.452
	NB	0.770	0.855	0.401	0.645	0.553	0.737
	ADA	0.814	0.812	0.430	0.548	0.514	0.552
	XGB	0.807	0.797	0.432	0.492	0.547	0.533
	NN	0.828	0.855	0.498	0.658	0.598	0.688
XGBoosting (xgb)	SVC	0.828	0.797	0.481	0.486	0.561	0.563
	RF	0.814	0.797	0.446	0.492	0.549	0.533
	ET	0.825	0.826	0.468	0.586	0.547	0.600
	KNN	0.825	0.812	0.465	0.548	0.538	0.552
	LR	0.836	0.841	0.505	0.611	0.579	0.667
	DT	0.777	0.812	0.363	0.548	0.504	0.552
	NB	0.755	0.826	0.362	0.572	0.525	0.684
	ADA	0.818	0.797	0.444	0.492	0.528	0.533
	XGB	0.774	0.913	0.343	0.797	0.483	0.833
	NN	0.814	0.826	0.454	0.586	0.564	0.600

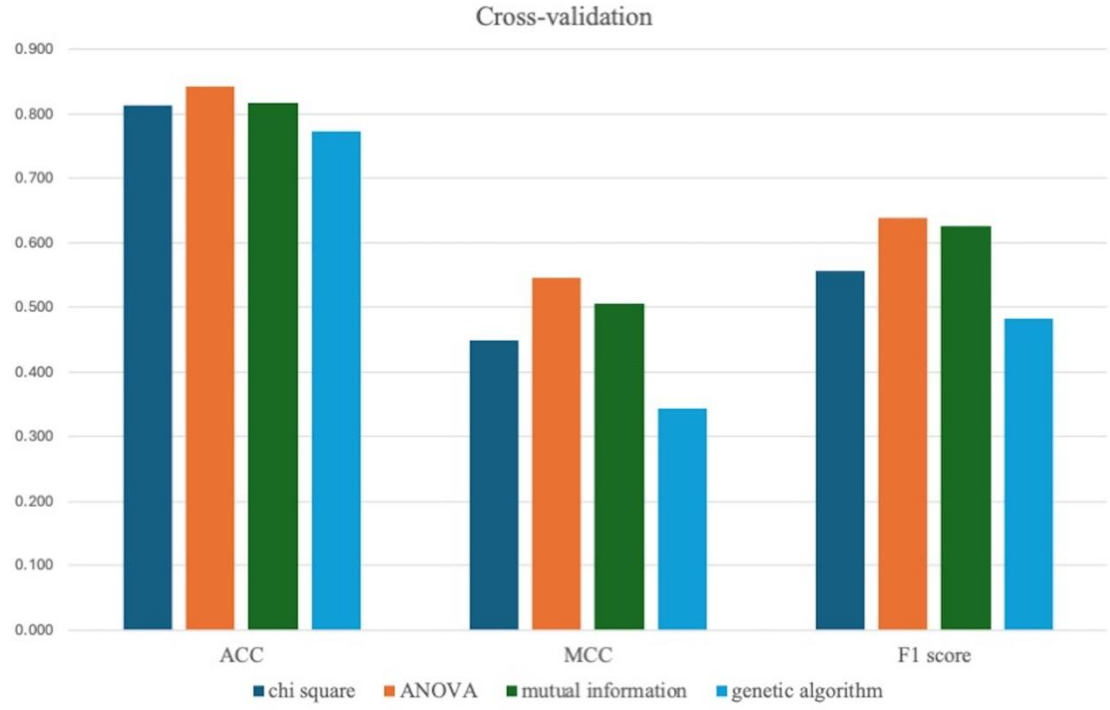

Figure 10 Model performance was compared using 4 feature selection methods - Chi-square, ANOVA, Mutual Information, and Genetic Algorithm - on a cross-validation dataset. Genetic Algorithm-selected models earned the highest prediction accuracy and F1-score, demonstrating their usefulness in feature optimization for this dataset.

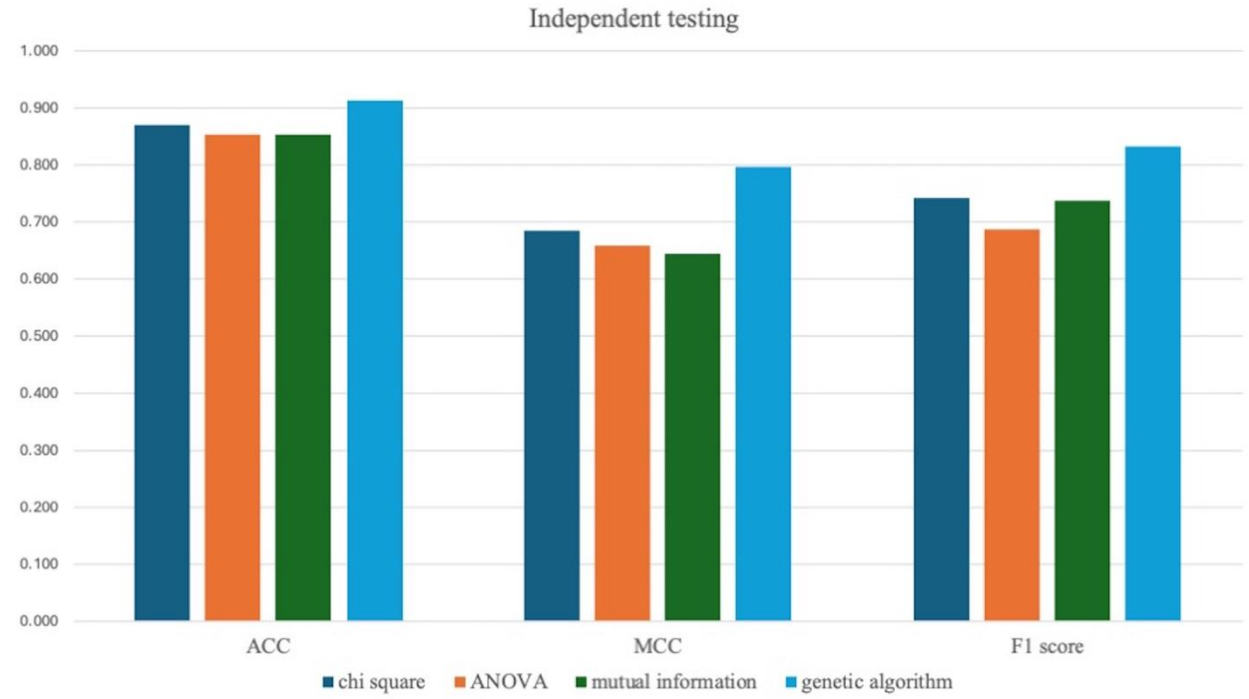

Figure 11 Comparison of predictive performance for the top models selected using 4 feature selection methods: Chi-square, ANOVA, mutual information, and genetic algorithms. Models were evaluated on an independent test set, showing accuracy, F1-score, and area under the ROC curve (AUC) to highlight differences in model effectiveness.

The model elements thoroughly and the significance of all features analyzed

Following this step, the impact of features are examined, which are the predictive probability values derived from a model constructed using 12 descriptors and 10 classifiers. Then utilize the genetic algorithm of the GAXGB model to choose the most suitable features. The SHAP value analytical findings are shown in **Figure 12**. There will be a total of 10 ranks. The initial significant characteristics include dpc_XGB, rcharge_XGB, apaac_ET, rcharge_RF, dpc_LR, paac_ET, paac_KNN, paac_SVM, rcharge_ET, and rsecond_XGB. Most of the red dots represent high-value attributes in the right pane. It demonstrates the impact on the model's forecast.

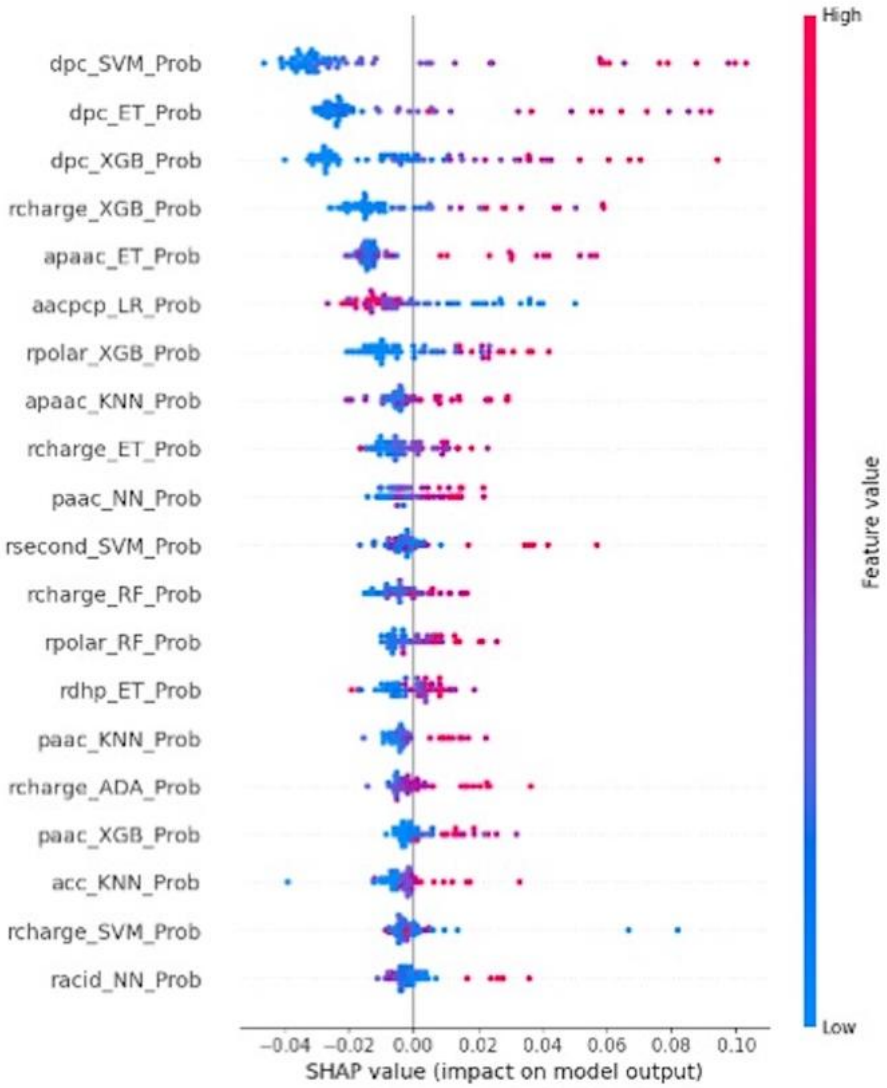

Figure 12 Feature importance analysis using SHAP values for the GAXGB model. The top 5 features, including AAC1 and DPC3, have the strongest influence on model predictions, indicating key sequence patterns associated with anti-HIV activity.

Discussion

Dataset and baseline model examination

A machine learning model to classify anti-HIV peptides are developed. A key factor in the method of learning is the level of quality of the data utilized. The GAXGB model was modified using research data from Jaru and the collaboration, which was collected from public databases. Subsequently, manual and CD-hit programs re applied to reduce data duplication, achieving an 80% threshold level. The cd-hit software is commonly used in bioinformatics research to generate non-redundant datasets. This is achieved by clustering sequences based on a specified identity threshold, such as 80% or 90%, to remove duplicate sequences. This step is essential for analyzing large-scale datasets [35].

The amino acid density analysis indicates that the anti-HIV peptide dataset contains significant amounts of amino acid sequences. The main amino acids include isoleucine (I), glutamic acid (E), lysine (K), glutamine (Q), tryptophan (W), glycine (G), and cysteine (C). [19]. In addition, examining the characteristics of the positive data set presents physicochemical components that are defined by small molecular sizes and short side chains. This leads to high flexibility and is commonly observed in regions of proteins with high rotation. Neutral qualities refer to features that do not generate either positive or negative electrical charges in an environment

with a neutral pH level, which is roughly pH 7. These amino acids do not directly induce chemical reactions associated with acidity or alkalinity [36,37]. The following step is to generate a baseline model by extracting data using 12 descriptors. Each descriptor's data is then utilized to construct a model with 10 classifiers, resulting in a cumulative count of 120 baseline models. During analysis of these models, it is apparent that the descriptors employed were developed a model to assess efficiency and found that AAC and DPC demonstrated elevated levels of ACC and MCC [38,39]. AAC stands for Amino Acid Composition, which calculates the amount of each of the 20 amino acids in a protein/peptide sequence, while DPC, or Dipeptide Composition, calculates the frequency of consecutive amino acid (dipeptide) pairings in a protein sequence.

Data engineering, feature selection and candidate model performance

Whenever the 120 baseline, models have been established, use this data to generate new data by computing the predictive probability based on the baseline model. The process of data engineering will generate new features based on the 120 baseline models. Next, apply the newly gathered data to build the model. The model was constructed using 10 classifiers and subsequently evaluated for its performance using an independent dataset. The Naive Bayes model achieved the highest rating score. Engineering data is using predictive probability values to construct models for classifying peptides based on their varied properties [40-42]. The building of the GAXGB model includes both data engineering and applying of 4 feature selection methods: Chi square, ANOVA, mutual information, and genetic algorithm. The genetic algorithm method generates the greatest assessment values for ACC, MCC, and F1-score in the independent testing set, showing higher efficiency. The use of the Chi-Square test and reduced redundancy for selecting features in machine learning models [43]. This method is used frequently to determine the correlation between categorical variables and their target labels. It has the advantage of simplifying data complexity while preserving significant features [43,44]. Similarly, in situations of datasets with a significant number of dimensions, ANOVA (Analysis of Variance) and Mutual Information methods are commonly used for selecting features. These techniques evaluate the significance and duplication of features, which is important in determining the most influential variables for a classification problem [43]. Genetic Algorithms (GA) use operations such as selection, crossover, and mutation to progressively enhance the quality of solutions. This leads to improved efficiency in measures like ACC (accuracy), MCC (Matthews correlation coefficient), and F1-score during model assessments [45,46]. Furthermore, the ability of GA to maintain a dynamic equilibrium between exploration (seeking new areas in the solution space) and exploitation (improving current solutions) is what enables them to generate improved assessment values [47]. While genetic algorithms (GA) are an effective technique for feature selection and model improvement, they have some drawbacks, including runtime, which varies based on population size, number of generations, and the complexity of fitness evaluation. In practice, using GA may necessitate sufficient hardware resources (e.g., multi-core CPUs or GPUs) and careful parameter optimization to ensure scalability when applied to huge datasets. (No. 1)

Conclusions

In the context of AIDS drug design, the GAXGB model was developed specifically for the classification of anti-HIV peptides. The global HIV/AIDS pandemic has persisted for approximately 5 decades. This model employs supervised machine learning techniques and utilizes data from an online database. It extracts features from 12 different descriptors and applies data processing methods to convert them into predictive probability values. The selected features are then refined through a genetic algorithm, and the model is trained using the XGBoost algorithm. Performance evaluation showed that the model achieves an accuracy (ACC) greater than 90%, a robust Matthews correlation coefficient (MCC), and an F1-score of approximately 80%. The developers anticipate that this model will be valuable for researchers seeking to advance studies in this area.

Code and data accessibility

Codes for **GAXGB** and all data generated are available on Github at https://github.com/Monster-Jaru/GAXGB (The publication will occur only after the printing process is completed.)

Acknowledgements

This research was supported by Department of Mathematics and Computer Science, Faculty of Science and Technology, Prince of Songkla University, Pattani Campus, Thailand.

Declaration of generative AI in scientific writing

The authors admit that generative AI technologies, such as QuillBot and ChatGPT by OpenAI, were used to prepare this work, particularly for grammar correction and language editing. AI was not used for data interpretation or content creation. The conclusions and content of this work are entirely the authors' responsibility.

CRediT author statement

Jaru Nikom: Writing - original draft, Visualization, Validation, Software, Methodology, Investigation, Formal analysis and Conceptualization. **Watshara Shoombuatong:** Supervision, Methodology and Conceptualization. **Phasit Charoenkwan:** Supervision, Software, Methodology, Investigation, Formal analysis, Conceptualization. **Salang Musikasuwan**: Writing - review & editing, Supervision, Software, Methodology, Investigation, Formal analysis and Conceptualization.

References

[1] C Chen, W He, TY Nassirou, A Nsabiyumva, X Dong, YMN Adedze and D Jin. Molecular characterization and genetic diversity of different genotypes of Oryza sativa and Oryza glaberrima. *Electronic Journal of Biotechnology* 2017; **30**, 48-57.

[2] D Zhou, SM Dai and Q Tong. COVID-19: A recommendation to examine the effect of hydroxychloroquine in preventing infection and progression. *Journal of Antimicrobial Chemotherapy* 2020; **75(7)**, 1667-1670.

[3] MC Aguilera-Puga, NL Cancelarich, MM Marani, C de la Fuente-Nunez and F Plisson. Accelerating the discovery and design of antimicrobial peptides with artificial intelligence. *Methods in Molecular Biology* 2024; **2714**, 329-352.

[4] S Sachdeva. Peptides as 'drugs': The journey so far. *International Journal of Peptide Research and Therapeutics* 2017; **23(1)**, 49-60.

[5] C Li, D Sutherland, SA Hammond, C Yang, F Taho, L Bergman, S Houston, RL Warren, T Wong and LM Hoang. AMPlify: Attentive deep learning model for discovery of novel antimicrobial peptides effective against WHO priority pathogens. B*MC genomics* 2022; **23(1)**, 77.

[6] N Schaduangrat, C Nantasenamat, V Prachayasittikul and W Shoombuatong. ACPred: A computational tool for the prediction and analysis of anticancer peptides. *Molecules* 2019; **24(10)**, 1973.

[7] W Shoombuatong, N Schaduangrat and C Nantasenamat. Unraveling the bioactivity of anticancer peptides as deduced from machine learning. *EXCLI journal 2018*; **17**, 734.

[8] P Charoenkwan, N Anuwongcharoen, C Nantasenamat, MM Hasan and W Shoombuatong. In silico approaches for the prediction and analysis of antiviral peptides: a review. *Current pharmaceutical design* 2021; **27(18)**, 2180-2188.

[9] AC Kaushik, A Mehmood, S Peng, YJ Zhang, X Dai and DQ Wei. A-CaMP: A tool for anti-cancer and antimicrobial peptide generation. *Journal of Biomolecular Structure and Dynamics* 2021; **39(1)**, 285-293.

[10] Y Pang, L Yao, JH Jhong, Z Wang and TY Lee. AVPIden: A new scheme for identification and functional prediction of antiviral peptides based on machine learning approaches. *Briefings in Bioinformatics* 2021; **22(6)**, bbab263.

[11] P Charoenkwan, W Chiangjong, VS Lee, C Nantasenamat, MM Hasan and W Shoombuatong. Improved prediction and characterization of anticancer activities of peptides using a novel flexible scoring card method. *Scientific Reports* 2021; **11(1)**, 3017.

[12] P Charoenkwan, N Schaduangrat, B Manavalan and W Shoombuatong. M3S-ALG: Improved and robust prediction of allergenicity of chemical compounds by using a novel multi-step stacking

strategy. *Future Generation Computer Systems* 2024; **162**, 107455.

[13] N Thakur, A Qureshi and M Kumar. AVPpred: Collection and prediction of highly effective antiviral peptides. *Nucleic Acids Research* 2012; **40(W1)**, W199-W204.

[14] JH Jhong, L Yao, Y Pang, Z Li, CR Chung, R Wang, S Li, W Li, M Luo and R Ma. dbAMP 2.0: Updated resource for antimicrobial peptides with an enhanced scanning method for genomic and proteomic data. *Nucleic Acids Research* 2022; **50(D1)**, D460-D470.

[15] A Qureshi, N Thakur, H Tandon and M Kumar. AVPdb: A database of experimentally validated antiviral peptides targeting medically important viruses. *Nucleic acids research* 2014; **42(D1)**, D1147-D1153.

[16] S Singh, K Chaudhary, SK Dhanda, S Bhalla, SS Usmani, A Gautam, A Tuknait, P Agrawal, D Mathur and GP Raghava. SATPdb: A database of structurally annotated therapeutic peptides. *Nucleic Acids Research* 2016; **44(D1)**, D1119-D1126.

[17] G Shi, X Kang, F Dong, Y Liu, N Zhu, Y Hu, H Xu, X Lao and H Zheng. DRAMP 3.0: An enhanced comprehensive data repository of antimicrobial peptides. *Nucleic Acids Research* 2022; **50(D1)**, D488-D496.

[18] M Pirtskhalava, AA Amstrong, M Grigolava, M Chubinidze, E Alimbarashvili, B Vishnepolsky, A Gabrielian, A Rosenthal, DE Hurt and M Tartakovsky. DBAASP v3: Database of antimicrobial/cytotoxic activity and structure of peptides as a resource for development of new therapeutics. *Nucleic Acids Research* 2021; **49(D1)**, D288-D297.

[19] A Qureshi, N Thakur and M Kumar. HIPdb: A database of experimentally validated HIV inhibiting peptides. *PloS One* 2013; **8(1)**, e54908.

[20] N Poorinmohammad and H Mohabatkar. 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* 2015; **21(1)**, 57-62.

[21] N Poorinmohammad, H Mohabatkar, M Behbahani and D Biria. 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* 2015; **21(1)**, 10-16.

[22] KC Chou. Pseudo amino acid composition and its applications in bioinformatics, proteomics and system biology. *Current Proteomics* 2009; **6(4)**, 262-274.

[23] M Esmaeili, H Mohabatkar and S Mohsenzadeh. Using the concept of Chou's pseudo amino acid composition for risk type prediction of human papillomaviruses. *Journal of Theoretical Biology* 2010; **263(2)**, 203-209.

[24] JA Suykens and J Vandewalle. Training multilayer perceptron classifiers based on a modified support vector method. *IEEE Transactions on Neural Networks* 1999; **10(4)**, 907-911.

[25] S Suthaharan. *Support vector machine*. *In*: S Suthaharan (Ed.). Machine learning models and algorithms for big data classification: Thinking with examples for effective learning. Springer, New York, 2016, p. 207-235.

[26] B Mathiyazhagan, J Liyaskar, AT Azar, HH Inbarani, Y Javed, NA Kamal and KM Fouad. Rough set based classification and feature selection using improved harmony search for peptide analysis and prediction of anti-HIV-1 activities. *Applied Sciences* 2022; **12(4)**, 2020.

[27] G Wang, X Li and Z Wang. APD3: The antimicrobial peptide database as a tool for research and education. *Nucleic Acids Research* 2016; **44(D1)**, D1087-D1093.

[28] Y Liu, Y Zhu, X Sun, T Ma, X Lao and H Zheng. DRAVP: A Comprehensive database of antiviral peptides and proteins. *Viruses* 2023; **15(4)**, 820.

[29] Y Huang, B Niu, Y Gao, L Fu and W Li. CD-HIT Suite: A web server for clustering and comparing biological sequences. *Bioinformatics* 2010; **26(5)**, 680-682.

[30] P Charoenkwan, N Schaduangrat and W Shoombuatong. StackTTCA: A stacking ensemble learning-based framework for accurate and high-throughput identification of tumor T cell antigens. *BMC Bioinformatics* 2023; **24(1)**, 301.

[31] Z Chen, P Zhao, F Li, A Leier, TT Marquez-Lago, Y Wang, GI Webb, AI Smith, RJ Daly and KC

Chou. iFeature: A python package and web server for features extraction and selection from protein and peptide sequences. *Bioinformatics* 2018; **34(14)**, 2499-2502.

[32] F Pedregosa, G Varoquaux, A Gramfort, V Michel, B Thirion, O Grisel, M Blondel, P Prettenhofer, R Weiss and V Dubourg. Scikit-learn: Machine learning in Python. *the Journal of Machine Learning Research* 2011; **12**, 2825-2830.

[33] J Kim and S Yoo. Software review: DEAP (distributed evolutionary algorithm in python) library. *Genetic Programming and Evolvable Machines* 2019; **20(1)**, 139-142.

[34] D Chicco and G Jurman. The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. *BMC Genomics* 2020; **21(1)**, 6.

[35] P Fourie. Cd-hit: A fast program for clustering and comparing large sets of protein or nucleotide sequences. *Plant Disease* 2020; **22**, 1658.

[36] TE Creighton. *Proteins: Structures and molecular properties*. 2nd ed. W. H. Freeman and Company, New York, 1993.

[37] M Fisher. Lehninger principles of biochemistry, 3rd edition; By David L. Nelson and Michael M. Cox. *The Chemical Educator* 2001; **6**, 69-70.

[38] P Charoenkwan, W Chiangjong, C Nantasenamat, MM Hasan, B Manavalan and W Shoombuatong. StackIL6: A stacking ensemble model for improving the prediction of IL-6 inducing peptides. *Briefings in Bioinformatics* 2021; **22(6)**, bbab172.

[39] V Laengsri, C Nantasenamat, N Schaduangrat, P Nuchnoi, V Prachayasittikul and W Shoombuatong. TargetAntiAngio: A sequence-based tool for the prediction and analysis of anti-angiogenic peptides. *International Journal of Molecular Sciences* 2019; **20(12)**, 2950.

[40] R Barrett, S Jiang and AD White. Classifying antimicrobial and multifunctional peptides with Bayesian network models. *Peptide Science* 2018; **110(4)**, e24079.

[41] M Lu and T Gibson. Development of predictive tools for anti-cancer peptide candidates using generative machine learning models. *The Journal of Young Investigators* 2021; **39(5)**, 60-64.

[42] J Lin, L Wen, Y Zhou, S Wang, H Ye, J Su, J Li, J Shu, J Huang and P Zhou. PepQSAR: A comprehensive data source and information platform for peptide quantitative structure - activity relationships. *Amino Acids* 2023; **55(2)**, 235-242.

[43] Y Wang and C Zhou. Feature selection method based on chi-square test and minimum redundancy. *In*: Proceedings of the 5th International Conference on Intelligent and Interactive Systems and Application, Shanghai, China. 2020, p. 171-178.

[44] E Asad, A Islam, A Alam and AF Mollah. Univariate feature fitness measures for classification problems: An empirical assessment. *In*: Proceedings of the 5th International Conference, AMLDA, Tamaulipas, Mexico. 2022, p. 13-26.

[45] B Alhijawi and A Awajan. Genetic algorithms: Theory, genetic operators, solutions, and applications. *Evolutionary Intelligence* 2024; **17(3)**, 1245-1256.

[46] T Alam, S Qamar, A Dixit and M Benaida. Genetic algorithm: Reviews, implementations, and applications. *International Journal of Engineering Pedagogy* 2020; **10(6)**, 57-77.

[47] D Yang, Z Yu, H Yuan and Y Cui. An improved genetic algorithm and its application in neural network adversarial attack. *Plos One* 2022; **17(5)**, e0267970.