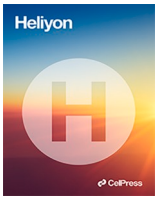
Heliyon 10 (2024) e36163

Contents lists available at ScienceDirect

Heliyon

journal homepage: www.cell.com/heliyon



Research article

Antiprotozoal peptide prediction using machine learning with effective feature selection techniques

Neha Periwal [a,1], Pooja Arora [b,1], Ananya Thakur [a], Lakshay Agrawal [c], Yash Goyal [d], Anand S. Rathore [b], Harsimrat Singh Anand [e], Baljeet Kaur [d,**], Vikas Sood [a,*]

[a] Department of Biochemistry, Jamia Hamdard, India
[b] Department of Zoology, Hansraj College, University of Delhi, India
[c] Freestand Sampling Pvt. Ltd, Delhi, India
[d] Department of Computer Science, Hansraj College, University of Delhi, India
[e] Faculty of Engineering and Mathematics, University of Waterloo, Canada

ARTICLE INFO

Keywords:
Peptide prediction
Machine learning
Antimicrobial peptides
Antiprotozoal peptides
Antiviral peptides
Non-AMP peptides
Feature selection

ABSTRACT

Background: Protozoal pathogens pose a considerable threat, leading to notable mortality rates and the ongoing challenge of developing resistance to drugs. This situation underscores the urgent need for alternative therapeutic approaches. Antimicrobial peptides stand out as promising candidates for drug development. However, there is a lack of published research focusing on predicting antimicrobial peptides specifically targeting protozoal pathogens. In this study, we introduce a successful machine learning-based framework designed to predict potential anti-protozoal peptides effective against protozoal pathogens.

Objective: The primary objective of this study is to classify and predict antiprotozoal peptides using diverse negative datasets.

Methods: A comprehensive literature review was conducted to gather experimentally validated antiprotozoal peptides, forming the positive dataset for our study. To construct a robust machine learning classifier, multiple negative datasets were incorporated, including (i) non-antimicrobial, (ii) antiviral, (iii) antibacterial, (iv) antifungal, and (v) antimicrobial peptides excluding those targeting protozoal pathogens. Various compositional features of the peptides were extracted using the pfeature algorithm. Two feature selection methods, SVC-L1 and mRMR, were employed to identify highly relevant features crucial for distinguishing between the positive and negative datasets. Additionally, five popular classifiers i.e. Decision Tree, Random Forest, Support Vector Machine, Logistic Regression, and XGBoost were used to build efficient decision models.

Results: XGBoost was the most effective in classifying antiprotozoal peptides from each negative dataset based on the features selected by the mRMR feature selection method. The proposed machine learning framework efficiently differentiate the antiprotozoal peptides from (i) non-antimicrobial (ii) antiviral (iii) antibacterial (iv) antifungal and (v) antimicrobial with accuracy of 97.27 %, 93.64 %, 86.36 %, 90.91 %, and 89.09 % respectively on the validation dataset.

* Corresponding author. Biochemistry Department School of Chemical and Life Science Jamia Hamdard, India.
** Corresponding author. Department of Computer Science Hansraj College University of Delhi, India.
E-mail addresses: baljeetkaur@hrc.du.ac.in (B. Kaur), v.sood@jamiahamdard.ac.in, vikas1101@gmail.com (V. Sood).
[1] NP and PA contributed equally to this study.

https://doi.org/10.1016/j.heliyon.2024.e36163
Received 14 June 2023; Received in revised form 9 August 2024; Accepted 11 August 2024
Available online 13 August 2024

Conclusion: The models are incorporated in a user-friendly web server (www.soodlab.com/appred) to predict the antiprotozoal activity of given peptides.

1. Introduction

Protozoal infections pose a profound threat to both human and animal health and have serious repercussions on the global economies (https://www.cdc.gov/parasites/about.html). These pathogens belong to the category of neglected tropical diseases and can lead to considerable mortality rates in endemic regions. For example, a notable proportion of deaths are attributed to *Plasmodium*, a parasitic protozoan responsible for Malaria (https://www.who.int/news-room/fact-sheets/detail/malaria). Another protozoan disease named Leishmaniasis is caused by *Leishmania donovani* and is responsible for an estimated 0.9 to 1.6 million cases each year (Leishmaniasis-PAHO/WHO|Pan American Health Organization). African Trypanosomiasis and Chagas disease caused by Trypanosoma species are other protozoan diseases that are a major concern in the endemic areas. In addition to the parasitic ones, free-living protozoa are also capable of causing diseases in humans. For instance, Acanthamoeba is a free-living unicellular pathogenic protozoan and is known to cause severe diseases of the eyes and central nervous system [1]. Another free-living flagellate pathogenic protozoa, *Naegleria fowleri* causes primary amoebic meningoencephalitis (PAM) that is prevalent both in developed and developing countries [2]. This water-borne amoeba has a mortality rate of 97 % since its emergence [3]. Protozoal pathogens are transmitted either directly or indirectly through contaminated food and water. Some pathogens are transmitted through vectors that carry the infection from an infected human to healthy ones [4]. Some of the factors like unhygienic living conditions, climate, and malnutrition lead to the frequent incidents of protozoal infections in various parts of the world [5,6]. Since protozoal pathogens are unicellular eukaryotes, hence they share common features with the mammalian systems. Therefore, some of the drugs targeting protozoa have been shown to cause severe toxicity in humans [7]. Compounding this issue, protozoal pathogens continue to develop novel means to evade host immunity and antiprotozoal drugs [8–12] leading to the emergence of drug resistance. Therefore, the identification of new drugs and novel druggable pathways is required to curb the rising menace of protozoal diseases.

Antimicrobial peptides are produced by various organisms and are one of the essential tools of the immune system. They are considered to be the first line of defence against several microbes. Naturally occurring antimicrobial peptides are active against a broad class of microorganisms including bacteria, virus and fungi [13,14]. They are short molecules (<100 amino acids long) [15] and are soluble in the aqueous environment [16] thereby making them an attractive drug candidate. Gramicidin is one of the antimicrobial peptides that has been successfully used in clinical settings as an alternative to antibiotics [17]. Another peptide commonly known as nisin has been shown to inhibit bacterial pathogens [18]. In addition to their antibacterial nature, several studies have successfully identified and characterized highly potent anti-protozoal peptides. These peptides act by disrupting protozoan cell membranes [19], cellular metabolism [20], and inducing cell death pathways [21–23]. Since these peptides are effective against protozoal pathogens, they can be used as one of the alternative approaches to curb these pathogens. However, the successful development of therapeutic peptides requires high throughput experimentations that are both time and resource-intensive. Recent advances in computational drug discovery have significantly expedited the drug discovery process [24], including the identification of potential bioactive peptides [25]. Therefore, we sought to build machine-learning based prediction models to classify antiprotozoal peptides from a diverse class of negative peptides including from highly diverse (non-antimicrobial peptides) to highly homologous (antiviral, antibacterial, antifungal, and antimicrobial) peptides. However, a comprehensive literature survey uncovered limited reports on the characterization of antiprotozoal peptides. Consequently, we conducted an extensive literature mining and curating of existing antimicrobial databases to assemble a positive dataset comprising experimentally validated antiprotozoal peptides.

Once the positive and negative datasets were curated, feature extraction was performed using the pfeature algorithm [26]. We explored two popular feature selection approaches including the Support Vector Classification with L1 regularization (SVC-L1) and minimum Redundancy Maximum Relevance (mRMR) to identify highly relevant features. Additionally, we conducted experiments with five different classifiers including Decision Tree (DT), Random Forest (RF), Support Vector Machine (SVM), Logistic Regression (LR), and XGBoost (XGB) to classify antiprotozoal peptides from each negative dataset. To facilitate the scientific community in predicting the antiprotozoal activity of peptides, we have developed a user-friendly web server where users can simply provide a list of peptides. The algorithm will then predict the antiprotozoal activity of the input peptides thereby enabling users to choose the most promising peptides for further validations. The novelty of the present study includes: (i) To the best of our knowledge, this is the first report that aims to use machine learning approaches for the classification of antiprotozoal peptides from non-antiprotozoal ones (ii) Inclusion of multiple negative datasets to make robust machine learning models and (iii) Inclusion of two feature selection tools to identify highly relevant and non-redundant features.

2. Related work

The advent of harnessing machine learning-based approaches for analysis of the copious amount of biological data has paved a new path for identifying and developing biopeptides. The prerequisite of drug discovery and development is the identification and validation of target molecules with required bioactivity [27]. Therefore the drug development pipeline becomes a very long process and the journey from lab to bedside can take decades. However, the process can be accelerated by using machine learning-based approaches [28]. Apart from the drug molecules, machine learning-based approaches have also been used successfully to predict and

Table 1
Gap areas in the antiprotozoal prediction research.

Authors	Year	Target organism	Bioactive Compound Database Against Target Organism	Aim/Drawbacks/Gap Area
Gulsen et al. [47]	2022	*Protozoa*	Secondary Metabolites	• Focus on screening antiparasitic secondary metabolites secreted from bacteria. •Screening done from the supernatants of 22 bacterial species only.
Mswahili et al. [48]	2021	*Plasmodium falciparum*	ChEMBL and PubChem	•Focus only on predicting antimalarial bioactivities, against Plasmodium falciparum, not considering other species of Plasmodium and potential protozoal targets. •Considering only two databases for antimalarial drugs
Liu et al. [49]	2020	*Plasmodium falciparum*	ChEMBL	•The aim was to develop classification models to predict the antimalarial activity of compounds against Plasmodium falciparum, not considering other species of Plasmodium and other protozoans. •Used a single database to extract compounds having antimalarial activity.
Moranga et al. [45]	2020	*Plasmodium species*	NA	•The aim was to develop a machine learning models for malaria detection using hematological parameters, not considering other protozoal disease. •Primarily focus on Ghanaian children, affect the applicability of developed models in diverse settings
Danishuddin et al. [50]	2019	*Plasmodium falciparum*	ChEMBL	•The aim was to build classification models for predicting antimalarial activity of compounds against only Plasmodium falciparum. •Used only a single database to collect experimentally verified compounds.
Egieyeh et al. [43]	2018	*Plasmodium falciparum*	ChEMBL, PubChem, manual curation from literature	•The paper aims to predict anti-plasmodium bioactivity of new natural compounds using machine learning classification model. •Included one protozoal parasite only.
Mason et al. [44]	2018	*Plasmodium falciparum*	NA	•The authors build machine learning models to identify novel combination of antimalarial drug that act synergistically against Plasmodium falciparum. •Considered only one species of Plasmodium.

design bioactive peptides which can induce Interferon-gamma (IFN gamma) [29], Interleukin-4 [30], Interleukin-10 [31], Interleukin-17 [32] and Interleukin-13 [33].

Additionally, use of machine-learning approaches for the prediction of antimicrobial peptides have been extensively investigated [34–36]. Several groups have developed machine learning models to predict the antiviral [37,38], antibacterial [39,40] and antifungal peptides [41,42] that constitute the subset of the antimicrobial peptides. Indeed various researchers have used machine learning-based approaches to successfully classify natural products having antimalarial activity [43], predict possible synergism among the anti-malarial drugs [44], classify malaria from other diseases [45], and improve malaria diagnosis [46]. However, to the best of our knowledge, no reports were found that aimed to use machine learning-based approaches to classify and predict the antiprotozoal peptides. This gap in the antiprotozoal peptide field prediction is summarized in Table 1 and prompted us to undertake the present study.

3. Materials and methods

We leveraged machine learning to classify antiprotozoal peptides from a diverse set of peptides including (i) non-antimicrobial (ii) antiviral (iii) antibacterial, (iv) antifungal, and (v) antimicrobial (excluding antiprotozoal) peptides. The major steps in this study include dataset preparation and pre-processing, feature extraction, feature selection, internal and external validation, building models, and design of a web-based prediction tool.

3.1. Dataset preparation and pre-processing

3.1.1. Original dataset

3.1.1.1. Positive dataset. A positive dataset consisting of experimentally validated antiprotozoal peptides targeting free-living as well as parasitic protozoa was manually curated from the research articles and existing antimicrobial databases including APD3 [51], DRAMP3 [52], ParaPep [53] and DBAASP [54]. Only those peptides that had linear conformation and contain natural amino acids were included in the positive dataset. We removed all the identical peptides and other peptides having length less than eight amino acids or greater than one hundred amino acids. We were thus successful in creating a positive dataset consisting of 275 experimentally validated antiprotozoal peptides for this study.

3.1.1.2. Negative datasets. The inclusion of multiple negative datasets has been reported to provide robustness to the study [38]. The negative datasets used in this study included (i) non-antimicrobial peptides (non-AMP), (ii) antiviral peptides, (iii) antibacterial peptides, (iv) antifungal peptides, and (v) antimicrobial peptides excluding antiprotozoal peptides. All these negative datasets were obtained from recent studies [55,56] and were found to be highly unbalanced as compared to the positive dataset. The non-AMP dataset comprised 6773 peptides, while there were 2001 peptides in the antiviral dataset, 3981 peptides in the antibacterial

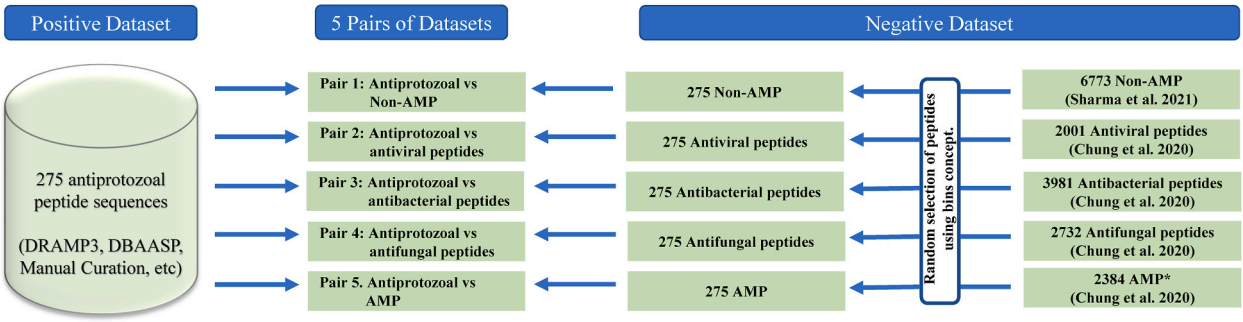


Fig. 1. Outlines the creation of five pairs, each with balanced positive and negative dataset. The positive dataset contains 275 experimentally validated antiprotozoal peptides. Various negative datasets consisting of (i) non-antimicrobial peptides (Non-AMP) (ii) antiviral peptides (iii) antibacterial peptides (iv) antifungal peptides and (v) antimicrobial peptides (excluding antiprotozoal peptides), were included. Balanced datasets in each pair were generated by randomly selecting 275 peptides from negative dataset using the bin strategy. AMP* indicates antimicrobial peptides excluding antiprotozoal peptides.

dataset, 2732 peptides in the antifungal dataset, and 2384 peptides in the antimicrobial dataset. In negative datasets, we retained only those peptides that had a length ranging from 8 to 100 amino acids.

3.1.2. Binning approach

The presence of a limited number of experimentally validated antiprotozoal peptides raised concerns about potential class imbalance. To construct robust models that are not influenced by the majority class, we performed under-sampling by employing binning approach [57], leading to the balancing of the positive and negative datasets. This approach entails establishing several empty bins according to peptide length. Peptides within the range of 8–25 amino acids were placed in the first bin, while those spanning from 26 to 50 amino acids were sorted in the second bin. The third bin included peptides with lengths between 51 and 75 amino acids, and the final bin was designated for peptides ranging from 76 to 100 amino acids. The negative datasets exhibited a higher abundance of sequences, resulting in a significantly greater number of sequences in each bin compared to the corresponding bins of the positive dataset. To handle this imbalance problem, we randomly select the peptides from each bin of the negative dataset, ensuring that the number of peptides is equivalent to the corresponding bin in the positive dataset. This process was repeated for all the above pre-defined bins. The peptides from each bin of the negative dataset were then combined to give a final negative dataset. This approach was repeated for each negative dataset to ensure an equal number of peptides compared to the positive class. The process of addressing class imbalance using the binning approach is described in Fig. 1.

3.2. Feature extraction

The Pfeature algorithm, extensively employed for extracting peptide features, served as the basis for our experiments [33,58,59]. Utilizing fifteen descriptors from the composition-based module of the Pfeature algorithm, we computed features for all the peptides. Each descriptor yields a distinct set of features for a peptide sequence. Consequently, a total of 9151 features were extracted for each peptide sequence. The fifteen descriptors of the composition-based feature module are described as follows.

3.2.1. Amino acid composition (AAC)

This descriptor computes the frequency of each amino acid in a peptide/protein sequence. Since there are 20 naturally occurring amino acids, thus this module yields 20 features for a sequence. Amino acid composition for each residue can be calculated using the following equation

$$AAC_i = \frac{N_i}{L}, i \epsilon \{A, C, D, \ldots . Y\}$$

where N_i is the count of amino acid i in the given peptide sequence and L represents the length of sequence.

3.2.2. Dipeptide composition (DPC)

This descriptor considers the coupling of adjacent amino acids and their positional information. DPC yields 400 features for a peptide sequence and can be computed usisng the following equation:

$$DPC_{i,j} = \frac{D_{i,j}}{L-1}, i, j \epsilon \{A, C, D, \ldots . Y\}$$

where $D_{i,j}$ represents the number of dipeptide consisting amino acid of type *i and j* in a peptide sequence and L is length of a sequence.

3.2.3. Tripeptide composition (TPC)

With 20 types of natural amino acid residues, there are 8000 (20*20*20) possible tripeptide combinations. TPC is 8000-dimensional feature vector and the frequency of each tripeptide in a peptide sequence can be computed using the following equation

$$TPC_i = \frac{T_i}{L-2}$$

where TPC_i represents the tripeptide composition of tripeptide i while T_i and L denote the number of tripeptides of type i and the length of the protein sequence, respectively.

3.2.4. Atom & bond composition (ATC & BTC)

This module computes different type of atom and bond composition in a peptide sequence. Atomic composition refers to the fraction of Carbon, Hydrogen, Nitrogen, Oxygen and Sulphur atoms present in a peptide sequence. For bond composition, a total number of bonds (including aromatic bonds), hydrogen bonds, single bonds, and double bonds are considered. These are nine dimensional feature vectors and can be described using the following equation:

$$ATC_i = \frac{A_i}{N}$$

$$BTC_i = \frac{B_i}{N}$$

where ATC_i represents the atomic composition of atom type i, with A_i denoting the number of atoms of type i and N representing the total number of atoms in a peptide sequence. Similarly, BTC_i represents the bond composition for bond type i, where B_i is the number of bonds of type i, and N is the total number of atoms in a peptide sequence.

3.2.5. Distance distribution of residue (DDOR)

This descriptor calculates the distribution of residue based on their distance from the N-terminal, C-terminal, and the inter-distances between identical residues in the given peptide sequence. DDOR is 20 dimensional feature vector and computed using the following equation

$$DDOR_i = \frac{(R_{NT})^2 + \sum_{j=1}^{N}(R_j)^2 + (R_{CT})^2}{(L - F_{i)} + 1}$$

where $DDOR_i$ and Fi represent the distance distribution and frequency of residue type i, R_{NT} and R_{CT} are the residue distance from the N-terminal and C-terminal respectively, N is the total number of inter-residue distances for type I, R is the inter-distance between residue type i, and L is the total length of the peptide sequence.

3.2.6. Residue repeat information (RRI)

RRI counts the number of consecutive runs of each amino acid type in a peptide sequence. It is a 20 dimensional feature vector which can be computed using the following formula.

$$RRI_i = \frac{\sum_{j=1}^{N}(R_j)^2}{\sum_{j=1}^{N} R_j}$$

where RRI_i and N represent the residue repeat information, and a maximum number of occurrences of residue i, respectively and R_j indicates the number of repeats in occurrence j for residue type i.

3.2.7. Shannon Entropy at peptide/protein Level (SE)

This descriptor computes the Shannon entropy of a peptide sequence by using the following expression

$$H(X) = -\sum_{i=1}^{20} pi\ log_2 p_i$$

where X refers to any peptide sequence while i corresponds to an amino acid in the sequence. The SE yield 1 feature for a peptide sequence.

3.2.8. Shannon Entropy at residue Level (SER)

This descriptor compute the Shannon entropy of 20 natural amino acid residues in a sequence. It computes 20 features for a given peptide sequence by using the following equations

$$p_i = \frac{c_i}{L}$$

$$H_i = pi\ log_2 p_i$$

Where C_i and H_i represents the count and entropy of residue i in the sequence, L denotes the total length of the sequence.

3.2.9. Shannon Entropy of physiochemical property (SEP)

This module calculates the Shannon entropy of a specific physiochemical property in a peptide and contributes 25 features for a peptide sequence. It can be computed by using the following formula

$$H_i = -p_i \log(p_i) - (1 - p_i)\log(1 - p_i)$$

where p_i is r_i/l.

H_i denotes the Shannon Entropy of a particular physiochemical property, l is the length of the sequence and has r_i instances of a property present in the sequence.

3.2.10. Conjoint Triad descriptors (CTD)

This scheme was initially presented by Ref. [60] where all twenty amino acids were divided into seven groups on the basis of their dipoles and volumes of the side chains: group 1 (A,G,V), group 2 (I,L,F,P), group 3 (Y,M, T,S), group 4 (H,N,Q,W), group 5 (R,K), group 6 (D,E), group 7 (C). The peptide sequence is analysed by calculating the frequency of three consecutive amino acids, resulting in output vectors with a dimension of 343. This can be illustrated using an example where a peptide sequence is denoted by binary vector (S_i, F_i) and S_i represents extracted feature space whereas F_i corresponds to the frequency vector. The values of F_i are related to the peptide length. To resolve this issue, a parameter 'd_i' is introduced to normalize f_i for each peptide sample, and it can be expressed as:

$$d_i = \frac{f_i - \min\left\{f_0 f_2, \ldots., f_{342}\right\}}{\max\left\{f_0 f_2, \ldots., f_{342}\right\}}$$

3.2.11. Composition-enhanced transition distribution (CeTD)

Composition, enhanced Transition, Distribution (CeTD), describes the pattern of amino acid distribution along the peptide sequence, on the basis of their physicochemical or structural properties. This analysis encompasses seven physiochemical properties: secondary structure, polarity, hydrophobicity, normalized van der Waals volume, polarizability, charge, and solvent accessibility. The Composition feature in CeTD, describes the percentage of a particular physiochemical property for each residue, the transition feature calculates the frequency of amino acids with one property followed by amino acids in another property, and the distribution feature characterizes the percentage of the peptide sequence containing fractions of amino acids with a specific property at varying chain lengths. The CeTD generates 189 vector for a peptide sequence.

3.2.12. Pseudo amino acid composition (PAAC)

This descriptor considers the sequence order correlation of any two residues in the peptide or protein sequence. In PAAC, λ represents the highest tier of correlation in the sequence, and in this study, λ was set to 3 reflecting the correlations up to three residues. Thus PAAC yields 23 dimensional feature vector for a peptide sequence.

3.2.13. Amphiphilic pseudo amino acid composition (APAAC)

APAAC is the modified version of PAAC and generates 29 features for a peptide sequence. The set of correlation factors (2 λ) in APAAC reflects distinct hydrophilicity and hydrophobicity distribution pattern along a peptide sequence.

$$P_C = \frac{f_c}{\sum_{r=1}^{20} f_r + w\sum_{j=1}^{2\lambda} \tau_j} \quad (1 < c < 20)$$

$$P_C = \frac{w\tau_u}{\sum_{r=1}^{20} f_r + w\sum_{j=1}^{2\lambda} \tau_j} \quad (21 < u < 20 + 2\lambda)$$

where w is the weighting factor.

3.2.14. Quasi-sequence order (QSO)

QSO captures the order based sequence information of a peptide sequence and generates 46 vectors. It computes the distance matrix between the 20 amino acids by using the Schneider-Wrede physiochemical distance matrix and Grantham chemical distance matrix. The following equation is used to calculate the quasi-sequence-order for each residue

$$X_r = \frac{f_r}{\sum_{r=1}^{20} f_r + w\sum_{d=1}^{nlag} \tau_d} \quad r = 1, 2, \ldots\ldots 20$$

$$X_d = \frac{w\tau_d - 20}{\sum_{r=1}^{20} f_r + w\sum_{d=1}^{nlag} \tau_d} \quad d = 21, 22, \ldots.30 + nlag$$

Where f_r represents the normalized occurrence of residue type r, w is a weighting factor and nlag is the maximum value of the lag.

3.2.15. Sequence order coupling number (SOCN)

This module yields 6-dimensional feature for a peptide sequence. It uses the Grantham and Schneider dissimilarity matrices to compute the d-th rank SOCN via the following equation.

$$\tau_d = \sum_{i=1}^{N-d} \left(d_{i,i+d}\right)^2 \; d = 1, 2, 3, \ldots\ldots, nlag$$

Where $d_{i,\, i+d}$ represents the distance between the two amino acids at positions (i) and (i + d), (nlag) is the maximum value of the lag, and (N) represent the length of the protein or peptide sequence.

3.3. Feature selection

Relevant features are a prerequisite for building a robust machine learning model. For any machine learning classifier, a small sample size with a high dimension of features results in building an overfitted model with poor generalization capabilities. The high dimensional attributes of the peptide sequences extracted from Pfeature algorithm may contain irrelevant and redundant features which may degrade the performance of the machine learning model. Therefore, it becomes necessary to retain only those features that are more discriminatory in nature. Feature selection thus becomes a crucial step in constructing an efficient machine learning model thereby improving the performance measures and reducing model complexity. Several studies have employed diverse feature selection methods to identify relevant features from different datasets like proteome data [61], transcriptome data [62], metabolome data [63–65], and metaproteome data [66]. In order to select the subset of non-redundant features that contribute toward an efficient machine learning model and classification of peptides with high confidence, two feature selection methods: SVC-L1 and mRMR were investigated in this study.

3.3.1. SVC-L1 feature selection

The L1-norm or L1 regularization adds a penalty equal to the sum of the absolute values of the coefficients to the loss function hence shrinking some parameters to zero. As a consequence, some variables do not play a role in the decision model, hence L1-norm helps to select a subset of features in a model. L1-norm in the formulation of the Support Vector Machine helps to select a subset of features. SVC-L1 is a linear model which is penalized by the L1 norm [67,68]. Let $\{y_i, \boldsymbol{x}_i\}_{i=1}^{n}$ be the data being considered, where $y_i \, \epsilon \, \{1, -1\}$ is the response variable, for the corresponding instance $\boldsymbol{x}_i = (x_0, x_1, \ldots, x_p)$, $x_0 = 1$ with respect to the intercept term. The SVM can be expressed as the following regularization problem:

$$\min \frac{1}{n} \sum_{i=1}^{n} \max\left(0, \left(1 - y_i \boldsymbol{x}_i{}^T \boldsymbol{\beta}\right)\right)$$

Where $\max\left(0, \left(1 - y_i \boldsymbol{x}_i{}^T \boldsymbol{\beta}\right)\right)$ is the hinge loss function, $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_p)$ and λ is the regularization parameter.

By using L1 regularization, some features are forced to be excluded, hence building the model with only the relevant features. This encourages sparse models and helps to eliminate features that are redundant. This method helps reduce the complexity of the machine learning model.

3.3.2. mRMR feature selection

For effective feature selection, it is imperative that the relevant features that are selected are most discriminative to distinguish the positive and the negative class. However, the presence of redundant features can deteriorate the performance of the decision model. Hence, it becomes necessary to select the relevant as well as non-redundant features that help improve the effectiveness of the decision model. The mRMR is a feature selection method that chooses those features that are maximally relevant to the target class (c) while being minimally redundant with the chosen subset of features(S) [69]. Given two features, i and j, with marginal probabilities, $p(i)$ and $p(j)$, and joint probabilities $p(i,j)$, the mutual information ($I(i, j)$) is given by:

$$I(i,j) = \Sigma p(i,j) log \frac{p(i,j)}{p(i)p(j)}$$

In order to select a feature subset that satisfies both the minimal redundancy and maximum relevance simultaneously, mRMR method is denoted as

$$\begin{gathered} Max\ \{\ Relevance - Redundancy\ \} \\ or \\ max\{\ Relevance/Redundancy\ \} \end{gathered}$$

where

$$Relevance = \frac{1}{|s|} \sum_{i \epsilon S} \mathrm{I}(\mathrm{i}, \mathrm{c})$$

$$Redundancy = \frac{1}{|s|^2}\sum_{i,j\ \epsilon S} I(i,j)$$

3.4. Machine learning classifiers

3.4.1. Decision tree

Decision tree-based classifier is a prediction method which resembles a tree structure and a set of rules. The training data is repeatedly partitioned using some splitting criterion till all records in a partition belong to a single class. The splitting criterion decreases the entropy of the dataset set with each split. The information gain, gain ratio and entropy are the common splitting criterion used in building decision trees. At every node, an appropriate feature with the most suitable split point is chosen that minimizes the cost function. The leaf nodes make the final predictions. A test sample is predicted by navigating the tree as per the split conditions and reaching a leaf node with the predicted target class.

3.4.2. Random forest

Random forest algorithm is an ensemble of decision trees where each tree is formed from a different training set and hence each has a different performance. Prediction in random forest method is based on the collective decision of the participating decision tress and hence shows improved performance as compared to when only a single decision tree is modelled. Random forest relies on the majority vote of predictions from each tree and predicts the final target class accordingly. Increased number of trees in the random forest prevents the problem of overfitting.

3.4.3. Support vector machine

The support vector machine classifier [70] determines a decision boundary that maximizes the margin between the hyperplanes passing through the support vectors of the two classes, where samples of a given class are on either side of the hyperplanes.

The optimization problem is represented as:

$$min \frac{1}{2}||w||^2$$

$$st. \quad y_i(\boldsymbol{w}^t x_i + b) \geq 1 \ for \ i = 1, 2 \ldots\ldots n$$

where, x_i is the ith ($i = 1 \cdots n$) input sample of m dimension, y_i is either 1 or -1, each indicating the positive or negative class to which the sample x_i belongs. $\boldsymbol{w}$ is the normal vector to the hyperplane separating the training samples of the two classes and b is the bias. Non-linear decision boundaries are determined by the SVM using the appropriate kernel function.

3.4.4. Logistic regression

Logistic regression predicts the probability of the target variable using the logistic function *f(z)* defined as:

$$f(z) = \frac{1}{1 + e^{-z}}$$

where

$$z = \beta_0 + \beta_1 x_{.,1} + \beta_2 x_{.,2} + \ldots + \beta_r x_{.,r}$$

$x_{.,i}$ are the independent variables and β_i are coefficients that are estimated using the maximum likelihood estimation.

$$y = \begin{cases} 1 \ f(z) > threshold \\ 0 \ f(z) < threshold \end{cases}$$

y is the predicted binary classification label.

3.4.5. XGBoost (Extreme gradient boost)

The XGBoost is an ensemble of decision tree models where each tree is included one at a time to the ensemble to improve the prediction errors made by prior models. It uses a gradient descent algorithm and improves upon the errors of previously built models. Overfitting is controlled in XGBoost with the help of regularization parameters to select features based on the weak and strong features in the decision tree. Both random forest and XGBoost generally decrease the variance, while XGBoost is instrumental in improving the bias.

3.5. Internal and external validation

To train, test, and evaluate our prediction models, we performed the 10-fold cross-validation and external validation technique. As per the standard protocol, the entire dataset from each pair was split in a ratio of 80:20 to obtain 440 peptides that formed the internal validation dataset whereas 110 peptides formed the external validation dataset. The selection of optimum hyperparameters is crucial

Fig. 2. The proposed framework for the design of APPRED: The balanced dataset was divided into 80:20 as training and testing data. The 15 descriptors from compositional module of Pfeature was used to compute the features of peptide sequence. Feature selection tool either mRMR or SVC-L1 was used to select non-redundant and discriminatory features from the pool of 9151 features. 10-fold cross validation was performed on training dataset using five machine learning classifier. External validation was carried out on testing and independent dataset. *Abbreviations:* mRMR, minimum redundancy maximum relevance; ML, Machine Learning; DT, Decision Tree; RF, Random Forest; SVM, Support Vector Machine; LR, Logistic Regression; XGB, eXtreme Gradient Boosting.

for building the robust and effective models [42,71–73]. We then performed 10-fold cross-validation on the training dataset using optimal hyperparameters. In this process, the entire training dataset was divided into ten equal parts where nine parts comprise the training data and one part forms the testing data. Each of the ten parts has an equal number of positive and negative sequences. This process is then iterated ten times so that each part of the data can be used as testing data. The robustness and generalizability of the trained model were evaluated using the external validation dataset.

3.6. Evaluation parameters

For a binary class problem, we label the class under consideration as the positive class (P) (anti-protozoal peptides in the given problem) and the other class as the negative class (N). The number of positive samples correctly predicted by the classifier is referred to as TP (true positives). The number of negative samples predicted correctly by the classifier is referred to as TN (true negatives). The term false positive (FP) refers to the number of negative samples predicted as positive by the classifier (which is misclassification). The false negative (FN) refers to the number of positive samples that the classifier predicted as negative (which too is misclassification). A successful machine learning model maximizes the TN and TP and minimizes the FP and FN. It is important to note that in an experiment, the total positive samples, $P = TP + FN$. The total negative samples can be represented as $N = TN + FP$. The complete set of samples is thus $T + N = TP + FN + TN + FP$.

In this study, well-established threshold-dependent and independent parameters were used to check the efficiency of our prediction models. The most commonly used threshold-dependent parameters used are accuracy, sensitivity and specificity. We used the standard threshold independent parameter i.e. Area under Receiver Operating Characteristics (AUCROC) to measure the performance of a machine learning model.

Accuracy is a measure of the total correct predictions by a decision model. It can be defined as the percentage of true positives and true negatives over the total number of observations.

$$\text{Accuracy} = (\text{correct predictions}) \,/\, (\text{total samples})$$

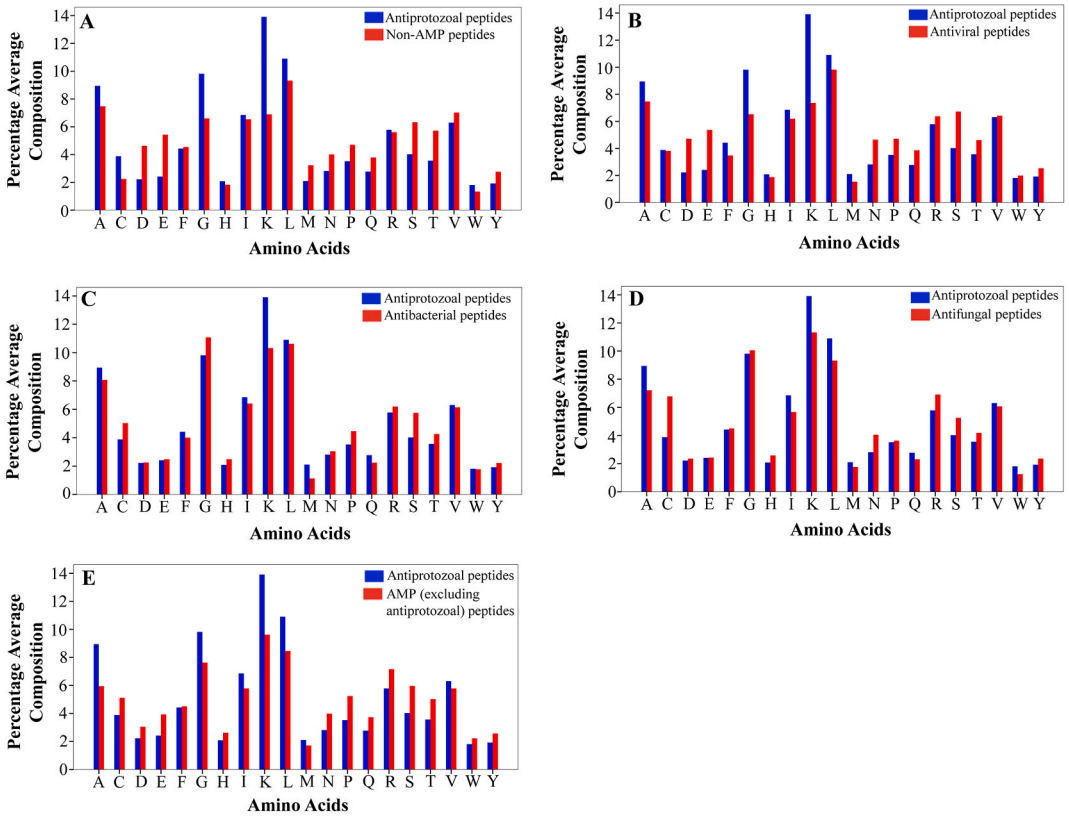


Fig. 3. Comparative analysis of the average amino acid composition between the positive and negative dataset of each pair: (A) antiprotozoal vs non-antimicrobial peptides, (B) antiprotozoal vs antiviral peptides, (C) antiprotozoal vs antibacterial peptides, (D) antiprotozoal vs antifungal peptides, and (E) antiprotozoal vs anti-microbial peptides excluding antiprotozoal peptides. The X-axis represents amino acids and the Y-axis represents the average composition of amino acids in percentage.

$$\text{Accuracy} = (\text{TP} + \text{TN}) \,/\, (\text{TP} + \text{FN} + \text{TN} + \text{FP})$$

Sensitivity is the metric that helps to measure the true positive rate (TPR). It identifies how well the model predicts the positives correctly. This is the ratio of true positive and the actual number of positives in the observations. If the number of true positives is high and the false negatives is low, it results in high sensitivity.

$$\text{Sensitivity} = \text{TP} \,/\, (\text{TP} + \text{FN})$$

Specificity is defined as the percentage of the negatives being correctly predicted by the algorithm. It is referred to as the true negative rate (TNR). When the negatives are predicted correctly and the false positives are low, specificity is high. Both sensitivity and specificity are separately indicative of the individual classes being predicted well.

$$\text{Specificity} = \text{TN} \,/\, (\text{TN} + \text{FP})$$

The *AUCROC score* is the area under the ROC curve drawn between the true positive rate (TPR) and false positive rate (FPR) across different decision thresholds. The x-axis of a ROC curve is the false positive rate, and the y-axis is the true positive rate. For each threshold, higher TPR and the lower FPR are desired, hence classifiers that have higher curves on the top left side are better. The value of AUCROC ranges from 0.5 to 1. The higher value of AUCROC suggests that the ML classifier is able to distinguish better among the positive and negative classes.

3.7. Design of web-based prediction tool

To assist the users in evaluating whether a given peptide might have antiprotozoal properties, we developed a user-friendly web server named “APPred”. The frontend of the web server was developed using the HTML, CSS, Bootstrap and Javascript whereas the flask framework was used for the backend. The users are prompted to submit a peptide sequence ranging from 8 to 100 amino acids in length. The input query is processed, and the output provides a probability score indicating the likelihood of the input sequence exhibiting antiprotozoal activity. The users are provided with the option to choose from a range of negative datasets, feature selection methods, and machine learning classifiers. The webserver offers two additional features: the design module and the protein scan

Table 2
The performance of machine learning models developed using SVC-L1 and mRMR selected features on training and validation dataset.

Classifier	Feature Selection	Sensitivity (mean ± SD)	Specificity (mean ± SD)	Accuracy (mean ± SD)	AUROC (mean ± SD)	Sensitivity	Specificity	Accuracy	AUCROC
		Training Dataset				Validation Dataset			
Decision Tree	SVC-L1	76.81 ± 9.41	76.81 ± 8.49	76.82 ± 6.80	0.83 ± 0.07	85.45	81.82	83.64	0.86
	mRMR	78.63 ± 6.11	80 ± 8.67	79.32 ± 6.13	0.85 ± 0.07	83.64	87.27	85.45	0.93
Random Forest	SVC-L1	82.27 ± 7.45	87.27 ± 8.33	84.77 ± 5.75	0.93 ± 0.04	90.91	90.91	90.91	0.97
	mRMR	85.0 ± 8.14	84.09 ± 5.47	84.54 ± 3.91	0.93 ± 0.02	94.55	90.91	92.73	0.98
SVM	SVC-L1	72.27 ± 10.84	80.45 ± 7.88	76.36 ± 7.55	0.83 ± 0.06	80.0	83.64	81.82	0.90
	mRMR	83.63 ± 6.49	79.09 ± 6.80	81.36 ± 4.7	0.89 ± 0.04	85.45	81.82	83.64	0.92
Logistic Regression	SVC-L1	87.27 ± 6.36	87.73 ± 10.57	87.95 ± 5.38	0.92 ± 0.03	89.09	90.91	90.0	0.95
	mRMR	82.72 ± 11.82	84.09 ± 5.08	83.41 ± 5.09	0.91 ± 0.05	89.09	89.09	89.09	0.96
XGBoost	SVC-L1	**88.18 ± 5.82**	87.73 ± 10.57	87.95 ± 6.35	0.94 ± 0.04	96.36	90.91	93.64	0.98
	mRMR	87.72 ± 4.56	**88.63 ± 9.58**	**88.18 ± 3.63**	**0.95 ± 0.02**	**98.18**	**96.36**	**97.27**	**0.99**

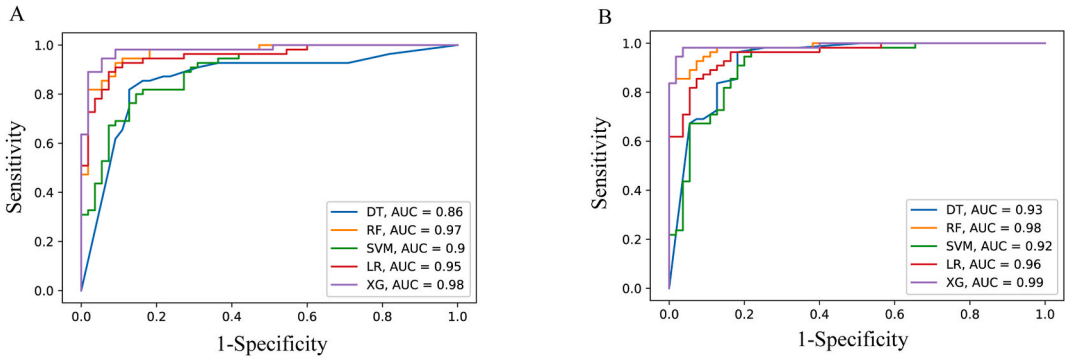


Fig. 4. AUCROC curve shows the performance of five models on validation dataset. These models were built to classify antiprotozoal from non-AMP peptides using features selected from (A) SVC-L1 and (B) mRMR. The X-axis represents the false positive rate i.e. 1-Specificity while Y-axis represents the true positive rate i.e. Sensitivity.

module. The design module facilitates the generation of single amino acid mutants by systematically mutating one residue of the peptide sequence at a time while keeping all other residues constant and then predicting the antiprotozoal activity of the resulting mutants. This procedure is repeated for each amino acid in a given peptide. The protein scan module generates all the overlapping peptides of the desired length from a given protein sequence and then predicts the antiprotozoal activity of all the generated peptides based on the probability score. Users are provided with the option to download and save the results for further analysis. The entire workflow of this study is depicted in Fig. 2.

4. Results

In this study, we created five negative datasets including non-antimicrobial peptides, antiviral peptides, antibacterial peptides, antifungal peptides and antimicrobial peptides (excluding antiprotozoal peptides). We then performed the classification of antiprotozoal peptides (positive dataset) from each negative dataset. For a better interpretation of the results, we created five pairs each consisting of the same positive dataset i.e. antiprotozoal peptides alongside a negative dataset selected from the aforementioned negative datasets. The five pairs include (i) antiprotozoal and non-antimicrobial peptides, (ii) antiprotozoal and antiviral peptides, (iii) antiprotozoal and antibacterial peptides, (iv) antiprotozoal and antifungal peptides and (v) antiprotozoal and antimicrobial (excluding antiprotozoal) peptides.

4.1. Compositional analysis of positive and negative datasets

We performed compositional analysis of the constituent amino acids for both the positive and negative datasets within each pair. Fig. 3 represents distinct amino acid patterns between the positive and negative datasets for each pair, delineating notable differences in the sequence composition. The results can be summarized as follows.

1. Comparative analysis of antiprotozoal with non-antimicrobial peptides (pair 1) revealed that several amino acids including Alanine (A), Cysteine (C), Glycine (G), Lysine (K), and Leucine (L) were enriched in antiprotozoal peptides whereas Aspartic Acid (D), Glutamic Acid (E), Methionine (M), Asparagine (N), Proline (P), Glutamine (Q), Serine (S), and Threonine (T) were over-represented in non-antimicrobial peptides as shown in Fig. 3A.
2. Fig. 3B illustrates the amino acid composition among antiprotozoal and antiviral peptides (pair 2). It can be observed that amino-acids Alanine (A), Phenylalanine (F), Glycine (G), Isoleucine (I), Lysine (K), and Leucine (L) were over-represented in antiprotozoal peptides whereas Aspartic acid (D), Glutamic acid (E), Asparagine (N), Proline (P), Glutamine (Q), Serine (S) and Threonine (T) were enriched in antiviral peptides.
3. In Fig. 3C, the frequency of amino acids Alanine (A), Lysine (K), and Methionine (M) was more prominent in antiprotozoal peptides whereas Cysteine (C), Glycine (G), Proline (P), Serine (S), and Threonine (T) formed the major constituents of the antibacterial peptides (pair 3).
4. A comparison of antiprotozoal and antifungal peptides (pair 4) revealed that certain amino acids such as Alanine (A), Isoleucine (I), Lysine (K), Leucine (L) were prominent constituents of antiprotozoal peptides whereas Cysteine (C), Asparagine (N), Arginine (R), and Serine (S) were enriched in antifungal peptides (Fig. 3D).
5. The amino acids Alanine (A), Glycine (G), Isoleucine (I), Lysine (K), and Leucine (L) were enriched in antiprotozoal peptides whereas Cysteine (C), Aspartic acid (D), Glutamic acid (E), Asparagine (N), Proline (P), Aspartic acid (Q), Arginine (R), Serine (S), and Threonine (T) were abundant in antimicrobial peptides (pair 5, Fig. 3E).

Table 3
The performance of machine learning models developed using SVC-L1 and mRMR selected features on training and validation dataset.

Classifier	Feature Selection	Sensitivity (mean ± SD)	Specificity (mean ± SD)	Accuracy (mean ± SD)	AUROC (mean ± SD)	Sensitivity	Specificity	Accuracy	AUCROC
		Training Dataset				Validation Dataset			
Decision Tree	SVC-L1	73.18 ± 6.57	69.09 ± 11.82	71.14 ± 6.10	0.75 ± 0.06	87.27	83.64	85.45	0.84
	mRMR	74.09 ± 8.63	75.91 ± 9.55	75.0 ± 4.98	0.82 ± 0.05	89.09	83.64	86.36	0.91
Random Forest	SVC-L1	80.91 ± 8.81	78.18 ± 8.81	79.54 ± 5.83	0.89 ± 0.05	2.73	85.45	89.09	0.96
	mRMR	76.36 ± 11.46	82.72 ± 7.27	79.55 ± 4.07	0.88 ± 0.03	92.73	87.27	90.0	0.96
SVM	SVC-L1	77.27 ± 12.36	66.36 ± 9.13	71.82 ± 8.75	0.77 ± 0.07	74.55	72.73	73.64	0.86
	mRMR	73.64 ± 9.71	74.09 ± 10.96	73.86 ± 6.68	0.84 ± 0.07	83.64	80.0	81.82	0.89
Logistic Regression	SVC-L1	80.91 ± 7.27	75.91 ± 10.17	78.41 ± 5.86	0.85 ± 0.06	89.09	85.45	87.27	0.92
	mRMR	73.64 ± 9.49	74.55 ± 8.18	74.55 ± 7.52	0.84 ± 0.04	89.09	81.82	85.45	0.92
XGBoost	SVC-L1	79.55 ± 11.36	**82.73 ± 9.19**	80.91 ± 6.03	0.87 ± 0.05	87.27	87.27	87.27	0.95
	mRMR	**83.18 ± 7.34**	81.36 ± 4.75	**82.72 ± 5.06**	**0.89 ± 0.03**	**94.55**	**92.73**	**93.64**	**0.98**

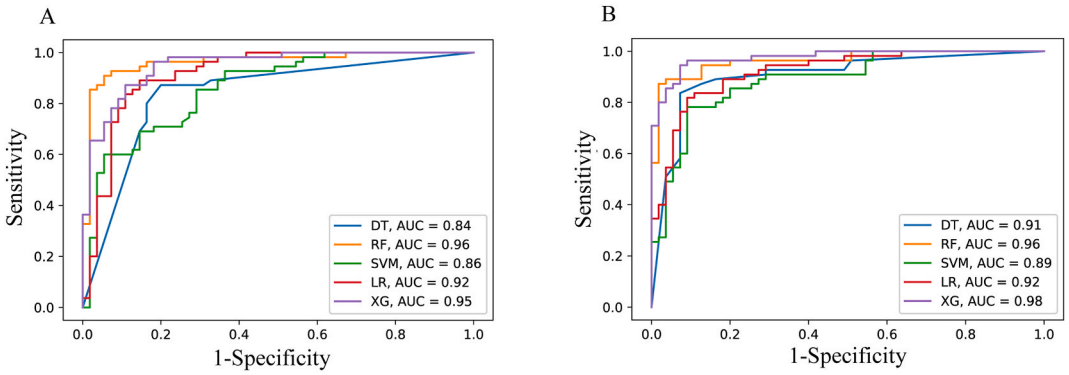


Fig. 5. AUCROC curve shows the performance of five models on validation dataset. These models were built to classify antiprotozoal from antiviral peptides using features selected from (A) SVC-L1 and (B) mRMR. The X-axis represents the false positive rate i.e. 1-Specificity while Y-axis represents the true positive rate i.e. Sensitivity.

4.2. Feature selection

The selection of relevant features becomes a prerequisite for building a robust machine-learning model. We explored two feature selection methods viz: SVC-L1 and mRMR to select highly relevant and non-redundant features. We evaluated the performance of the different feature sets extracted from SVC-L1 and mRMR respectively. For each method, we build machine learning based prediction models on the top (15, 16, 17,.30) features, respectively, and assess the performance on both training and validation datasets via sensitivity, specificity, accuracy, and AUCROC. We identified the feature set with the least number of features, that effectively discriminates the positive dataset from the negative dataset with high accuracy. This process was repeated for each pair, revealing that the optimum number of features in the feature set varies across pairs.

4.3. Machine learning-based prediction models

We developed machine learning-based prediction models with the optimum features using five different classifiers i.e. DT, RF, SVM, LR, and XGB. We performed 10-fold cross-validation and external validation on the training and validation dataset respectively. We evaluated the performance metrics on the training dataset via average sensitivity, specificity, accuracy, and AUCROC. For each experimental dataset pair, we build the five machine-learning models using the optimum features selected by the SVC-L1 and the mRMR methods separately. The hyperparameters that were used to build the robust model are described in Supplementary Table 1.

4.3.1. A. Experimental results of pair 1: antiprotozoal vs non-antimicrobial peptides

The first set of experiments was performed with antiprotozoal peptides as positive and non-antimicrobial peptides as negative dataset (pair 1). We build the models on the top 25 optimum features selected from SVC-L1 and mRMR.

The following is observed.

i. During internal validation, we attained the maximum average specificity (88.63 %), accuracy (88.18 %), and AUCROC (0.95) using the discriminatory features obtained from mRMR feature selection method in conjunction with the XGB classifier (Table 2). However, the maximum average sensitivity (88.18 %) was achieved from the 25 optimum features obtained using SVC-L1 feature selection method with XGB classifier.
ii. During external validation, we achieved 98.18 % sensitivity, 96.36 % specificity, 97.27 % accuracy, and a 0.99 AUCROC using features obtained through the mRMR feature selection method combined with the XGB classifier (Table 2).
iii. The AUCROC curve establishes the superior performance of 25 optimum features selected using mRMR (Fig. 4B) in comparison to SVC-L1 (Fig. 4A) leading to the effective classification of antiprotozoal peptides from non-antimicrobial ones.

4.3.2. B. Experimental results of pair 2: antiprotozoal vs antiviral peptides

The second set of experiments involved antiprotozoal peptides as the positive dataset and antiviral peptides as negative dataset (pair 2). The classification models were built on top 25 optimum features selected from SVC-L1 and mRMR. The following is observed.

i. For the training dataset, we attained the model's highest performance, with an average sensitivity of 83.18 %, accuracy of 82.72 %, and AUCROC of 0.89. This model was built by leveraging the optimum feature selected via the mRMR method in conjunction with the XGB classifier (Table 3).
ii. We achieved 94.55 % sensitivity, 92.73 % specificity, 93.64 % accuracy, and 0.98 AUCROC with the validation dataset. The model was built on features selected by mRMR in conjunction with XGB classifier (Table 3).

Table 4
The performance of machine learning models developed using SVC-L1 and mRMR selected features on training and validation dataset.

Classifier	Feature Selection	Sensitivity (mean ± SD)	Specificity (mean ± SD)	Accuracy (mean ± SD)	AUROC (mean ± SD)	Sensitivity	Specificity	Accuracy	AUCROC
		Training Dataset				Validation Dataset			
Decision Tree	SVC-L1	64.09 ± 11.57	63.18 ± 7.99	63.64 ± 5.47	0.66 ± 0.05	69.09	70.91	70.0	0.67
	mRMR	65.91 ± 7.4	65.91 ± 12.06	65.91 ± 8.07	0.7 ± 0.07	74.55	69.09	71.82	0.79
Random Forest	SVC-L1	64.09 ± 16.82	68.64 ± 7.72	66.36 ± 7.45	0.72 ± 0.06	74.54	69.09	71.82	0.79
	mRMR	72.27 ± 7.17	65.91 ± 10.61	69.09 ± 4.22	0.75 ± 0.05	80.0	81.82	80.91	0.88
SVM	SVC-L1	58.63 ± 8.96	62.72 ± 10.90	60.68 ± 6.59	0.63 ± 0.07	54.54	60.0	57.27	0.61
	mRMR	69.09 ± 11.09	65.0+±1.14	67.05 ± 7.76	0.74 ± 0.09	69.09	72.73	70.90	0.79
Logistic Regression	SVC-L1	66.36 ± 10.20	64.09 ± 7.17	65.22 ± 4.87	0.72 ± 0.04	60.0	70.91	65.45	0.72
	mRMR	68.64 ± 6.88	64.55 ± 9.49	66.50 ± 5.75	0.74 ± 0.06	61.82	72.73	67.27	0.73
XGBoost	SVC-L1	65.45 ± 7.38	68.18 ± 8.86	66.81 ± 4.99	0.72 ± 0.05	74.54	70.91	72.73	0.83
	mRMR	**69.55 ± 9.76**	**69.09 ± 9.04**	**69.32 ± 5.5**	**0.75 ± 0.05**	**85.45**	**87.27**	**86.36**	**0.87**

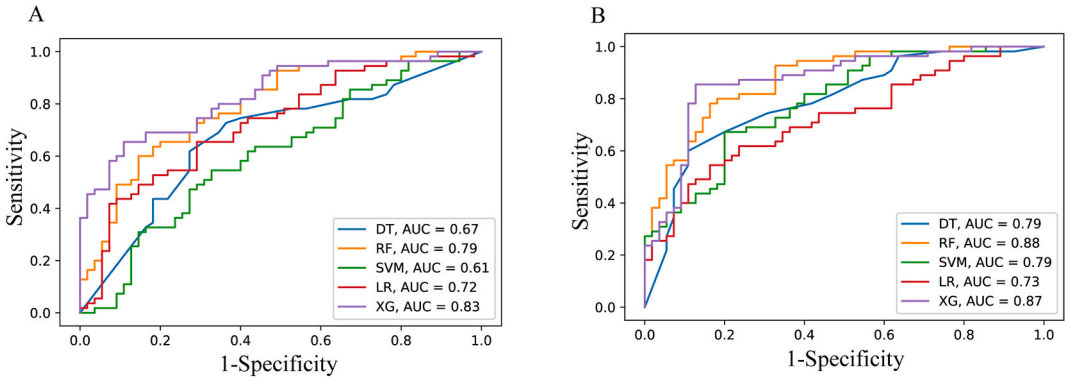


Fig. 6. AUCROC curve shows the performance of five models on validation dataset. These models were built to classify antiprotozoal from antibacterial peptides using features selected from (A) SVC-L1 and (B) mRMR. The X-axis represents the false positive rate i.e. 1-Specificity while Y-axis represents the true positive rate i.e. Sensitivity.

iii. The AUCROC curve establishes the superiority of the mRMR feature selection (Fig. 5B) as compared to SVC-L1 feature selection method (Fig. 5A) pointing to the efficient and reliable classification of antiprotozoal peptides from the antiviral ones.

4.3.3. C. Experimental results of pair 3: antiprotozoal vs antibacterial peptides

The third set of experiments utilized antiprotozoal peptides as the positive dataset and antibacterial peptides as negative dataset (pair 3). Classification models were built using 23 and 25 optimal features selected via SVC-L1 and mRMR respectively, employing five different machine learning classifiers.

The following is observed.

i. For the training dataset, the model built with mRMR selected optimum features in tandem with XGB classifier achieved the maximum average sensitivity (69.55 %), specificity (69.09 %) accuracy (69.32 %), and AUCROC (0.75) (Table 4).
ii. The performance of the models was assessed on validation datasets, revealing that model built with the XGB classifier using the top 25 optimal features selected by mRMR outperformed other achieving 85.45 % sensitivity, 87.27 % specificity, 86.36 % accuracy, and 0.87 AUCROC. (Table 4).
iii. The AUCROC curves provides further evidence supporting the superiority of mRMR feature selection methods (Fig. 6B) over the feature selection by SVC-L1 (Fig. 6A).

4.3.4. D. Experimental results of pair 4: antiprotozoal vs antifungal peptides

The fourth set of experiments employed antiprotozoal peptides as positive and antifungal peptides as negative dataset (pair 4). We build machine learning based prediction models by employing five different machine learning classifiers, using 26 and 25 optimal features selected via SVC-L1 and mRMR respectively.

The following is observed.

i. The 10-fold cross-validation using the XGB classifier on the mRMR selected optimum features achieved a maximum average sensitivity, specificity, accuracy and AUCROC of 75.91 %, 73.18 %, 74.55 % and 0.82 (Table 5**).**
ii. During external validation, the performance metrics of all models were evaluated and found that the model that performed well in internal validation, demonstrated equally impressive performance on the validation dataset with sensitivity, specificity, accuracy and AUCROC are 89.09 %, 92.73 %, 90.91 %, and 0.93 respectively (Table 5).
iii. The AUCROC curves further provide additional evidence of superiority of model built with mRMR selected features combined with the XGB classifier, achieving a score of 0.93 (Fig. 7B), compared to a score of 0.83 obtained with SVC-L1 selected features in conjunction with RF classifier (Fig. 7A).

4.3.5. E. Experimental results of pair 5: antiprotozoal vs antimicrobial peptides

In the fifth series of experiments, antiprotozoal peptides were used as the positive dataset, while antimicrobial peptides (excluding antiprotozoal peptides) as negative dataset (pair 5). We employed five machine learning classifiers on 28 and 25 best features, selected via SVC-L1 and mRMR respectively, to build our models.

The following is observed.

i. During internal validation, the model built using the XGB classifier on mRMR selected optimum features achieved a good performance metrics over the other models, with average sensitivity, specificity, accuracy and AUROC of 77.27 %, 73.18 %, 73.41 %, 0.79 respectively (Table 6).

Table 5
The performance of machine learning models developed using SVC-L1 and mRMR selected features on training and validation dataset.

Classifier	Feature Selection	Sensitivity (mean ± SD)	Specificity (mean ± SD)	Accuracy (mean ± SD)	AUROC (mean ± SD)	Sensitivity	Specificity	Accuracy	AUCROC
		Training Dataset				Validation Dataset			
Decision Tree	SVC-L1	65.91 ± 8.91	68.18 ± 12.69	67.04 ± 7.82	0.68 ± 0.07	72.73	67.27	70.0	0.75
	mRMR	70.45 ± 8.44	71.36 ± 9.32	70.91 ± 5.16	0.73 ± 0.05	80.0	80.0	80.0	0.85
Random Forest	SVC-L1	69.82 ± 9.27	72.27 ± 5.37	71.54 ± 5.54	0.81 ± 0.05	76.36	69.09	72.72	0.83
	mRMR	75.89 ± 9.33	70.0 ± 10.20	72.95 ± 6.22	0.81 ± 0.07	85.45	83.64	84.55	0.91
SVM	SVC-L1	70.91 ± 7.10	56.81 ± 10.61	63.86 ± 5.96	0.70 ± 0.06	63.64	65.45	64.54	0.72
	mRMR	69.55 ± 8.64	59.55 ± 15.27	64.55 ± 8.14	0.67 ± 0.09	67.28	67.28	67.28	0.69
Logistic Regression	SVC-L1	68.18 ± 8.86	67.27 ± 7.99	67.95 ± 5.69	0.72 ± 0.05	63.64	67.27	65.45	0.74
	mRMR	76.82 ± 8.24	67.27 ± 8.08	72.04 ± 5.38	0.79 ± 0.06	78.18	74.55	76.36	0.81
XGBoost	SVC-L1	71.36 ± 11.5	71.81 ± 7.27	71.59 ± 5.94	0.80 ± 0.06	76.36	70.91	73.64	0.81
	mRMR	**75.91 ± 7.34**	**73.18 ± 9.19**	**74.55 ± 5.06**	**0.82 ± 0.04**	**89.09**	**92.73**	**90.91**	**0.93**

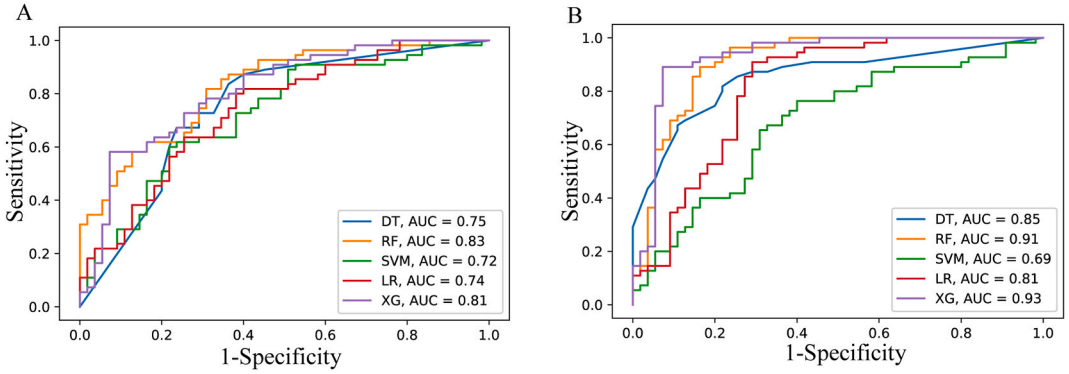


Fig. 7. ROC curve shows the performance of five models on validation dataset. These models were built to classify antiprotozoal from antifungal peptides using features selected from (A) SVC-L1 and (B) mRMR. The X-axis represents the false positive rate i.e. 1-Specificity while Y-axis represents the true positive rate i.e. Sensitivity.

ii. For external validation, all the trained models were accessed on validation dataset, revealing that the model which demonstrated superior performance during 10-fold cross validation, performed equally well with sensitivity, specificity, accuracy and AUCROC of 87.27 %, 90.91 %, 89.09 % and 0.92 respectively (Table 6).
iii. The AUCROC curves confirms the supremacy of mRMR feature selection methods with the XGB classifier (Fig. 8B) with a score of 0.92 as compared to SVC-L1 feature selection with other classifiers (Fig. 8A).

Thus we can conclude that our proposed machine-learning models were successful in classifying peptides with antiprotozoal activity from diverse set of the peptides. For training and validation datasets, it was consistently noted that the XGB classifier yielded superior performance metrics when utilizing best features selected via the mRMR algorithm. The box plot analysis of optimal features revealed that those selected by mRMR were more contrasting and differed in their median value and the interquartile spread across positive and negative datasets as compared to the SVC-L1 method (Supplementary Figs. 1–5). This may be attributed to the ability of the mRMR algorithm to select the relevant as well as the non-redundant features, a property that is instrumental in building robust classification models.

5. Interpretation of the best predictor model in each pair via lime/SHAP analysis

Comprehending the biological relevance of extracted features can pose a challenge, as machine learning models are often considered "black boxes'' due to their intricate mechanisms. Shapley Additive Explanation Algorithm (SHAP) and Local Interpretable Model-agnostic Explanations (LIME) are employed to evaluate the contribution of each feature to the model [74,75] SHAP is a global interpretation method that measures the contribution of each feature by aggregating its Shapley values. We assessed the impact of each feature on the best predictor model output by analysing SHAP values for each feature in the validation dataset through density scatter plot. LIME analysis elucidates how individual features contribute to predictions by assuming that complex models have explainable relationships in the local space of the dataset. It simplifies models through feature matrix permutations and constructs a similarity matrix to measure distances between query and perturbed sequences, also evaluating model significance per instance. LIME and SHAP analysis for one pair (antiprotozoal vs non-antimicrobial peptides) are shown in Fig. 9A and B respectively. However, the analysis for other pairs are illustrated in the Supplementary Figs. 6–9.

6. Conclusions

Apart from causing morbidity and mortality, protozoal pathogens are also getting immune to the existing drugs. The increasing immunity to the available drugs as well as the dearth of efficacious treatment against protozoan diseases highlights the urgent need for the identification of antiprotozoal regimens [76]. Therefore there is a need for the development of new therapeutics against these pathogens.

In this quest, several molecules including nanobodies [77] and nanotraps [78] have been reported to be effective against a wide range of micro-organisms. Additionally, computational methods and fragment-based drug design approaches have been used to identify new antimicrobial agents [79].

Apart from this new generation of antimicrobials, antimicrobial peptides form a class of naturally occurring compounds that have been shown to inhibit a broad range of micro-organisms. The fact that some of the peptides targeting Hepatitis [80,81], Influenza [82, 83], and Human Immunodeficiency Virus-1 [84–86] are in the preclinical or clinical phase further attests to their importance. It motivates toward designing and testing antiprotozoal peptides.

Therefore, we undertook a systematic approach and designed this study to leverage machine learning-based approaches to predict antiprotozoal peptides. To build a robust dataset for the proposed work, we performed extensive manual curation of experimentally

Table 6
The performance of machine learning models developed using SVC-L1 and mRMR selected features on training and validation dataset.

Classifier	Feature Selection	Sensitivity (mean ± SD)	Specificity (mean ± SD)	Accuracy (mean ± SD)	AUROC (mean ± SD)	Sensitivity	Specificity	Accuracy	AUCROC
		Training Dataset				Validation Dataset			
Decision Tree	SVC-L1	61.36 ± 8.19	67.27 ± 10.84	64.54 ± 6.68	0.68 ± 0.07	63.64	70.91	67.27	0.71
	mRMR	65.45 ± 9.58	67.73 ± 10.05	66.59 ± 4.2	0.68 ± 0.08	70.91	74.55	72.73	0.78
Random Forest	SVC-L1	69.09 ± 8.33	**76.36 ± 8.33**	72.73 ± 6.89	0.80 ± 0.07	76.36	80.0	78.18	0.86
	mRMR	72.27 ± 6.57	74.09 ± 10.17	73.18 ± 5.55	**0.81 ± 0.05**	83.64	83.64	83.64	0.89
SVM	SVC-L1	68.18 ± 10.56	64.54 ± 9.27	66.36 ± 3.77	0.70 ± 0.06	65.45	65.45	65.45	0.69
	mRMR	71.36 ± 7.89	70.0 ± 6.80	70.68 ± 5.12	0.76 ± 0.05	74.55	78.18	76.36	0.82
Logistic Regression	SVC-L1	74.55 ± 7.10	68.18 + 10.56	71.36 + 5.94	0.77 ± 0.07	72.72	74.55	73.64	0.82
	mRMR	70 ± 11.35	73.18 ± 8.49	72.05 ± 5.38	0.79 ± 0.05	76.36	81.82	79.09	0.83
XGBoost	SVC-L1	72.73 ± 7.61	69.55 ± 10.77	72.95 ± 6.46	0.80 ± 0.07	78.18	83.64	80.91	0.88
	mRMR	**77.27 ± 9.96**	73.18 ± 10.25	**73.41 ± 6.75**	0.79 ± 0.07	**87.27**	**90.91**	**89.09**	**0.92**

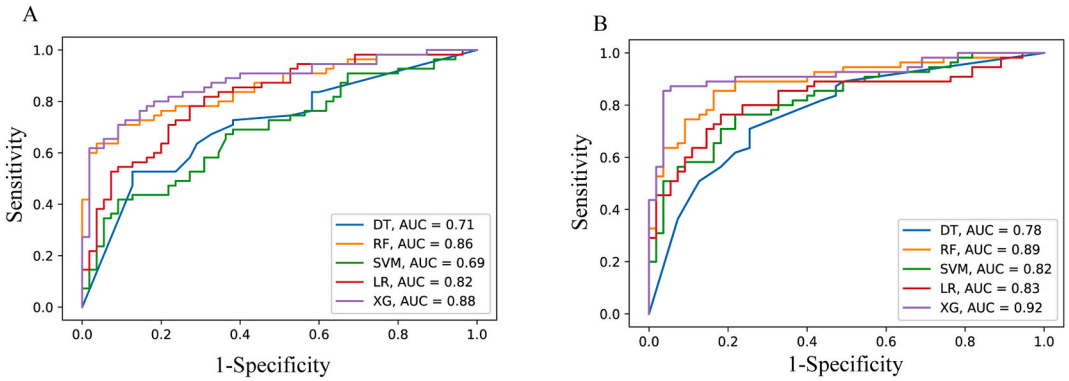


Fig. 8. ROC curve shows the performance of five models on validation dataset. These models were built to classify antiprotozoal from AMP peptides using features selected from (A) SVC-L1 and (B) mRMR. The X-axis represents the false positive rate i.e. 1-Specificity while Y-axis represents the true positive rate i.e. Sensitivity.

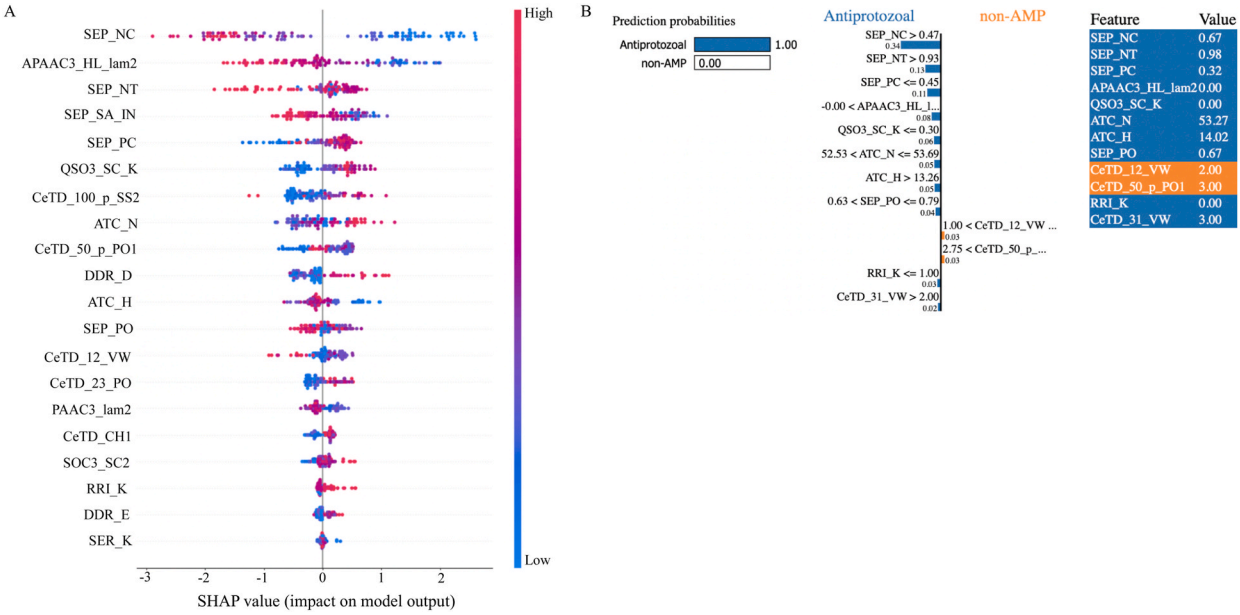


Fig. 9. (A) SHAP and (B) LIME analysis of robust model for classifying antiprotozoal from non-antimicrobial peptides.

verified antiprotozoal peptides to create a positive dataset. It has been reported that including multiple negative datasets leads to the generation of efficient models [38]. Thus we included five different negative datasets in this study. For the sake of simplification, we created five pairs of the datasets with each pair consisting of the same positive dataset i.e. antiprotozoal peptides and different negative data. The pfeature algorithm was used to compute 9151 features of the peptide sequences. Two feature selection tools viz. SVC-L1 and mRMR were then used to identify the relevant and non-redundant features. To avoid overfitting of models, we evaluated the performance of the different feature sets extracted from SVC-L1 and mRMR and found that the optimum features vary from pair to pair. We observed that the XGB classifier exhibited superior performance with optimal features selected through the mRMR algorithm in both the training and validation datasets. The results demonstrate that we were able to classify antiprotozoal peptides from all negative datasets in an efficient way. The box plot analysis revealed that features selected by mRMR for classifying antiprotozoal peptides exhibited greater diversity and contrast compared to those chosen by SVC-L1, potentially enhancing the model efficiency. Specifically, mRMR-selected features predominantly comprised combinations of two or three enriched amino acids, whereas SVC-L1 features mainly consisted of single enriched amino acids. Besides conducting external validation, we tested the efficacy of our top-performing model on an independent dataset (Supplementary Table 2). Surprisingly, our model demonstrated impressive performance (Supplementary Table 3), instilling confidence in its capability to effectively generalize to unseen data provided by the user.

To facilitate the research in this field, we developed a webserver "APPred" where the users can ascertain whether the query peptide possesses antiprotozoal properties or not. The peptides with the best-predicted scores can then be validated in the wet lab for their antiprotozoal properties.

Challenges and future prospects

In the present study, machine learning models were used to classify antiprotozoal peptides from other classes of peptides including non-antimicrobial, antiviral, antibacterial, antifungal and antimicrobial peptides. As with any good model, each new piece of information appears to raise a number of unanticipated and intriguing questions. We acknowledge one of the lacunae of this study is the limited number of experimentally validated peptides spanning all the protozoal pathogens. Therefore, to make a model more robust, we require a greater number of positive instances to elucidate with great accuracy. Unfortunately, that has not been completely achieved in the present study due to the non-availability of the data.

Additionally, our future aim is to make a multiclass model for peptides from diverse protozoal pathogens. The proposed study could not achieve this because the majority of the experimentally validated peptides belonged to 2–3 protozoal pathogens only. For instance, around 43.2 % of the experimental antiprotozoal peptides peptide pool belonged to the *Leishmania species* whereas 21.30 % and 21.90 % of the peptides were against *Trypanosoma species* and *Plasmodium species* respectively. Consequently, more studies are required to further validate and characterize the peptides against other protozoal pathogens to enable us to make a multiclass model.

Data availability statement

The positive and negative datasets used in the training and testing of the models can be downloaded from www.soodlab.com/appred.

Funding

NP is thankful to UGC for PhD fellowship. VS is the recipient of UGC-FRP award and Start-up grant from UGC.

CRediT authorship contribution statement

Neha Periwal: Writing – original draft, Methodology, Investigation, Formal analysis, Data curation. **Pooja Arora:** Writing – review & editing, Writing – original draft, Investigation, Formal analysis. **Ananya Thakur:** Investigation, Data curation. **Lakshay Agrawal:** Software, Investigation. **Yash Goyal:** Software, Investigation. **Anand S. Rathore:** Methodology, Investigation, Data curation. **Harsimrat Singh Anand:** Software, Investigation. **Baljeet Kaur:** Writing – review & editing, Writing – original draft, Supervision, Methodology, Investigation, Formal analysis. **Vikas Sood:** Writing – review & editing, Writing – original draft, Supervision, Resources, Investigation, Funding acquisition, Conceptualization.

Declaration of competing interest

The authors declare no conflict of interest.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.heliyon.2024.e36163.

References

[1] R. Siddiqui, N.A. Khan, Biology and pathogenesis of Acanthamoeba, Parasit Vectors 5 (2012) 6.
[2] M. Jahangeer, Z. Mahmood, N. Munir, et al., Naegleria fowleri: sources of infection, pathophysiology, diagnosis, and management; a review, Clin. Exp. Pharmacol. Physiol. 47 (2020) 199–212.
[3] M. Naveed, U. Ali, T. Aziz, et al., Development and immunological evaluation of an mRNA-based vaccine targeting Naegleria fowleri for the treatment of primary amoebic meningoencephalitis, Sci. Rep. 14 (2024) 767.
[4] J. Farrar, P. Hotez, T. Junghanss, et al., Manson's Tropical Diseases E-Book, 2013.
[5] I. Gupta, P. Guin, Communicable diseases in the south-east asia region of the world health organization: towards a more effective response, Bull. World Health Organ. 88 (2010) 199–205.
[6] Z.A. Bhutta, J. Sommerfeld, Z.S. Lassi, et al., Global burden, distribution, and interventions for infectious diseases of poverty, Infectious diseases of poverty 3 (2014) 1–7.
[7] K. Ohnishi, N. Sakamoto, K. Kobayashi, et al., Subjective adverse reactions to metronidazole in patients with amebiasis, Parasitol. Int. 63 (2014) 698–700.
[8] X. Su, K.D. Lane, L. Xia, et al., Plasmodium genomics and genetics: new insights into malaria pathogenesis, drug resistance, epidemiology, and evolution, Clin. Microbiol. Rev. 32 (2019) e00019, 19.
[9] A. Ponte-Sucre, F. Gamarro, J.-C. Dujardin, et al., Drug resistance and treatment failure in leishmaniasis: a 21st century challenge, PLoS Neglected Trop. Dis. 11 (2017) e0006052.
[10] M.-C.N. Laffitte, P. Leprohon, B. Papadopoulou, et al., Plasticity of the Leishmania genome leading to gene copy number variations and drug resistance, F1000Research 5 (2016).
[11] J.-M. Ubeda, D. Légaré, F. Raymond, et al., Modulation of gene expression in drug resistant Leishmania is associated with gene amplification, gene deletion and chromosome aneuploidy, Genome biology 9 (2008) 1–16.
[12] R.L. Dunne, A.D. Linda, P. Upcroft, et al., Drug resistance in the sexually transmitted protozoan Trichomonas vaginalis, Cell Res. 13 (2003) 239–249.

[13] Y. Huan, Q. Kong, H. Mou, et al., Antimicrobial peptides: classification, design, application and research progress in multiple fields, Front. Microbiol. (2020) 2559.
[14] H.X. Luong, T.T. Thanh, T.H. Tran, Antimicrobial peptides–Advances in development of therapeutic applications, Life Sci. 260 (2020) 118407.
[15] M. Pasupuleti, A. Schmidtchen, M. Malmsten, Antimicrobial peptides: key components of the innate immune system, Crit. Rev. Biotechnol. 32 (2012) 143–171.
[16] J.K. Boparai, P.K. Sharma, Mini review on antimicrobial peptides, sources, mechanism and recent applications, Protein Pept. Lett. 27 (2020) 4–16.
[17] J.M. David, A.K. Rajasekaran, Gramicidin A: a new mission for an old antibiotic, Journal of kidney cancer and VHL 2 (2015) 15.
[18] A. Prince, P. Sandhu, P. Ror, et al., Lipid-II independent antimicrobial mechanism of nisin depends on its crowding and degree of oligomerization, Sci. Rep. 6 (2016) 1–15.
[19] M. Torrent, D. Pulido, L. Rivas, et al., Antimicrobial peptide action on parasites, Curr. Drug Targets 13 (2012) 1138–1147.
[20] S.E.C. Maluf, C. Dal Mas, E. Oliveira, et al., Inhibition of malaria parasite Plasmodium falciparum development by crotamine, a cell penetrating peptide from the snake venom, Peptides 78 (2016) 11–16.
[21] C.M. Adade, I.R. Oliveira, J.A. Pais, et al., Melittin peptide kills Trypanosoma cruzi parasites by inducing different cell death pathways, Toxicon 69 (2013) 227–239.
[22] I.C.J. Bandeira, D. Bandeira-Lima, C.P. Mello, et al., Antichagasic effect of crotalicidin, a cathelicidin-like vipericidin, found in Crotalus durissus terrificus rattlesnake's venom gland, Parasitology 145 (2018) 1059–1064.
[23] L. Giovati, C. Santinoli, C. Mangia, et al., Novel activity of a synthetic decapeptide against Toxoplasma gondii tachyzoites, Front. Microbiol. 9 (2018) 753.
[24] A. Zhavoronkov, Y.A. Ivanenkov, A. Aliper, et al., Deep learning enables rapid identification of potent DDR1 kinase inhibitors, Nat. Biotechnol. 37 (2019) 1038–1040.
[25] C.T. Madsen, J.C. Refsgaard, F.G. Teufel, et al., Combining mass spectrometry and machine learning to discover bioactive peptides, Nat. Commun. 13 (2022) 6235.
[26] A. Pande, S. Patiyal, A. Lathwal, et al., Computing wide range of protein/peptide features from their sequence and structure, bioRxiv (2019) 599126.
[27] J. Vamathevan, D. Clark, P. Czodrowski, et al., Applications of machine learning in drug discovery and development, Nat. Rev. Drug Discov. 18 (2019) 463–477.
[28] S. Dara, S. Dhamercherla, S.S. Jadav, et al., Machine learning in drug discovery: a review, Artif. Intell. Rev. 55 (2022) 1947–1999.
[29] S.K. Dhanda, P. Vir, G.P. Raghava, Designing of interferon-gamma inducing MHC class-II binders, Biol. Direct 8 (2013) 1–15.
[30] S.K. Dhanda, S. Gupta, P. Vir, et al., Prediction of IL4 inducing peptides, Clin. Dev. Immunol. 2013 (2013).
[31] G. Nagpal, S.S. Usmani, S.K. Dhanda, et al., Computer-aided designing of immunosuppressive peptides based on IL-10 inducing potential, Sci. Rep. 7 (2017) 1–10.
[32] S. Gupta, P. Mittal, M.K. Madhu, et al., IL17eScan: a tool for the identification of peptides inducing IL-17 response, Front. Immunol. 8 (2017) 1430.
[33] P. Arora, N. Periwal, Y. Goyal, et al., iIL13Pred: improved prediction of IL-13 inducing peptides using popular machine learning classifiers, BMC Bioinf. 24 (2023) 141.
[34] J. Xu, F. Li, A. Leier, et al., Comprehensive assessment of machine learning-based methods for predicting antimicrobial peptides, Briefings Bioinf. 22 (2021) bbab083.
[35] A. Capecchi, X. Cai, H. Personne, et al., Machine learning designs non-hemolytic antimicrobial peptides, Chem. Sci. 12 (2021) 9221–9232.
[36] C.M. Van Oort, J.B. Ferrell, J.M. Remington, et al., AMPGAN v2: machine learning-guided design of antimicrobial peptides, J. Chem. Inf. Model. 61 (2021) 2198–2207.
[37] Y. Pang, L. Yao, J.-H. Jhong, et al., AVPIden: a new scheme for identification and functional prediction of antiviral peptides based on machine learning approaches, Briefings Bioinf. 22 (2021) bbab263.
[38] Y. Pang, Z. Wang, J.-H. Jhong, et al., Identifying anti-coronavirus peptides by incorporating different negative datasets and imbalanced learning strategies, Briefings Bioinf. 22 (2021) 1085–1095.
[39] V. Singh, S. Shrivastava, S. Kumar Singh, et al., StaBle-ABPpred: a stacked ensemble predictor based on biLSTM and attention mechanism for accelerated discovery of antibacterial peptides, Briefings Bioinf. 23 (2022) bbab439.
[40] R. Sharma, S. Shrivastava, S. Kumar Singh, et al., Deep-ABPpred: identifying antibacterial peptides in protein sequences using bidirectional LSTM with word2vec, Briefings Bioinf. 22 (2021) bbab065.
[41] V. Singh, S. Shrivastava, S. Kumar Singh, et al., Accelerating the discovery of antifungal peptides using deep temporal convolutional networks, Briefings Bioinf. 23 (2022).
[42] A. Ahmad, S. Akbar, S. Khan, et al., Deep-AntiFP: prediction of antifungal peptides using distanct multi-informative features incorporating with deep neural networks, Chemometr. Intell. Lab. Syst. 208 (2021) 104214.
[43] S. Egieyeh, J. Syce, S.F. Malan, et al., Predictive classifier models built from natural products with antimalarial bioactivity using machine learning approach, PLoS One 13 (2018) e0204644.
[44] D.J. Mason, R.T. Eastman, R.P. Lewis, et al., Using machine learning to predict synergistic antimalarial compound combinations with novel structures, Front. Pharmacol. (2018) 1096.
[45] C.M. Morang'a, L. Amenga–Etego, S.Y. Bah, et al., Machine learning approaches classify clinical malaria outcomes based on haematological parameters, BMC Med. 18 (2020) 1–16.
[46] M. Poostchi, K. Silamut, R.J. Maude, et al., Image analysis and machine learning for detecting malaria, Transl. Res. 194 (2018) 36–55.
[47] S.H. Gulsen, E. Tileklioglu, E. Bode, et al., Antiprotozoal activity of different Xenorhabdus and Photorhabdus bacterial secondary metabolites and identification of bioactive compounds using the easyPACId approach, Sci. Rep. 12 (2022) 10779.
[48] M.E. Mswahili, G.L. Martin, J. Woo, et al., Antimalarial drug predictions using molecular descriptors and machine learning against Plasmodium falciparum, Biomolecules 11 (2021) 1750.
[49] Q. Liu, J. Deng, M. Liu, Classification models for predicting the antimalarial activity against Plasmodium falciparum, SAR QSAR Environ. Res. 31 (2020) 313–324.
[50] null Danishuddin, G. Madhukar, M.Z. Malik, et al., Development and rigorous validation of antimalarial predictive models using machine learning approaches, SAR QSAR Environ. Res. 30 (2019) 543–560.
[51] G. Wang, X. Li, Z. Wang, APD3: the antimicrobial peptide database as a tool for research and education, Nucleic Acids Res. 44 (2016) D1087–D1093.
[52] G. Shi, X. Kang, F. Dong, et al., Dramp 3.0: an enhanced comprehensive data repository of antimicrobial peptides, Nucleic Acids Res. 50 (2022) D488–D496.
[53] D. Mehta, P. Anand, V. Kumar, et al., ParaPep: a web resource for experimentally validated antiparasitic peptide sequences and their structures, Database 2014 (2014) bau051.
[54] M. Pirtskhalava, A.A. Amstrong, M. Grigolava, et al., DBAASP v3: database of antimicrobial/cytotoxic activity and structure of peptides as a resource for development of new therapeutics, Nucleic Acids Res. 49 (2021) D288–D297.
[55] C.-R. Chung, T.-R. Kuo, L.-C. Wu, et al., Characterization and identification of antimicrobial peptides with different functional activities, Briefings Bioinf. 21 (2020) 1098–1114.
[56] R. Sharma, S. Shrivastava, S. Kumar Singh, et al., AniAMPpred: artificial intelligence guided discovery of novel antimicrobial peptides in animal kingdom, Briefings Bioinf. 22 (2021) bbab242.
[57] S.S. Usmani, S. Bhalla, G.P.S. Raghava, Prediction of antitubercular peptides from sequence information using ensemble classifier and hybrid features, Front. Pharmacol. 9 (2018) 954.
[58] L.D. Naorem, N. Sharma, G.P.S. Raghava, A web server for predicting and scanning of IL-5 inducing peptides using alignment-free and alignment-based method, Comput. Biol. Med. 158 (2023) 106864.
[59] A. Dhall, S. Patiyal, N. Sharma, et al., Computer-aided prediction and design of IL-6 inducing peptides: IL-6 plays a crucial role in COVID-19, Briefings Bioinf. 22 (2021) 936–945.
[60] J. Shen, J. Zhang, X. Luo, et al., Predicting protein–protein interactions based only on sequences information, Proc. Natl. Acad. Sci. USA 104 (2007) 4337–4341.

[61] F. Li, Y. Zhou, Y. Zhang, et al., POSREG: proteomic signature discovered by simultaneously optimizing its reproducibility and generalizability, Briefings Bioinf. 23 (2022).
[62] Q. Yang, B. Li, J. Tang, et al., Consistent gene signature of schizophrenia identified by a novel feature selection strategy from comprehensive sets of transcriptomic data, Briefings Bioinf. 21 (2020) 1058–1068.
[63] J. Fu, Y. Zhang, Y. Wang, et al., Optimization of metabolomic data processing using NOREVA, Nat. Protoc. 17 (2022) 129–151.
[64] Q. Yang, Y. Wang, Y. Zhang, et al., NOREVA: enhanced normalization and evaluation of time-course and multi-class metabolomic data, Nucleic Acids Res. 48 (2020) W436–W448.
[65] B. Li, J. Tang, Q. Yang, et al., NOREVA: normalization and evaluation of MS-based metabolomics data, Nucleic acids research 45 (2017) W162–W170.
[66] J. Tang, J. Fu, Y. Wang, et al., ANPELA: analysis and performance assessment of the label-free quantification workflow for metaproteomic studies, Briefings Bioinf. 21 (2020) 621–636.
[67] P.S. Bradley, O.L. Mangasarian, Feature Selection via Concave Minimization and Support Vector Machines, vol. 98, 1998, pp. 82–90.
[68] B. Peng, L. Wang, Y. Wu, An error bound for l1-norm support vector machine coefficients in ultra-high dimension, J. Mach. Learn. Res. 17 (2016) 1–26.
[69] H. Peng, F. Long, C. Ding, Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy, IEEE Trans. Pattern Anal. Mach. Intell. 27 (2005) 1226–1238.
[70] C. Cortes, V. Vapnik, Support-vector networks, Mach. Learn. 20 (1995) 273–297.
[71] S. Akbar, M. Hayat, M. Tahir, et al., cACP-DeepGram: classification of anticancer peptides via deep neural network and skip-gram-based word embedding model, Artif. Intell. Med. 131 (2022) 102349.
[72] S. Akbar, S. Khan, F. Ali, et al., iHBP-DeepPSSM: identifying hormone binding proteins using PsePSSM based evolutionary features and deep learning approach, Chemometr. Intell. Lab. Syst. 204 (2020) 104103.
[73] S. Akbar, M. Hayat, M. Iqbal, et al., iACP-GAEnsC: evolutionary genetic algorithm based ensemble classification of anticancer peptides by utilizing hybrid feature space, Artif. Intell. Med. 79 (2017) 62–70.
[74] S. Akbar, A. Raza, T. Al Shloul, et al., pAtbP-EnC: identifying anti-tubercular peptides using multi-feature representation and genetic algorithm based deep ensemble model, IEEE Access (2023).
[75] A. Ahmad, S. Akbar, M. Tahir, et al., iAFPs-EnC-GA: identifying antifungal peptides using sequential and evolutionary descriptors based multi-information fusion and ensemble learning approach, Chemometr. Intell. Lab. Syst. 222 (2022) 104516.
[76] E.J.A. Luna, S.R.S.L.D.C. Campos, Vaccine development against neglected tropical diseases, Cad. Saúde Pública 36 (2020) e00215720.
[77] Y. Mei, Y. Chen, J.P. Sivaccumar, et al., Research progress and applications of nanobody in human infectious diseases, Front. Pharmacol. 13 (2022) 963978.
[78] M. Chen, J. Rosenberg, X. Cai, et al., Nanotraps for the containment and clearance of SARS-CoV-2, Matter 4 (2021) 2059–2082.
[79] Z. Breijyeh, R. Karaman, Design and synthesis of novel antimicrobial agents, Antibiotics 12 (2023) 628.
[80] P. Bogomolov, N. Voronkova, K. Schoeneweis, et al., A Proof-Of-Concept Phase IIa Clinical Trial to Treat Chronic HBV/HDV with the Entry Inhibitor Myrcludex B, vol. 63, 2016, p. 121A, 121A.
[81] H. Wedemeyer, K. Schoneweis, P. Bogomolov, et al., Final results of a multicenter, open-label phase 2 clinical trial (MYR203) to assess safety and efficacy of myrcludex B in combination with PEG-interferon Alpha 2a in patients with chronic HBV/HDV co-infection, J. Hepatol. 70 (2019) e81.
[82] H. Badani, R.F. Garry, T.G. Voss, et al., Mechanism of action of flufirvitide, a peptide inhibitor of influenza virus infection, Biophys. J. 106 (2014) 707a.
[83] S. Skalickova, Z. Heger, L. Krejcova, et al., Perspective of use of antiviral peptides against influenza virus, Viruses 7 (2015) 5428–5442.
[84] D. Yu, X. Ding, Z. Liu, et al., Molecular mechanism of HIV-1 resistance to sifuvirtide, a clinical trial–approved membrane fusion inhibitor, J. Biol. Chem. 293 (2018) 12703–12718.
[85] L. Li, Y. Ben, S. Yuan, et al., Efficacy, stability, and biosafety of sifuvirtide gel as a microbicide candidate against HIV-1, PLoS One 7 (2012) e37381.
[86] X. Yao, H. Chong, C. Zhang, et al., Broad antiviral activity and crystal structure of HIV-1 fusion inhibitor sifuvirtide, J. Biol. Chem. 287 (2012) 6788–6796.