IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 27 October 2024, accepted 23 December 2024, date of publication 26 December 2024, date of current version 2 January 2025.

Digital Object Identifier 10.1109/ACCESS.2024.3523068

RESEARCH ARTICLE

Anticancer Peptides Classification Using Long-Short-Term Memory With Novel Feature Representation

NAZER AL TAHIFAH[1,2], MUHAMMAD SOHAIL IBRAHIM[3], ERUM REHMAN[4], NAVEED AHMED[5], ABDUL WAHAB[6], AND SHUJAAT KHAN[1,2]

[1]Department of Computer Engineering, College of Computing and Mathematics, King Fahd University of Petroleum & Minerals, Dhahran 31261, Saudi Arabia
[2]SDAIA-KFUPM Joint Research Center for Artificial Intelligence, King Fahd University of Petroleum & Minerals, Dhahran 31261, Saudi Arabia
[3]Department of Mechanical Systems Engineering, Kumoh National Institute of Technology, Gumi-si 39177, South Korea
[4]Department of Mathematics, Nazarbayev University, Astana 010000, Kazakhstan
[5]Center for Applied Mathematics and Bio-Informatics (CAMB), Department of Mathematics and Natural Sciences, Gulf University for Science and Technology (GUST), Mubarak Al-Abdullah 32093, Kuwait
[6]Department of Mathematics, College of Science, Sultan Qaboos University, Muscat 123, Oman

Corresponding author: Shujaat Khan (shujaat.khan@kfupm.edu.sa)

The work of Nazer Al Tahifah and Shujaat Khan was supported by the Saudi Data and AI Authority (SDAIA) and King Fahd University of Petroleum & Minerals (KFUPM) under SDAIA-KFUPM Joint Research Center for Artificial Intelligence under Grant JRC-AI-RFP-16. The work of Erum Rehman was supported by Nazarbayev University, Kazakhstan through the Faculty Development Competitive Research Grant Program (FDCRGP) under Grant 1022021FD2914. The work of Naveed Ahmed was supported by Gulf University for Science and Technology (GUST) Internal Seed under Grant Case no 74.

ABSTRACT Cancer treatment is a challenging endeavor because of the intricacy, heterogeneity, and diversity of cancer causes. Comprehensive therapeutic approaches are crucial for cancer treatment. *Anticancer peptides* (ACPs) present a potentially effective therapeutic option. However, the extensive identification and synthesis of these peptides present a persistent difficulty that calls for the creation of effective prediction techniques. Existing techniques either suffer from low accuracy or employ high-dimensional feature sets, frequently producing sparse features and leading to ineffective model designs. This work presents a novel set of features and a *long-short-term-memory* (LSTM)-based classification strategy to create an efficient model. The suggested feature set includes three new and two modern feature extraction methods. The binary profile feature and k-mer sparse matrix of the reduced amino acid alphabet are part of the modern feature set. The combination of the *composition of the K-spaced side chain pairs* (CKSSCP), the *composition of the K-spaced electrically charged side chain pairs* (CKSECSCP), and the *combination of [pk(CO2H)] + [pk(NH2)] + [pk(R)] + [isoelectric point]* is used to derive the novel features. The suggested LSTM model is trained using the combined feature set. The trials are carried out with a k-fold cross-validation method on benchmark datasets. The results indicate that the proposed model outperforms alternative ACP classification techniques in terms of *Mathew's correlation coefficient* (MCC) and accuracy. The *ACP740* dataset with 5-folds yields an MCC score of 75%, which is 12%, 11%, 3%, and 8% greater than those of the ACP-DL, ACP-DA, ACP-MHCNN, and ACP-KSRC approaches, respectively. For the *ACP344* dataset with 10-folds, the proposed method achieves an MCC score of 85.14%, which is 23% and 2% higher than the MCC scores of ACP-DL and SAP methods, respectively. Better classification performance offered by the proposed approach could help identify new ACPs and better understand their structural and chemical characteristics. The source code and the datasets are available on the author's GitHub page (https://github.com/Shujaat123/ACP-LSTM-NFR).

The associate editor coordinating the review of this manuscript and approving it for publication was Mohammad Zia Ur Rahman.

INDEX TERMS Anticancer peptides (ACPs), composition of K-spaced amino-acid pairs (CKSAAP), long-short-term-memory (LSTM), composition of the K-spaced side chain pairs (CKSSCP), composition of the K-spaced electrically charged side chain pairs (CKSECSCP), isoelectric point (pI).

I. INTRODUCTION

Healthy cells in the human body develop and reproduce through division, an important process for maintaining the body's population of healthy cells. Cells die as they age or become damaged, making room for new, healthy ones. However, cancer disrupts the process of healthy cell reproduction through an uncontrolled expansion of damaged cells. When damaged, these cells multiply uncontrollably, potentially leading to grave effects, including death [1]. Several procedures are used to restrict unregulated cell development, such as surgery, radiation therapy, and chemotherapy [2]. However, these procedures typically have major adverse effects and can cause severe discomfort [3].

The discovery of peptides significantly advanced the medical field in the avenues of disease treatment [4] and drug discovery [5]. Several well-known peptides have emerged, such as *anti-oxidant peptides* (AoP) [6], *anti-microbial peptides* (AMPs) [7], *anti-cancer peptides* (ACPs) [8], and others. These peptides, like others, typically consist of a chain of several amino acids derived from natural sources such as animals and plants. Advances in synthetic biology have also significantly expanded the range and functionality of peptides used in research and therapeutics. This unique therapeutic approach is recommended due to its specific targeting of cancer cells. Unlike chemotherapy, this approach selectively targets damaged cells, resulting in minimal adverse effects and excellent precision in treatment [9].

ACPs are difficult to distinguish from other peptides because of their significant intra-class variability and low inter-class separability [10]. Given the diversity of amino acid combinations in ACPs, standard identification procedures may be inadequate. Developing alternative approaches to differentiate ACPs from other peptides is critical for cancer treatment research [11]. Alternative techniques, such as machine learning-based classification strategies [12], [13], can improve the efficiency and accuracy of ACP discovery.

Researchers have exploited the widespread applicability of AI to tackle the vast search space of ACPs. Recently, many researchers have used machine learning models to classify and predict protein/peptide sequences [14], [15]. These works investigated various ways of classifying unknown sequences. The study in [16] used a *long-short-term memory* (LSTM) network in conjunction with *k-mer sparse matrix* and *binary profile features* (bof), achieving promising classification results. Other strategies used the *composition of K-spaced amino-acid pairs* (CKSAAP) as a feature extraction method, along with other learning algorithms [14], [17], [18]. In a similar context, multiple ACP classification algorithms were developed [19], [20], exhibiting promising performance in ACP prediction. Further research and experimentation are deemed necessary. For instance, the ACP-KSRC study [19] uses sparse representation classifier [21] and found that feature extraction approaches such as CKSAAP significantly improved performance. However, large feature space dimensions require the use of dimension reduction techniques such as *principal component analysis* (PCA), which can be limited in preserving large variations captured by the CKSAAP feature set. In contrast, feature extraction approaches such as k-mer sparse matrix or bpf do not necessarily require dimension reduction techniques. By adding additional characteristics to these methods, the overall resilience of the feature set can be improved.

II. PROPOSED APPROACH

The proposed approach is based on the utilization of LSTM networks, with a focus on feature extraction, aiming to surpass the performance of existing machine learning models. This section will comprehensively outline each step of the classification process, as illustrated in the block diagram in Fig. 1, while also showcasing the proposed feature extraction approach.

A. DATASET

Researchers used a variety of datasets, including those presented in [11], [22], and [23]. We evaluated two datasets and compared our findings to state-of-the-art approaches using different evaluation protocols. The first dataset, reported by Chen et al. [22] and Wei et al. [23], is *ACP740*. It comprises 740 peptide sequences, with 376 classified as ACPs and the remaining 364 as non-ACPs. The second dataset, *ACP344*, presented in [11], consists of 344 peptide sequences, with 138 labeled as ACPs and the remaining 206 designated as non-ACPs. The *ACP740* dataset was evaluated using a 5-fold cross-validation protocol. In comparison, the *ACP344* dataset was evaluated using both 5-fold and 10-fold cross-validation.

B. FEATURES ENCODING

There are several approaches to feature-encoding peptide sequences, and no one approach is superior because they all have unique characteristics and advantages of their own. Different strategies might work better with different machine learning algorithms. For example, a strategy used in the ACP-KSRC model [19], when applied to LSTM [16], might not perform well or better as in the ACP-KSRC [19]. All feature encoding techniques, however, contribute to the advancement of cancer treatment research by improving our understanding of the characteristics of the ACP sequences, independent of the classification algorithm employed. Five feature encoding techniques–some new, some adapted from previous approaches–are explained in this study and used in the suggested strategy. These methods are combined to create an extensive set of features to optimize the overall performance of the ACP classification.

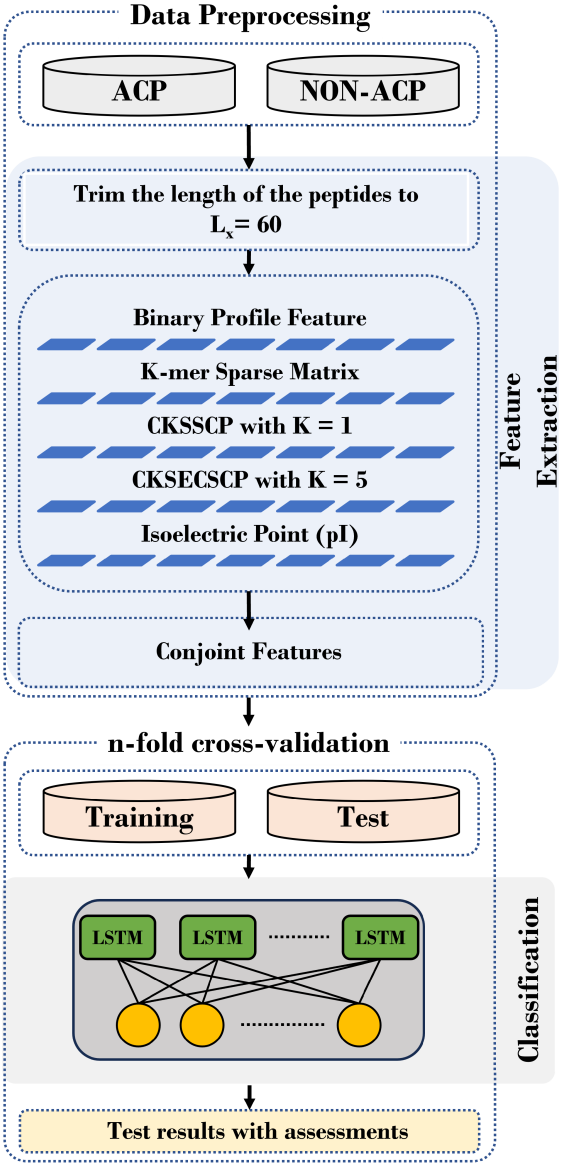


FIGURE 1. The flowchart of the proposed approach.

1) BINARY PROFILE

Analyzing the peptide sequence representation before delving into the composition of the co-joint feature set is crucial. These sequences are composed of a selection from the 20 available amino acids, each represented by a character in the alphabet (A, C, D, E, F, G, H, I, K, L, M, N, P, Q, R, S, T, V, W, and Y). Consequently, a peptide sequence can be accurately represented using the amino acid alphabet characters as

$$\mathrm{P} := p_1, p_2, p_3, \ldots, p_{L_x}. \tag{1}$$

Here, p_n, for $n = 1, \cdots, L_x$, represents the nth amino acid in the peptide sequence P of length L_x. Each amino acid is encoded as a string of 0's and 1's in the *binary profile*. Specifically, the A amino acid can be encoded as $V(A) = (1, 0, 0, \cdots, 0)$, the C amino acid as $V(C) = (0, 1, 0, \cdots, 0)$, and so forth, until Y is encoded as $V(Y) = (0, \cdots, 0, 0, 1)$. This encoding furnishes the vector

$$\mathrm{BP}(L_x) := [V(p_1), V(p_2), V(p_3), \cdots, V(p_{L_x})].$$

It is suggested in [16] to use BP(7) instead of encoding the entire peptide sequence using $BP(L_x)$. Thus, limiting the features for any peptide sequence encoded to a 140-dimensional vector.

2) K-MER SPARSE MATRIX

The second characteristic incorporated into the co-joint set is the k-mer sparse matrix [16]. In this encoding, 20 distinct amino acids are grouped depending on the side chain and the dipole moments. This leads to the 7 groups [Ala, Gly, Val], [Ile, Leu, Phe, Pro], [Tyr, Met, Thr, Ser], [His, Asn, Gln, Tpr], [Arg, Lys], [Asp, Glu], and [Cys] (see [24], [25], [26]). Consequently, the peptide sequence consists of just 7 group memberships. Next, we iterate through amino acids in the sequence, replacing each of the 20 amino acids with the new 7-alphabet set corresponding to the relevant group membership. An example of k-mers feature extraction with $L_x = 10$ sequence length and $k = 3$ amino-acid groups is shown in Fig. 2, yielding a 343×8 k-mers sparse matrix. In general, for a peptide sequence of length L_x, there would be 7^k possible k-mer and $(L_x - k + 1)$ steps to iterate through the peptide sequence, transforming it into a $7^k \times (L_x - k + 1)$ k-mer sparse matrix S. In this study, we set $k = 3$ and define the k-mer sparse matrix S as

$$S := (q_{ij})_{7^k \times (L_x - k + 1)},$$
$$q_{ij} := \begin{cases} 1, & \text{if } m_j m_{j+1} m_{j+2} = k\text{-mer}(i) \\ 0, & \text{otherwise} \end{cases}$$

A low-rank feature set comprising the raw data from the original sequence can be created by transforming the resulting S matrix. To further decrease the dimensions of S, the singular value decomposition (SVD) [27] is applied, resulting in a 343-dimensional row vector.

3) COMPOSITION OF THE K-SPACED SIDE CHAIN PAIRS (CKSSCP)

As described in Section II-B1, special alphabet characters represent each amino acid in protein/peptide sequences [28]. Since there are 20 amino acids in total, there are $20^2 = 400$ potential pairings of the amino acids. Using this information, we define

$$\hat{\boldsymbol{\psi}}_0 := [\psi_{AA}, \psi_{AC}, \psi_{AD}, \ldots, \psi_{YY}]^T \in \mathbb{R}^{400},$$

where each component ψ_{xy} of the $\hat{\boldsymbol{\psi}}_0$ is a possible pair of amino acids. However, as was already noted, a vector of that size becomes problematic if the dimension of the attributes exceeds the number of samples in the dataset. We remind that the datasets *ACP740* and *ACP344* considered here have 740 and 344 samples, respectively. To this end, we use the side chain feature [29] to separate the amino acids into

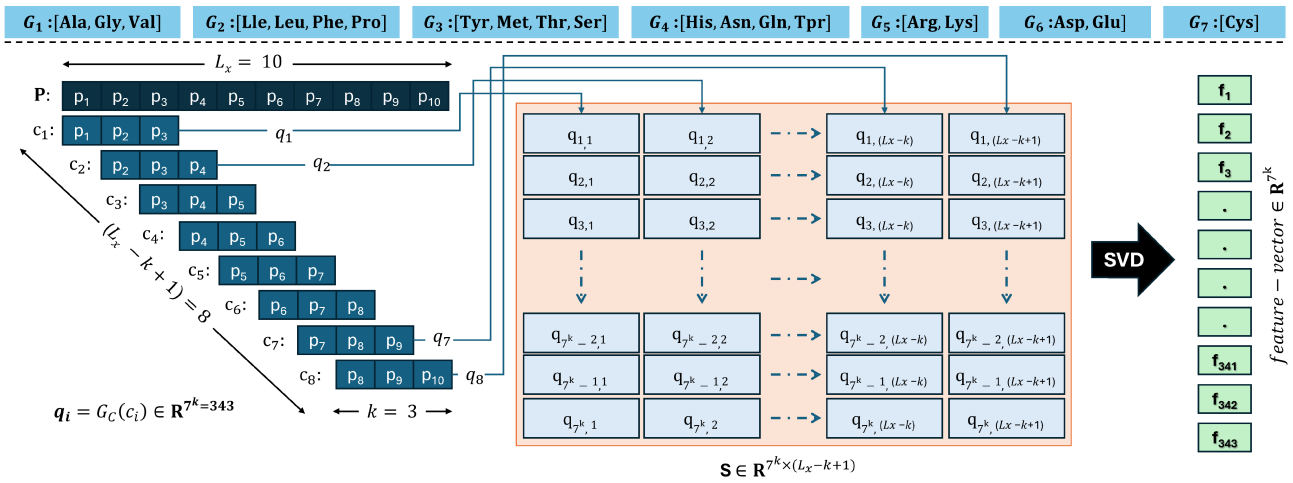


FIGURE 2. Illustration of the k-mers feature extraction for a sequence of length $L_x = 10$, with $k = 3$.

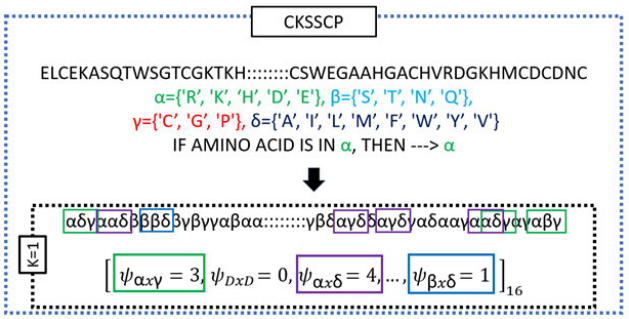


FIGURE 3. Illustration of the CKSSCP with $k = 1$.

4 groups, $\alpha = [R, K, H, D, E]$, $\beta = [S, T, N, Q]$, $\gamma = [C, G, P]$, and $\delta = [A, I, L, M, F, W, Y, V]$. This leads to

$$\boldsymbol{\psi}_0 := [\psi_{\alpha\alpha}, \psi_{\alpha\beta}, \psi_{\alpha\gamma}, \cdots, \psi_{\delta\delta}]^T \in \mathbb{R}^{16}.$$

For the gap between the pairs $k = 4$, we also form

$$\boldsymbol{\psi}_1 := [\psi_{\alpha x\alpha}, \psi_{\alpha x\beta}, \psi_{\alpha x\gamma}, \cdots \psi_{\delta x\delta}]^T \in \mathbb{R}^{16},$$
$$\boldsymbol{\psi}_2 := [\psi_{\alpha xx\alpha}, \psi_{\alpha xx\beta}, \psi_{\alpha xx\gamma}, \cdots, \psi_{\delta xx\delta}]^T \in \mathbb{R}^{16},$$
$$\Psi_k := [\boldsymbol{\psi}_0, \boldsymbol{\psi}_1, \boldsymbol{\psi}_2, \cdots, \boldsymbol{\psi}_k]^T \in \mathbb{R}^{16(k+1)}.$$

Fig. 3 shows an illustration for ψ_1 where $k = 1$.

4) COMPOSITION OF THE K-SPACED ELECTRICALLY CHARGED SIDE CHAIN PAIRS (CKSECSCP)

The structure of the CKSECSCP feature is similar to the CKSSCP feature, but it takes a different approach. This feature is primarily concerned with determining if amino acids have positive, negative, or neutral charges. Compared to CKSSCP, this feature only targets two groups: $\alpha_1 := [R, K, H]$ and $\alpha_2 = [D, E]$, where α_1 represents positive charges and α_2 represents negative charges [29]. Thus, there are five groups overall that are created by further classifying the α group into α_1 and α_2 based on the side chain groups $\alpha = [R, K, H, D, E]$, $\beta = [S, T, N, Q]$, $\gamma = [C, G, P]$, and $\delta = [A, I, L, M, F, W, Y, V]$. We construct a normalized

TABLE 1. The combination of all possible attributes in $\hat{\psi}_0$.

$\psi_{\alpha_1\alpha_1}$	$\psi_{\alpha_2\alpha_1}$	$\psi_{\beta\alpha_1}$	$\psi_{\gamma\alpha_1}$	$\psi_{\delta\alpha_1}$
$\psi_{\alpha_1\alpha_2}$	$\psi_{\alpha_2\alpha_2}$	$\psi_{\beta\alpha_2}$	$\psi_{\gamma\alpha_2}$	$\psi_{\delta\alpha_2}$
$\psi_{\alpha_1\beta}$	$\psi_{\alpha_2\beta}$	$\psi_{\beta\beta}$	$\psi_{\gamma\beta}$	$\psi_{\delta\beta}$
$\psi_{\alpha_1\gamma}$	$\psi_{\alpha_2\gamma}$	$\psi_{\beta\gamma}$	$\psi_{\gamma\gamma}$	$\psi_{\delta\gamma}$
$\psi_{\alpha_1\delta}$	$\psi_{\alpha_2\delta}$	$\psi_{\beta\delta}$	$\psi_{\gamma\delta}$	$\psi_{\delta\delta}$

vector

$$\hat{\boldsymbol{\psi}}_0 := \frac{1}{N_0}\left[\frac{\psi_{\alpha_1\alpha_1}}{\text{count}(\alpha_1\alpha_1)}, \frac{\psi_{\alpha_1\alpha_2}}{\text{count}(\alpha_1\alpha_2)}, \cdots, \frac{\psi_{\delta\delta}}{\text{count}(\delta\delta)}\right]^T \in \mathbb{R}^{25},$$

where ψ_{xy} is the electrical charge of the k-spaced pair, k is the gap between the pairs and $N_0 := L_x - (k + 1)$. The quantity count(xy) is the total number of occurrences of the xy pair in the sequence. This formulation produces a vector of size 25; however, only 16 of these 25 attributes are pertinent, and the remaining 9 consistently sum to 0, lowering model performance. For instance, $\psi_{\gamma\gamma}$ is always 0 because both groups have no charge. Table 1 facilitates selecting attributes from the 25-vector.

Table 1 shows just 16 charged pairings (involving α_1 and α_2), resulting in a total of 16 characteristics. After normalizing the new attribute selection, we set

$$\boldsymbol{\psi}_0 := \frac{1}{N_0}\left[\frac{\psi_{\alpha_1\alpha_1}}{\text{count}(\alpha_1\alpha_1)}, \frac{\psi_{\alpha_1\alpha_2}}{\text{count}(\alpha_1\alpha_2)}, \cdots, \frac{\psi_{\delta\delta}}{\text{count}(\delta\delta)}\right]^T \in \mathbb{R}^{16}.$$

Hence, the feature follows the same process as CKSSCP, incorporating gaps to produce a larger vector. Here, we set

$$\psi_1 := \frac{1}{N_1}\left[\frac{\psi_{\alpha_1x\alpha_1}}{\text{count}(\alpha_1x\alpha_1)}, \frac{\psi_{\alpha_1x\alpha_2}}{\text{count}(\alpha_1x\alpha_2)}, \cdots, \frac{\psi_{\delta x\delta}}{\text{count}(\delta x\delta)}\right]^T \in \mathbb{R}^{16},$$
$$\psi_2 := \frac{1}{N_2}\left[\frac{\psi_{\alpha_1xx\alpha_1}}{\text{count}(\alpha_1xx\alpha_1)}, \frac{\psi_{\alpha_1xx\alpha_2}}{\text{count}(\alpha_1xx\alpha_2)}, \cdots, \frac{\psi_{\delta xx\delta}}{\text{count}(\delta xx\delta)}\right]^T \in \mathbb{R}^{16},$$
$$\Psi_k := [\psi_0, \psi_1, \psi_2, \ldots, \psi_k]^T \in \mathbb{R}^{16(k+1)}.$$

An illustration of the implementation of these equations for $k = 1$ is furnished in Fig. 4.

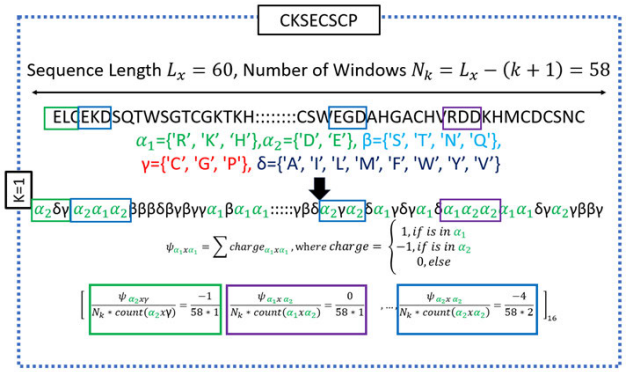


FIGURE 4. Illustration of the CKSECSCP with $k = 1$.

5) ISOELECTRIC POINT FEATURE

In addition to the previously discussed features, the *isoelectric point feature* (pI) is quite important. Amino acids have complex structures and constant parameters [30]. In this feature extraction approach, we use four parameters, *pk(CO2H)*, *pk(NH2)*, *pk(R)*, and the *isoelectric point*. Using these parameters individually results in a huge attribute size, potentially reducing model performance. To this end, we define a combination ψ_x of these parameters as

$$\psi_x := \alpha \times [\text{pk}(CO2H)]_x + \beta \times [\text{pk}(NH2)]_x + \gamma \times [\text{pk}(R)]_x + \delta \times [\text{isoelectric point}]_x.$$

The combination ψ_x represents the final value of the isoelectric point for the amino acid x in the peptide sequence, where α, β, γ, and δ are constants that can be adjusted to fine-tune the model. In this study, α, β, γ, and δ are set to be 1, 2, 4, and 8, respectively. The pI feature vector, Ψ, which contains the relative pI value for each amino acid ψ_x, is constructed as

$$\Psi := \left[\frac{\psi_A \times N_A}{L_x}, \frac{\psi_C \times N_C}{L_x}, \frac{\psi_D \times N_D}{L_x}, \ldots, \frac{\psi_Y \times N_Y}{L_x}\right]^T \in \mathbb{R}^{20},$$

where N_x is the sum of the count of ψ_x. Fig. 5 illustrates an example of the pI feature vector Ψ.

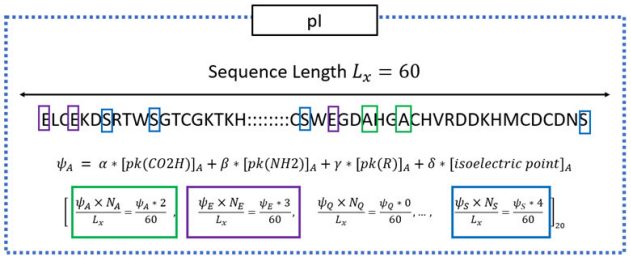


FIGURE 5. Illustration of the calculation for the pI feature.

C. MODEL ARCHITECTURE

Recurrent neural networks (RNNs) have grown in prominence over the past decade. RNNs, like LSTM networks, have found use in a variety of fields, including speech recognition and natural language processing (NLP). Neurons in a typical RNN design are interconnected nodes that collaborate to accomplish complex tasks; nonetheless, RNNs have limitations. When incorporating new information, RNNs often transform current information rather than replace it entirely which may alter the new information in the process. LSTM operates similarly to RNNs but has a key advantage: it can overwrite old information as needed, which is useful in a variety of applications. LSTM operates on the basis of the *forget gate*, the *input gate*, and the *output gate*, each with a specialized role. Initially, the model learns to determine whether the information circulating inside the model should be discarded or retained. The forget gate analyzes two inputs, h_{t-1} and x_t, and decides whether to forget or retain the information if the output equals 0 or 1, respectively. The output of the forget gate is calculated as

$$f_t := \sigma(W_f \times [h_{t-1}, x_t] + b_f),$$

where W_f and b_f are the weights and bias, respectively. Here, h_{t-1} is the output value of the preceding cell, x_t is the input value of the current cell, and

$$\sigma(x) = \frac{1}{1 + e^{-x}},$$

is the *sigmoid* activation function.

The input gate plays a pivotal role in determining the relevance of newly introduced information. This process consists of three sequential steps. The first step functions similarly to the forget gate, acting as a filter for new incoming information. The second step generates a vector containing all potential values that could be written into the state cell. Finally, in the last step, the outputs from the preceding two processes are multiplied and then added to the state cell. Therefore, upon the completion of these steps, the input gate model, described by the equations

$$\begin{aligned} i_t &:= \sigma(W_i \times [h_{t-1}, x_t] + b_i), \\ \tilde{C}_t &:= \tanh(W_c \times [h_{t-1}, x_t] + b_c), \\ C_t &:= f_t \times C_{t-1} + i_t \times \tilde{C}_t, \end{aligned}$$

assures that the input information is indeed beneficial. Here, W_i, W_c, b_i, and b_c, are weights and bias, respectively. This process determines the new value of the cell state as

$$\begin{aligned} o_t &:= \sigma(W_o \times [h_{t-1}, x_t] + b_o), \\ h_t &:= o_t \times \tanh(C_t), \end{aligned}$$

where h_t is the output, and W_o and b_o are the weight and bias of the output function, respectively.

Finally, the output gate implementation can be described in three steps. The first step is to generate a vector to scale the values in the cell state to the range of -1 to $+1$ using the tanh activation function. The second step is to regulate the output values by establishing a filter using h_{t-1} and x_t. The final step is to multiply the output values from steps 1 and 2, resulting in the output value h_t.

The proposed model has 128 neurons in the LSTM layer, followed by a dense layer with a *sigmoid* activation function. To prevent overfitting, a dropout layer with a rate of 0.25 is added following the LSTM layer. During model compilation, binary cross-entropy is used as the loss function alongside the *Adam* optimizer [31].

D. EVALUATION PROTOCOL

The performance of the proposed technique is evaluated using 5-fold and 10-fold cross-validation protocols. The *ACP740* dataset is subjected to a 5-fold cross-validation protocol, while *ACP344* dataset is subjected to both 5-fold and 10-fold cross-validation protocols. The performance of the proposed technique is compared with the state-of-the-art classification methods in terms of sensitivity, specificity, accuracy, balanced accuracy, Matthew's correlation coefficient (MCC), Youden's index, and F1 Score. These metrics are respectively defined as

$$
\begin{aligned}
S_n &:= \frac{TP}{TP+FN},\\
S_p &:= \frac{TN}{TN+FP},\\
\text{Acc.} &:= \frac{TP+TN}{TP+TN+FP+FN},\\
\text{Bal. Acc.} &:= \frac{S_n+S_p}{2},\\
MCC &:= \frac{TPTN-FPFN}{\sqrt{\Delta}},\\
\text{YI} &:= S_n+S_p-1,\\
\text{F1 Score} &:= 2*\frac{\text{Precision}*S_n}{\text{Precision}+S_n},
\end{aligned}
$$

where TP and FP are the true and false positives, TN and FN are the true and false negatives, and

$$
\begin{aligned}
\Delta &:= (TP+FP)(TN+FN)(TP+FN)(TN+FP),\\
\text{Precision} &:= \frac{TP}{TP+FP}.
\end{aligned}
$$

III. EXPERIMENTAL RESULTS

This section describes the experiments carried out to assess the performance of the suggested technique. It also includes a feature space analysis and relevant experiments to evaluate the robustness and effectiveness of the different feature encoding strategies discussed.

A. PERFORMANCE COMPARISION OF FEATURE ENCODING SCHEMES

We conduct an ablation study to evaluate how different feature encoding techniques discussed in the previous section affect the discriminating performance of the proposed model. The proposed model is trained on individual feature encoding techniques for the ablation experiment. The performance of the proposed model is measured, assessed, and compared to the other feature encoding techniques employed in this study. Furthermore, the input feature space for each feature encoding approach is displayed and compared to the proposed combined feature encoding method.

We performed 5-fold cross-validation on both *ACP344* and *ACP740* datasets. Each dataset has seen a total of 6 trials. The proposed model is trained using various feature encoding schemes, including k-mer, BPF, CKSSCP, CKSECSCP, pI, a mixture of CKSSCP and CKSECSCP, and lastly, a combination of all five schemes mentioned previously. The performance evaluation methods outlined in Section II-D are used to compare the results of these experiments.

Table 2 compares the performance of the proposed model to each feature encoding scheme combination on the *ACP344* dataset. Interestingly, the proposed feature encoding strategy has the highest MCC score of 0.85, resulting in greater performance over the other feature encoding strategies. Similarly, the suggested scheme achieves a higher Youden's index of 0.84, compared to 0.83 for the CKSSCP + pI + CKSECSCP features. Since the difference in CKSSCP + pI + CKSECSCP and proposed feature set is very small, we perform statistical tests [32]. The t-statistic is the ratio of the difference in a number's estimated value from its assumed value to its standard error. Higher values of the t-score indicate that a large difference exists between the two sample sets and generally, a t-statistic of 2 or higher and p-value of less than 0.05 is considered to be statistically significant. For the given results we obtain t-statistics of -2.656 and the p-value of 0.008 suggesting a significant improvement. Furthermore, the suggested input feature encoding strategy achieves improved balanced accuracy and sensitivity, which contributes to better feature representation than the listed encoding schemes.

TABLE 2. Comparison of different feature encoding schemes for *ACP344* database.

Features	Sn (%)	Sp (%)	Acc. (%)	Bal. Acc. (%)	MCC	YI	F1-score
K-mer	46.32	95.63	75.86	70.97	0.51	0.42	0.6
BPF	76	86.91	82.55	81.45	0.64	0.63	0.78
CKSSCP	73.2	85.93	80.8	79.56	0.61	0.6	0.76
CKSECSCP	45.74	86.37	70.06	66.05	0.36	0.33	0.55
pI	85.52	94.65	90.98	90.09	0.82	0.81	0.89
CKSSCP+ pI + CKSECSCP	86.29	96.59	92.45	91.44	0.84	0.83	0.91
Proposed	87.01	96.61	92.74	91.81	0.85	0.84	0.91

The experiment using the *ACP740* database shows a similar performance trend. Table 3 shows the results of an ablation study on the *ACP740* database, subjected to 5-fold cross-validation. It is worth noting that the performance of the model trained on the combination of two or more features is promising as compared to the model trained on individual features. For example, while the proposed model performs poorly on individual CKSECSCP features as compared to all other individual feature encoding schemes, incorporating the CKSECSCP feature into the combination of the different encoding schemes contributes to the overall performance improvements of the proposed model. To further validate the difference in CKSSCP + pI + CKSECSCP and proposed feature set, we perform statistical tests and for the

TABLE 3. Comparison of different feature encoding schemes for *ACP740* database.

Features	Sn (%)	Sp (%)	Acc. (%)	Bal. Acc. (%)	MCC	YI	F1-score
K-mer	73.4	67.85	70.67	70.63	0.42	0.41	0.72
BPF	81.92	79.39	80.67	80.66	0.62	0.62	0.82
CKSSCP	80.05	79.12	79.59	79.58	0.6	0.6	0.8
CKSECSCP	67.54	58.5	63.1	63.02	0.27	0.27	0.65
pI	83.25	79.94	81.62	81.59	0.64	0.64	0.83
CKSSCP + pI + CKSECSCP	86.71	86.25	86.48	86.48	0.73	0.73	0.87
Proposed	87.25	87.63	87.43	87.44	0.75	0.75	0.87

given results we obtain t-statistics of -4.28 and p-value of 1.99×10^{-5} suggesting a significant improvement.

We also performed a visualization analysis of the specified feature encoding schemes for the *ACP344* database. Because the input feature space has a large dimension, we use the *t-stochastic neighbor embedding* (tSNE) [33] to generate its 2-dimensional projection. Figure 6 illustrates tSNE scatter plots for different input feature encoding schemes. It is evident from the Figs. 6a, 6b, 6c, and 6d that the respective tSNE projections of k-mer, BPF, CKSSCP, and CKSECSCP have little to no class separation. However, the tSNE projections resulting from the combination of multiple feature encoding schemes show adequate class separability; see Figs. 6f and 6g. Particularly, the feature space shown in Fig. 6f has adequate inter-class separation but also exhibits slightly higher intra-class disparities compared to the proposed feature encoding scheme as depicted in Fig. 6g. The analysis in Fig. 6 aligns with the findings in Table 2.

The input feature space tSNE presented in Fig. 6f and Fig. 6g look somewhat visually similar, however, it is worth noting that these tSNE embeddings present the input space. To ascertain the performance of the proposed method, the tSNE embeddings of the latent space for these two cases was also plotted and presented in Fig. 7. It can be observed that the train and test sets in Fig. 7b are linearly separable while the train and test latent spaces presented in Fig. 7a, although separate, do not follow a similar linear separability compared to the proposed technique.

B. ACP CLASSIFICATION PERFORMANCE EVALUATION

This section describes various experiments carried out to validate our approach. The *ACP740* and *ACP344* datasets [19] have been used to test the proposed method. Each dataset has its own set of parameters, optimized for the best performance. The *ACP344* dataset is evaluated using 5-fold and 10-fold cross-validation protocols. The *ACP740* dataset is evaluated using 5-fold cross-validation. Both CKSSCP and CKSECSCP employed a single gap parameter, resulting in two parameters. On the other hand, α, β, γ, and δ were the parameters used for the *pI* feature. Furthermore, the timestep and the *sigmoid* activation function were used in the deep LSTM algorithm.

Table 4 provides an overview of the hyper-parameters taken into account in the proposed method. Here, n denotes the number of folds in the k-fold cross-validation experiment. The parameters for the CKSSCP and CKSECSCP features are shown in columns *gap1* and *gap2*, respectively. Both datasets maintain the same *timesteps* and epoch values. The batch size varies due to the variation in the sample sizes of the datasets. For both datasets, a sigmoid activation function was employed. Finally, the four parameters in the isoelectric point feature were kept identical for both datasets.

TABLE 4. Summary of the hyper-parameters.

Dataset	n	gap1	gap2	Batch size	Epochs	α	β	γ	δ
ACP344	5	1	5	32	30	1	2	4	8
ACP344	10	1	5	32	30	1	2	4	8
ACP740	5	1	5	64	30	1	2	4	8

1) ACP344 DATESET

The initial experiment on the *ACP344* dataset used a 5-fold cross-validation. Table 5 summarizes the findings. The mean accuracy was determined to be 92.74%, with a standard deviation of 2.23%. Similarly, the mean MCC was calculated to be 85.00%, with a standard deviation of 4.00%.

TABLE 5. 5-fold cross-validation for *ACP344*.

Fold #	Sn (%)	Sp (%)	Acc. (%)	Bal. Acc. (%)	MCC	YI	F1-score
1	85.71	97.56	92.75	91.64	0.85	0.83	0.91
2	92.86	95.12	94.20	93.99	0.88	0.88	0.93
3	75.00	97.56	88.40	86.30	0.76	0.73	0.84
4	92.60	95.24	94.20	94.00	0.88	0.88	0.93
5	88.89	97.56	94.12	93.22	0.88	0.86	0.92
Mean	87.01	96.61	92.74	91.81	0.85	0.84	0.91
Std	6.55	1.17	2.23	2.89	0.04	0.05	0.03

In Table 6, the results of the 5-fold cross-validation experiment are compared with other contemporary techniques. The performance of the proposed method was better or comparable to the listed techniques. In terms of accuracy, the proposed method performs better than the *ZH* method [11], the *AntiCP(model2)* [34], and the *ACP-KSRC* [19] with margins of 3.32%, 6.72%, and 3.02%, respectively. Similarly, in terms of the MCC, the proposed method performs better than the *ZH* method, *AntiCP*(model2), and *ACP-KSRC* with margins of 8.97%, 19.72%, and 4.94%, respectively. It is noteworthy, nevertheless, that the *Li* method [35], the *IACP* [35], and the *EnACP* [35] performed marginally better than the proposed method in terms of the MCC.

TABLE 6. Comparison of the proposed method with contemporary methods on *ACP344* dataset using 5-fold cross-validation.

Methods	Sn (%)	Sp (%)	Acc. (%)	Bal. Acc. (%)	MCC	YI	F1-score	Year
ZH Method [11], [35]	85.23	92.73	89.76	88.00	0.78	0.77	0.89	2014
Li Method [35], [22]	90.60	96.70	94.25	93.05	0.87	0.87	0.93	2016
IACP [35], [36]	88.00	89.10	94.80	93.05	0.89	0.87	0.91	2016
EnACP [35]	92.20	98.10	95.41	95.00	0.91	0.90	0.94	2020
AntiCP(model2) [34]	81.32	90.02	86.90	85.07	0.71	0.71	0.82	2021
ACP-KSRC [19]	96.07	82.97	90.02	89.10	0.81	0.79	0.92	2023
Proposed	87.01	96.61	92.74	91.81	0.85	0.84	0.91	2024

Similarly, *ACP344* dataset was also subjected to a 10-fold cross-validation, and the specific outcomes are presented in Table. 7. In the 10-fold cross-validation experiment, the

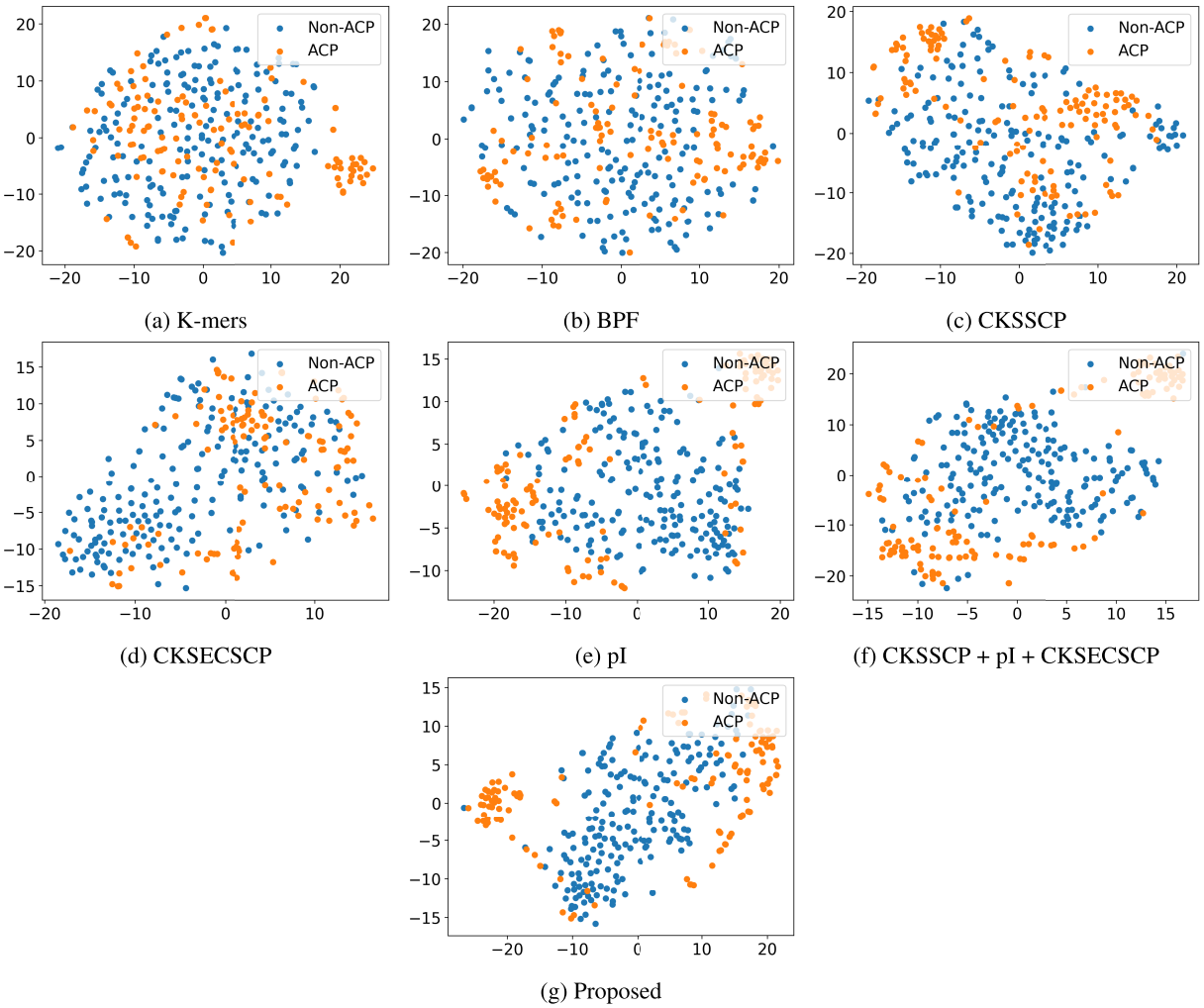


FIGURE 6. Feature embedding visualization of *ACP344* dataset using (a) *k*-mers, (b) binary-profile features (BPF), (c) CKSSCP, (d) CKSECSCP, (e) pI, (f) CKSSCP + pI+ CKSECSCP, and (g) the proposed feature set.

TABLE 7. 10-fold cross-validation for *ACP344*.

Fold #	Sn (%)	Sp (%)	Acc. (%)	Bal. Acc. (%)	MCC	YI	F1-score
1	92.86	100	97.14	96.42	0.94	0.93	0.96
2	92.86	90.5	91.43	91.67	0.83	0.83	0.90
3	85.71	100	94.29	92.86	0.88	0.86	0.92
4	78.57	95.24	88.57	86.90	0.76	0.74	0.85
5	92.86	90.00	91.18	91.43	0.82	0.83	0.90
6	85.71	95.00	91.18	90.36	0.82	0.81	0.89
7	92.86	100	97.06	96.43	0.94	0.93	0.96
8	71.43	100	88.23	85.71	0.77	0.71	0.83
9	92.31	95.24	94.12	93.77	0.87	0.87	0.92
10	92.31	100	97.06	96.15	0.94	0.92	0.96
Mean	87.75	96.6	93.02	92.17	0.86	0.84	0.91
Std	7.10	3.81	3.23	3.59	0.06	0.07	0.04

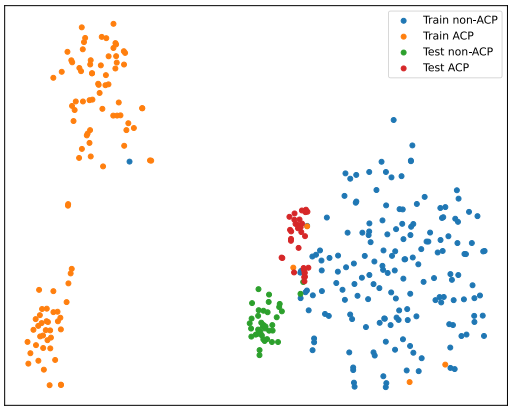
suggested approach outperformed the 5-fold cross-validation experiment.

A comparison of the 10-fold cross-validation experiment results for the proposed and state-of-the-art approaches is shown in Table 8. It is observed that the suggested approach performs better than existing methods in terms of the MCC and that its accuracy and Youden's index results are at par with the approach outlined in [19].

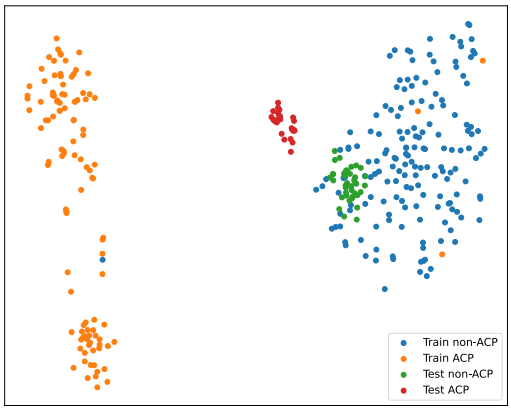
TABLE 8. Comparison of the proposed method with contemporary methods on *ACP344* dataset using 10-fold cross-validation.

Methods	Sn (%)	Sp (%)	Acc. (%)	Bal. Acc. (%)	MCC	YI	F1-score	Year
SAP [37]	86.23	95.63	91.86	90.93	0.83	0.81	0.89	2020
ACP-LDF (SVM) [38]	87.70	96.10	92.73	91.90	0.84	0.83	0.92	2020
ACP-LDF (LibD3C) [38]	85.50	96.10	92.15	91.05	0.83	0.82	0.92	2020
ACP-LDF (RF) [38]	86.20	97.10	92.70	91.65	0.84	0.83	0.92	2020
ACP-DL [16]	75.82	86.32	82.16	81.07	0.62	0.62	0.77	2022
ACP-KSRC [19]	97.07	86.97	93.02	91.89	0.85	0.84	0.94	2023
Proposed	87.75	96.60	93.02	92.17	0.86	0.84	0.91	2024

Since the ACP detection is primarily a classification problem, the area under the curve (AUC) of the receiver operating characteristics (ROC) curve can also be used to evaluate the performance of the classification model. AUC serves as a comprehensive metric to evaluate the overall classification performance. Figs. 8a and 8b presents

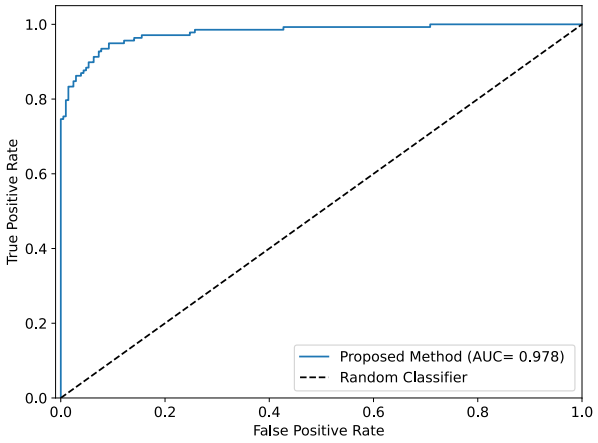


(a) CKSSCP + pI + CKSECSCP

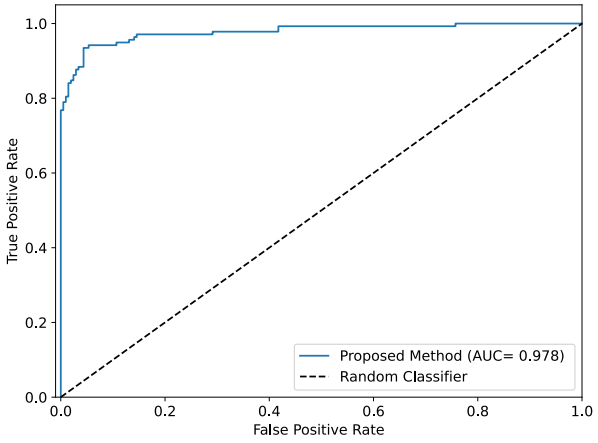


(b) Proposed

FIGURE 7. Latent-space feature embedding visualization of different classifiers trained on *ACP344* dataset using (a) CKSSCP + pI+ CKSECSCP, and (b) the proposed feature set.



(a) 5-fold cross-validation



(b) 10-fold cross-validation

FIGURE 8. Receiver operating characteristics curves for (a) 5-fold and (b) 10-fold cross-validation on *ACP344* dataset.

the AUC-ROC curves and AUC scores for the 5-fold and 10-fold cross-validation experiments carried out on *ACP344* dataset. Interestingly, both the experiments achieved the same AUC score of 97.8%. This may be due to the reason that both the experiments achieved almost the same average performance in-terms of sensitivity and specificity as listed in Tables 5 and 7.

2) ACP740 DATASET

Using 5-fold cross-validation, the suggested approach is verified on the *ACP740* dataset, by recording every fold in the study. The complete experimental findings are shown in Table 9, which include mean and standard deviation for each of the seven assessment criteria: sensitivity, specificity, accuracy, balanced accuracy, MCC, Youden's index, and F1 Score. In particular, the mean accuracy was 87.25% on average, with a standard deviation of 5.83%, and the MCC value was 75.00% on average, with a 6% standard deviation.

Table 10 compares the performance of the proposed method with the state-of-the-art techniques on the *ACP740* dataset. With a *sensitivity* score S_n that is just 1.89% lower than ACP-MHCNN [20], the proposed method had the second-best score. However, it outperformed the listed methods on other evaluation criteria. In particular, the accuracy of the proposed method was superior to that of the ACP-DL [16], ACP-DA [39], ACP-MHCNN [20], and ACP-KSRC [19] by margins of 7.3%, 6.58%, 1.66%, and 4.19%, respectively. Likewise, the suggested approach demonstrated superior MCC performance compared to the ACP-DL, ACP-DA, ACP-MHCNN, and ACP-KSRC, by margins of 19.05%, 17.19%, 4.16%, and 11.94%, respectively. Table 10 compares the ACP prediction methods across all 7 evaluation criteria.

TABLE 9. 5-fold cross-validation for *ACP740*.

Fold #	Sn (%)	Sp (%)	Acc. (%)	Bal. Acc. (%)	MCC	YI	F1-score
1	81.59	83.33	82.43	82.46	0.65	0.65	0.83
2	89.33	87.70	88.51	88.50	0.77	0.77	0.89
3	89.33	89.04	89.19	89.19	0.78	0.78	0.89
4	80.00	90.41	85.13	85.21	0.71	0.70	0.85
5	96.00	87.67	91.90	91.83	0.84	0.84	0.92
Mean	87.25	87.63	87.43	87.44	0.75	0.75	0.87
Std	5.83	2.37	3.30	3.26	0.06	0.06	0.03

TABLE 10. Comparison of the proposed method with contemporary methods on *ACP740* dataset using 5-fold cross-validation.

Methods	Sn (%)	Sp (%)	Acc. (%)	Bal. Acc. (%)	MCC	YI	F1-score	Year
ACP-DL [16]	82.61	80.59	81.48	83.30	0.63	0.62	0.71	2019
ACP-DA [39]	86.98	83.26	82.03	85.12	0.64	0.70	0.85	2021
ACP-MHCNN [20]	88.90	83.10	86.00	86.00	0.72	0.71	0.86	2021
ACP-KSRC [19]	86.23	81.62	83.91	83.94	0.67	0.67	0.84	2023
Proposed	87.25	87.63	87.43	87.44	0.75	0.75	0.87	2024

We also present the AUC-ROC curve for the 5-fold cross-validation experiment conducted on *ACP740* dataset. The AUC-ROC curve is presented in Fig. 9, where the proposed method achieved an AUC score of 94% for 5-fold cross-validation experiment on *ACP740* dataset.

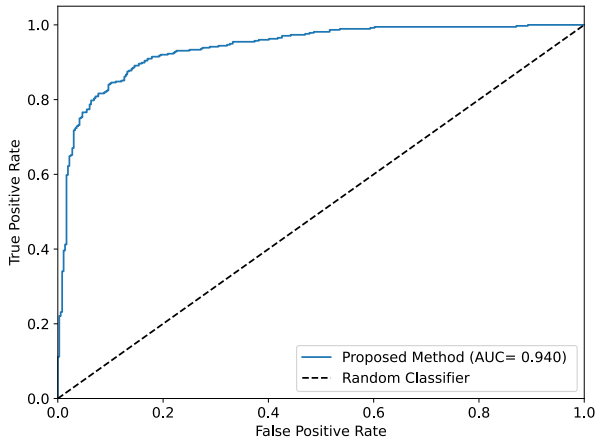


FIGURE 9. Receiver operating characteristics curves for 10-fold cross-validation on *ACP740* dataset.

C. ABLATION STUDY

1) COMPARISON WITH BASELINE CLASSIFIERS

In this section, we compare the performance of the newly designed feature set using the proposed LSTM-based model, and several baseline machine learning and deep learning classifiers. The comparison is performed on the *ACP344* and *ACP740* datasets, using both 5-fold and 10-fold cross-validation experiments. The baseline models include Random Forest [40], Support Vector Machine (SVM) [41], Logistic Regression [42], Multilayer Perceptron (MLP) [43], Naïve Bayes [44], K-Nearest Neighbors (KNN) [45], and Decision Tree [46]. The detailed metrics, including Sensitivity (Sn), Specificity (Sp), Accuracy (Acc), Balanced Accuracy (Bal. Acc), Matthews Correlation Coefficient (MCC), Youden's Index (YI), F1-score, and Area Under the Curve (AUC), are presented in Tables 11, 12, and 13.

The proposed method significantly outperforms the baseline classifiers across several key metrics. For instance, on the *ACP344* dataset, the proposed LSTM-based method achieves the highest AUC of 0.978 in both 10-fold and 5-fold cross-validation experiments, demonstrating superior discriminative ability compared to other models like Random Forest (AUC: 0.969) and SVM (AUC: 0.974). The proposed method also excels in terms of Accuracy, with values of 93.02% (in 10-fold) and 92.71% (in 5-fold), outperforming other models such as the MLP (92.73%) and Logistic Regression (88.66%).

The MCC score for the proposed method also indicates strong classification performance, particularly when compared to baseline classifiers. For instance, the MCC reaches 0.86 in the 10-fold cross-validation on the *ACP344* dataset, outperforming models such as SVM (0.828) and KNN (0.804). These improvements across multiple metrics suggest that the integration of multiple feature encoding schemes within the LSTM architecture provides a robust approach for anticancer peptide classification, enhancing both sensitivity and specificity.

TABLE 11. Comparison with baseline machine learning and deep learning classifiers on the *ACP344* dataset using a 10-fold cross-validation experiment.

Model	Sn (%)	Sp (%)	Acc. (%)	Bal. Acc. (%)	MCC	YI	F1-score	AUC
Random Forest	78.96	97.57	90.12	88.26	0.80	0.77	0.86	0.97
SVM	86.87	94.67	91.56	90.77	0.83	0.82	0.89	0.97
Logistic Regression	83.19	92.21	88.66	87.70	0.77	0.75	0.85	0.94
MLP	86.98	96.60	92.73	91.79	0.85	0.84	0.90	0.97
Naïve Bayes	80.27	78.55	79.29	79.41	0.58	0.59	0.76	0.80
KNN	82.75	95.64	90.42	89.20	0.80	0.78	0.87	0.95
Decision Tree	81.21	86.33	84.28	83.77	0.68	0.68	0.81	0.84
Proposed Method	87.75	96.60	93.02	92.17	0.86	0.84	0.91	0.98

TABLE 12. Comparison with baseline machine learning and deep learning classifiers on the *ACP344* dataset using a 5-fold cross-validation experiment.

Model	Sn (%)	Sp (%)	Acc. (%)	Bal. Acc. (%)	MCC	YI	F1-score	AUC
Random Forest	78.17	98.06	90.11	88.12	0.80	0.76	0.86	0.97
SVM	87.59	95.62	92.44	91.61	0.84	0.83	0.90	0.98
Logistic Regression	82.46	90.75	87.49	86.61	0.74	0.73	0.84	0.94
MLP	83.97	96.59	91.57	90.28	0.83	0.81	0.89	0.97
Naïve Bayes	78.15	76.70	77.32	77.42	0.54	0.55	0.73	0.79
KNN	81.85	96.12	90.41	88.99	0.80	0.78	0.87	0.95
Decision Tree	76.16	84.48	81.13	80.32	0.61	0.61	0.76	0.80
Proposed Method	87.01	96.61	92.71	91.81	0.85	0.84	0.91	0.98

TABLE 13. Comparison with baseline machine learning and deep learning classifiers on the *ACP740* dataset using a 5-fold cross-validation experiment.

Model	Sn (%)	Sp (%)	Acc (%)	Bal. Acc (%)	MCC	YI	F1	AUC
Random Forest	79.78	87.91	83.78	83.85	0.68	0.68	0.83	0.92
SVM	85.64	84.07	84.86	84.85	0.70	0.70	0.85	0.92
Logistic Regression	85.37	84.61	85.00	84.99	0.70	0.70	0.85	0.91
MLP	86.17	82.98	84.59	84.57	0.69	0.69	0.85	0.93
Naïve Bayes	69.14	83.52	76.22	76.33	0.53	0.53	0.75	0.82
KNN	84.31	84.90	84.59	84.61	0.69	0.69	0.85	0.92
Decision Tree	74.46	67.56	71.08	71.01	0.42	0.42	0.72	0.71
Proposed Method	87.25	87.63	87.43	87.44	0.75	0.75	0.87	0.94

2) ANALYZING THE EFFECT OF LARGE NEGATIVE SET ON THE PERFORMANCE OF THE PROPOSED TECHNIQUE

In this section, we analyze the effect of increasing the number of negative samples on the performance of the proposed LSTM-based method. The experiments are conducted using

TABLE 14. Effect of negative samples on the performance of the proposed technique. The experiment is performed using positive sequences from the *ACP740* dataset and negative samples from mACPpred 2.0 [47] (Without SMOTE).

Ratio	Sn (%)	Sp (%)	Acc (%)	Bal. Acc (%)	MCC	YI	F1-score	AUC
1:1	86.43	83.24	84.85	84.84	0.70	0.70	0.85	0.88
1:2	84.30	92.83	89.99	88.57	0.78	0.77	0.85	0.92
1:3	70.78	93.16	87.56	81.97	0.67	0.64	0.74	0.82
1:4	65.21	94.21	88.40	79.71	0.63	0.59	0.69	0.81
1:5	64.68	94.73	89.72	79.71	0.63	0.59	0.68	0.83
1:6	61.49	94.90	90.12	78.19	0.61	0.56	0.65	0.83
1:7	62.81	95.97	91.82	79.39	0.63	0.59	0.66	0.85
1:8	63.09	96.24	92.55	79.67	0.62	0.59	0.65	0.84
1:9	58.56	95.04	91.38	76.80	0.57	0.54	0.60	0.82
1:10	61.49	96.44	93.25	78.97	0.62	0.58	0.64	0.83

TABLE 15. Effect of negative samples on the performance of the proposed technique. The experiment is performed using positive sequences from the *ACP740* dataset and negative samples from mACPpred 2.0 [47] (With SMOTE).

Ratio	Sn (%)	Sp (%)	Acc (%)	Bal. Acc (%)	MCC	YI	F1-score	AUC
1:1	84.57	83.25	83.91	83.91	0.68	0.68	0.84	0.88
1:2	83.51	91.37	88.75	87.44	0.75	0.75	0.84	0.90
1:3	75.02	90.86	86.90	82.94	0.66	0.66	0.75	0.82
1:4	72.14	91.82	87.87	81.98	0.64	0.64	0.70	0.86
1:5	69.74	93.30	89.36	81.52	0.64	0.63	0.69	0.88
1:6	71.34	93.35	90.20	82.34	0.64	0.65	0.69	0.91
1:7	67.34	95.44	91.92	81.39	0.64	0.63	0.68	0.88
1:8	65.74	95.11	91.84	80.43	0.60	0.61	0.65	0.83
1:9	62.81	94.47	91.30	78.64	0.58	0.57	0.61	0.85
1:10	66.28	95.40	92.75	80.84	0.62	0.62	0.65	0.87

376 positive sequences from the *ACP740* dataset and 3760 negative samples from (training and independent)-sets of mACPpred 2.0 [47]. The results are shown in Tables 14 and 15, where we examine how different negative-to-positive ratios influence the model's sensitivity, specificity, and other key metrics.

As the negative-to-positive ratio increases, the sensitivity (Sn) decreases, indicating that the model becomes more inclined to classify negative samples correctly at the expense of positive samples. For example, with a 1:1 ratio, sensitivity is 0.864, but it drops to 0.585 when the negative ratio increases to 1:9. However, specificity (Sp) improves significantly as the model becomes more effective at identifying negative samples, rising from 0.832 to 0.950 as the negative ratio increases.

To handle this imbalance between positive and negative samples, we employed the SVM-SMOTE technique [48], [49], which synthetically generates positive samples to balance the dataset. After applying SVM-SMOTE, the sensitivity improves, particularly for higher negative-to-positive ratios. For instance, at a 1:9 ratio, the sensitivity increases from 0.585 (without SMOTE) to 0.628 (with SMOTE), and the AUC shows corresponding improvements, reflecting the model's enhanced ability to generalize across imbalanced data. The use of SVM-SMOTE also stabilizes the balanced accuracy (Bal. Acc), preventing the model from becoming overly biased towards the negative class. However, model is still biased towards negative samples indicating the need for more ground truth positive samples.

IV. CONCLUSION

In this study, we proposed a novel long-short-term-memory (LSTM)-based anticancer peptide (ACP) classification technique in which we integrated multiple feature encoding schemes to enhance the discriminative power of the model. We designed a strong feature representation by combining the advantages of several feature encoding approaches, including k-mer, BPF, CKSSCP, CKSECSCP, and pI. We demonstrated the efficacy of the proposed method using extensive experimentation and ablation studies.

The experimental findings on the *ACP344* and *ACP740* datasets demonstrated that the proposed method works better than individual feature encoding schemes. In particular, it achieved better accuracy, sensitivity, and Matthews correlation coefficient (MCC) than other modern techniques. The performance improvements were particularly notable when multiple feature encoding schemes were combined, highlighting the importance of feature integration in the ACP classification tasks. The results of the combined feature encoding strategy in improving class separability and lowering intra-class variability were further validated by the *t-stochastic neighbor embedding* (tSNE) visualizations of the feature space. Furthermore, 5-fold and 10-fold cross-validation protocols were used to assess the robustness of the proposed approach. The results showed consistent and dependable performance of the proposed method across several validation configurations, confirming its generalizability.

As a future direction, we aim to investigate the use of Transformers, such as BioBERT or protein-specific pretrained models, for ACP classification. A comparative analysis with LSTM in terms of accuracy, scalability, and computational cost will provide further insights into the trade-offs between these architectures.

In summary, the performance of the ACP classification models can be significantly improved by the integration of different feature encoding approaches. The proposed method for ACP classification is promising and has potential applications in the broader field of peptide-based therapeutics. To further enhance performance, future research could investigate applying this technique to different peptide classification applications and include different feature encoding approaches.

REFERENCES

[1] S. Basith, M. Cui, S. J. Y. Macalino, and S. Choi, "Expediting the design, discovery and development of anticancer drugs using computational approaches," *Current Medicinal Chem.*, vol. 24, no. 42, pp. 4753–4778, Jan. 2018.

[2] J.-X. Su, S.-J. Li, X.-F. Zhou, Z.-J. Zhang, Y. Yan, S.-L. Liu, and Q. Qi, "Chemotherapy-induced metastasis: Molecular mechanisms and clinical therapies," *Acta Pharmacologica Sinica*, vol. 44, no. 9, pp. 1725–1736, Sep. 2023.

[3] I. Kaur, M. N. Doja, and T. Ahmad, "Data mining and machine learning in cancer survival research: An overview and future recommendations," *J. Biomed. Informat.*, vol. 128, Apr. 2022, Art. no. 104026.
[4] A. F. G. Cicero, F. Fogacci, and A. Colletti, "Potential role of bioactive peptides in prevention and treatment of chronic diseases: A narrative review," *Brit. J. Pharmacol.*, vol. 174, no. 11, pp. 1378–1394, Jun. 2017.
[5] M. Muttenthaler, G. F. King, D. J. Adams, and P. F. Alewood, "Trends in peptide drug discovery," *Nature Rev. Drug Discovery*, vol. 20, no. 4, pp. 309–325, Apr. 2021.
[6] T.-B. Zou, T.-P. He, H.-B. Li, H.-W. Tang, and E.-Q. Xia, "The structure-activity relationship of the antioxidant peptides from natural proteins," *Molecules*, vol. 21, no. 1, p. 72, Jan. 2016.
[7] A. Jafari, A. Babajani, R. Sarrami Forooshani, M. Yazdani, and M. Rezaei-Tavirani, "Clinical applications and anticancer effects of antimicrobial peptides: From bench to bedside," *Frontiers Oncol.*, vol. 12, Feb. 2022, Art. no. 819563.
[8] Y. Liscano, J. Oñate-Garzón, and J. P. Delgado, "Peptides with dual antimicrobial-anticancer activity: Strategies to overcome peptide limitations and rational design of anticancer peptides," *Molecules*, vol. 25, no. 18, p. 4245, Sep. 2020.
[9] N. Shao, L. Yuan, P. Ma, M. Zhou, X. Xiao, Z. Cong, Y. Wu, G. Xiao, J. Fei, and R. Liu, "Heterochiral β-peptide polymers combating multidrug-resistant cancers effectively without inducing drug resistance," *J. Amer. Chem. Soc.*, vol. 144, no. 16, pp. 7283–7294, Apr. 2022.
[10] X. Xu, C. Li, X. Yuan, Q. Zhang, Y. Liu, Y. Zhu, and T. Chen, "ACP-DRL: An anticancer peptides recognition method based on deep representation learning," *Frontiers Genet.*, vol. 15, Apr. 2024, Art. no. 1376486.
[11] Z. Hajisharifi, M. Piryaiee, M. Mohammad Beigi, M. Behbahani, and H. Mohabatkar, "Predicting anticancer peptides with Chouas pseudo amino acid composition and investigating their mutagenicity via Ames test," *J. Theor. Biol.*, vol. 341, pp. 34–40, Jan. 2014.
[12] M. Usman, S. Khan, S. Park, and A. Wahab, "AFP-SRC: Identification of antifreeze proteins using sparse representation classifier," *Neural Comput. Appl.*, vol. 34, no. 3, pp. 2275–2285, Feb. 2022.
[13] S. Khan, I. Naseem, R. Togneri, and M. Bennamoun, "RAFP-pred: Robust prediction of antifreeze proteins using localized analysis of n-Peptide compositions," *IEEE/ACM Trans. Comput. Biol. Bioinf.*, vol. 15, no. 1, pp. 244–250, Jan. 2018.
[14] M. Usman, S. Khan, and J.-A. Lee, "AFP-LSE: Antifreeze proteins prediction using latent space encoding of composition of k-spaced amino acid pairs," *Sci. Rep.*, vol. 10, no. 1, p. 7197, Apr. 2020.
[15] M. Usman, S. Khan, S. Park, and J.-A. Lee, "AoP-LSE: Antioxidant proteins classification using deep latent space encoding of sequence features," *Current Issues Mol. Biol.*, vol. 43, no. 3, pp. 1489–1501, Oct. 2021.
[16] H.-C. Yi, Z.-H. You, X. Zhou, L. Cheng, X. Li, T.-H. Jiang, and Z.-H. Chen, "ACP-DL: A deep learning long short-term memory model to predict anticancer peptides using high-efficiency feature representation," *Mol. Therapy Nucleic Acids*, vol. 17, pp. 1–9, Sep. 2019.
[17] S. Park, S. Khan, and A. Wahab, "E3-targetPred: Prediction of E3-target proteins using deep latent space encoding," 2020, *arXiv:2007.12073*.
[18] U. M. Al-Saggaf, M. Usman, I. Naseem, M. Moinuddin, A. A. Jiman, M. U. Alsaggaf, H. K. Alshoubaki, and S. Khan, "ECM-LSE: Prediction of extracellular matrix proteins using deep latent space encoding of k-Spaced amino acid pairs," *Frontiers Bioeng. Biotechnol.*, vol. 9, Oct. 2021, Art. no. 752658.
[19] E. Fazal, M. S. Ibrahim, S. Park, I. Naseem, and A. Wahab, "Anticancer peptides classification using kernel sparse representation classifier," *IEEE Access*, vol. 11, pp. 17626–17637, 2023.
[20] S. Ahmed, R. Muhammod, Z. H. Khan, S. Adilina, A. Sharma, S. Shatabda, and A. Dehzangi, "ACP-MHCNN: An accurate multi-headed deep-convolutional neural network to predict anticancer peptides," *Sci. Rep.*, vol. 11, no. 1, p. 23676, Dec. 2021.
[21] I. Naseem, S. Khan, R. Togneri, and M. Bennamoun, "ECMSRC: A sparse learning approach for the prediction of extracellular matrix proteins," *Current Bioinf.*, vol. 12, no. 4, pp. 361–368, Jul. 2017.
[22] W. Chen, H. Ding, P. Feng, H. Lin, and K.-C. Chou, "IACP: A sequence-based tool for identifying anticancer peptides," *Oncotarget*, vol. 7, no. 13, pp. 16895–16909, Mar. 2016.
[23] L. Wei, C. Zhou, H. Chen, J. Song, and R. Su, "ACPred-FL: A sequence-based predictor using effective feature representation to improve the prediction of anti-cancer peptides," *Bioinformatics*, vol. 34, no. 23, pp. 4007–4016, Dec. 2018.
[24] A. Tyagi, P. Kapoor, R. Kumar, K. Chaudhary, A. Gautam, and G. P. S. Raghava, "In silico models for designing and discovering novel anticancer peptides," *Sci. Rep.*, vol. 3, no. 1, pp. 1–8, Oct. 2013.
[25] V. Suresh, L. Liu, D. Adjeroh, and X. Zhou, "RPI-pred: Predicting ncRNA-protein interaction using sequence and structural information," *Nucleic Acids Res.*, vol. 43, no. 3, pp. 1370–1379, Feb. 2015.
[26] X. Pan, Y.-X. Fan, J. Yan, and H.-B. Shen, "IPMiner: Hidden ncRNA-protein interaction sequential pattern mining with stacked autoencoder for accurate computational prediction," *BMC Genomics*, vol. 17, no. 1, pp. 1–14, Dec. 2016.
[27] T. G. Kolda and D. P. O'Leary, "A semidiscrete matrix decomposition for latent semantic indexing information retrieval," *ACM Trans. Inf. Syst.*, vol. 16, no. 4, pp. 322–346, Oct. 1998.
[28] P.-A. Binz, "Proteomics standards initiative extended FASTA format," *J. Proteome Res.*, vol. 18, no. 6, pp. 2686–2692, Jun. 2019.
[29] D. Steward. (2019). *Essential Amino Acids: Chart, Abbreviations and Structure*. [Online]. Available: http://www.technologynetworks.com/appliedsciences/articles/essential-amino-acids-chart-abbreviations-and-structure-324357
[30] D. R. Lide, *CRC Handbook Chemistry and Physics*, 85 th ed. Boca Raton, FL, USA: CRC Press, Jun. 2004, p. 2656.
[31] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2014, *arXiv:1412.6980*.
[32] T. Dahiru, "P-value, a true test of statistical significance? A cautionary note," *Ann. Ibadan Postgraduate Med.*, vol. 6, no. 1, pp. 21–26, Mar. 2011.
[33] L. V. D. Maaten and G. E. Hinton, "Visualizing data using t-SNE," *J. Mach. Learn. Res.*, vol. 9, no. 86, pp. 2579–2605, Jan. 2008.
[34] B. Manavalan, S. Basith, T. H. Shin, S. Choi, M. O. Kim, and G. Lee, "MLACP: Machine-learning-based prediction of anticancer peptides," *Oncotarget*, vol. 8, no. 44, pp. 77121–77136, Sep. 2017.
[35] R. Ge, G. Feng, X. Jing, R. Zhang, P. Wang, and Q. Wu, "EnACP: An ensemble learning model for identification of anticancer peptides," *Frontiers Genet.*, vol. 11, p. 760, Jul. 2020.
[36] S. Vijayakumar and L. Ptv, "ACPP: A web server for prediction and design of anti-cancer peptides," *Int. J. Peptide Res. Therapeutics*, vol. 21, no. 1, pp. 99–106, Mar. 2015.
[37] L. Xu, G. Liang, L. Wang, and C. Liao, "A novel hybrid sequence-based model for identifying anticancer peptides," *Genes*, vol. 9, no. 3, p. 158, Mar. 2018.
[38] Q. Li, W. Zhou, D. Wang, S. Wang, and Q. Li, "Prediction of anticancer peptides using a low-dimensional feature model," *Frontiers Bioeng. Biotechnol.*, vol. 8, p. 892, Aug. 2020.
[39] X.-G. Chen, W. Zhang, X. Yang, C. Li, and H. Chen, "ACP-DA: Improving the prediction of anticancer peptides using data augmentation," *Frontiers Genet.*, vol. 12, Jun. 2021, Art. no. 698477.
[40] A. Cutler, D. R. Cutler, and J. R. Stevens, "Random forests," *Mach. Learn.*, vol. 45, pp. 157–175, Jan. 2012.
[41] C. Cortes and V. Vapnik, "Support-vector networks," *Mach. Learn.*, vol. 20, no. 3, pp. 273–297, Sep. 1995.
[42] D. W. Hosmer, S. Lemeshow, and R. X. Sturdivant, *Applied Logistic Regression*, vol. 398. Hoboken, NJ, USA: Wiley, 2013.
[43] G. E. Hinton, "Connectionist learning procedures," in *Machine Learning*, vol. 40. Amsterdam, The Netherlands: Elsevier, 1989, pp. 185–234.
[44] I. Rish, "An empirical study of the naive Bayes classifier," in *Proc. IJCAI Workshop Empirical Methods Artif. Intell.*, Jan. 2001, pp. 41–46.
[45] T. M. Cover and P. E. Hart, "Nearest neighbor pattern classification," *IEEE Trans. Inf. Theory*, vols. IT-13, no. 1, pp. 21–27, Jan. 1967.
[46] J. R. Quinlan, "Induction of decision trees," *Mach. Learn.*, vol. 1, no. 1, pp. 81–106, Mar. 1986.
[47] V. K. Sangaraju, N. T. Pham, L. Wei, X. Yu, and B. Manavalan, "MACPpred 2.0: Stacked deep learning for anticancer peptide prediction with integrated spatial and probabilistic feature representations," *J. Mol. Biol.*, vol. 436, no. 17, Sep. 2024, Art. no. 168687.
[48] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: Synthetic minority over-sampling technique," *J. Artif. Intell. Res.*, vol. 16, pp. 321–357, Jun. 2002.
[49] H. M. Nguyen, E. W. Cooper, and K. Kamei, "Borderline over-sampling for imbalanced data classification," *Int. J. Knowl. Eng. Soft Data Paradigms*, vol. 3, no. 1, p. 4, 2011.

NAZER AL TAHIFAH received the degree in computer engineering from King Fahd University of Petroleum & Minerals (KFUPM). He was a Machine Learning Research Assistant with the SDAIA-KFUPM Joint Research Center for Artificial Intelligence (JRC-AI). In that role, he focused on creating classification models by utilizing advanced computational methods, generative modeling, and natural language processing. He has a keen interest in biomedical research, driving his passion for innovative solutions in this field.

MUHAMMAD SOHAIL IBRAHIM received the B.E. degree in electronic engineering from Iqra University, Karachi, Pakistan, in 2012, the M.E. degree in telecommunications from N. E. D. University of Engineering and Technology, Pakistan, in 2016, and the Ph.D. degree in computer engineering from Chosun University, South Korea, in 2024. From 2013 to 2019, he was a Lecturer with the Faculty of Engineering, Science, and Technology, Iqra University. From 2013 to 2014, he also was a Research Assistant with the Embedded Systems Research Group, Karachi Institute of Economics and Technology, Pakistan. He is currently affiliated with the Microfluidics and Nano-Bio Systems Laboratory, Department of Mechanical Systems Engineering, Kumoh National Institute of Technology, South Korea. His research interests include deep learning, computer vision, bioinformatics, and medical image processing. He was honored with the 2020 Highly Cited Review Paper award from Applied Energy, Elsevier, for his paper titled "Machine Learning Driven Smart Electric Systems: Current Trends and New Perspectives."

ERUM REHMAN received the Ph.D. degree in statistics from the Southwestern University of Finance and Economics, China, in 2020, and the Ph.D. degree in applied economics from Shandong University of Finance and Economics, China, in 2022. She is currently an Assistant Professor with Nazarbayev University, Kazakhstan. She has published extensively in leading journals, including *Frontiers in Public Health*, *Psychology Research and Behavior Management*, and *Journal of Pharmaceutical Policy and Practice*. Her research on the psychological dynamics of pandemic resilience, the socio-economic influences on mental health, and the correlation between self-compassion and emotional regulation has been widely recognized. She was a recipient of research grants from Nazarbayev University, supporting her work on sustainable development and healthcare policy. Her research interests include the epidemiology of communicable and non-communicable diseases, with a focus on cardiovascular health, breast cancer, and mental disorders.

NAVEED AHMED received the Ph.D. degree from Otto-von-Guerke University, Magdeburg, Germany, in 2011. He has been an Associate Professor with the Gulf University for Science and Technology (GUST), since November 2023. Prior to this role, he was an Assistant Professor with GUST, from 2019 to 2023, and Lahore University of Management Sciences (LUMS), from 2018 to 2019. From 2013 to 2018, he was a Research Associate with the Weierstrass Institute for Applied Analysis and Stochastics (WIAS), Berlin, Germany. He was a Postdoctoral Researcher with the University of Kassel, Germany, from 2011 to 2013. Since January 2024, he has been leading the Center for Applied Mathematics and Bioinformatics (CAMB). He has published numerous articles in prestigious journals, such as *SIAM Journal on Numerical Analysis*, *IMA Journal of Numerical Analysis*, *Computer Methods in Applied Mechanics and Engineering*, and *Journal of the Acoustical Society of America*. His research interests include numerical analysis, scientific computing, stabilized finite element methods for convection-dominated problems, turbulent flows, and machine learning.

ABDUL WAHAB received the M.S. degree from Université Sorbonne (Pierre et Marie Curie Campus), Paris, France, in 2009, and the Ph.D. degree from École Polytechnique Paris, France, in 2011. Since 2024, he has been an Associate Professor with Sultan Qaboos University, Muscat, Oman. Prior to this, he was an Associate Professor with Nazarbayev University, Astana, Kazakhstan, from 2020 to 2024. He also was a Research Professor with Korea Advanced Institute of Sciences and Technology (KAIST), Daejeon, South Korea, from 2016 to 2018, and as a Postdoctoral Researcher with Université Denis Diderot, in 2012. In 2015, he was awarded Korea Research Fellowship by the National Research Foundation of Korea (NRF). He has published numerous articles in leading journals. Additionally, he co-authored the book *Mathematical Methods in Elasticity Imaging*, published by Princeton University Press, USA, in 2015. His research interests include inverse problems, mathematical imaging, wave propagation, and machine learning.

SHUJAAT KHAN received the Ph.D. degree from the Department of Bio and Brain Engineering, Korea Advanced Institute of Science and Technology (KAIST), Daejeon, South Korea, in 2022. He is currently an Assistant Professor with the Department of Computer Engineering and a fellow with the Saudi Data and AI Authority (SDAIA) and King Fahd University of Petroleum & Minerals (KFUPM) under the SDAIA-KFUPM Joint Research Center for Artificial Intelligence (JRC-AI), KFUPM, Dhahran, Saudi Arabia. Prior to joining KFUPM, he was a Senior AI Scientist with the Digital Technology and Innovation, Siemens Medical Solutions USA, Inc. He was a Researcher with the Synergistic Bioinformatics (SynBi) and the Bio-Imaging, Signal Processing Learning (BISPL) Laboratories, KAIST. His research interests include machine learning, optimization, inverse problems, and signal processing, with a focus on biomedical and bioinformatics applications.