information

Article

Metaverse Applications in Bioinformatics: A Machine Learning Framework for the Discrimination of Anti-Cancer Peptides

Sufyan Danish [1], Asfandyar Khan [2], L. Minh Dang [3], Mohammed Alonazi [4], Sultan Alanazi [5], Hyoung-Kyu Song [3] and Hyeonjoon Moon [1,*]

[1] Department of Computer Science and Engineering, Sejong University, Seoul 05006, Republic of Korea; sufyandanish@sju.ac.kr
[2] Institute of Computer Science and Information Technology, The Agriculture University Peshawar, Peshawar 25000, Pakistan; asfandyar@aup.edu.pk
[3] Department of Information and Communication Engineering and Convergence Engineering for Intelligent Drone, Sejong University, Seoul 05006, Republic of Korea; minhdl@sejong.ac.kr (L.M.D.); songhk@sejong.ac.kr (H.-K.S.)
[4] Department of Information Systems, College of Computer Engineering and Sciences, Prince Sattam Bin Abdulaziz University, Al-Kharj 16273, Saudi Arabia; mn.alonazi@psau.edu.sa
[5] Department of Computer Science, College of Computer Engineering and Sciences, Prince Sattam Bin Abdulaziz University, Al-Kharj 16273, Saudi Arabia; sa.alanazi@psau.edu.sa
* Correspondence: hmoon@sejong.ac.kr

Abstract: Bioinformatics and genomics are driving a healthcare revolution, particularly in the domain of drug discovery for anticancer peptides (ACPs). The integration of artificial intelligence (AI) has transformed healthcare, enabling personalized and immersive patient care experiences. These advanced technologies, coupled with the power of bioinformatics and genomic data, facilitate groundbreaking developments. The precise prediction of ACPs from complex biological sequences remains an ongoing challenge in the genomic area. Currently, conventional approaches such as chemotherapy, target therapy, radiotherapy, and surgery are widely used for cancer treatment. However, these methods fail to completely eradicate neoplastic cells or cancer stem cells and damage healthy tissues, resulting in morbidity and even mortality. To control such diseases, oncologists and drug designers highly desire to develop new preventive techniques with more efficiency and minor side effects. Therefore, this research provides an optimized computational-based framework for discriminating against ACPs. In addition, the proposed approach intelligently integrates four peptide encoding methods, namely amino acid occurrence analysis (AAOA), dipeptide occurrence analysis (DOA), tripeptide occurrence analysis (TOA), and enhanced pseudo amino acid composition (EPseAAC). To overcome the issue of bias and reduce true error, the synthetic minority oversampling technique (SMOTE) is applied to balance the samples against each class. The empirical results over two datasets, where the accuracy of the proposed model on the benchmark dataset is 97.56% and on the independent dataset is 95.00%, verify the effectiveness of our ensemble learning mechanism and show remarkable performance when compared with state-of-the-art (SOTA) methods. In addition, the application of metaverse technology in healthcare holds promise for transformative innovations, potentially enhancing patient experiences and providing novel solutions in the realm of preventive techniques and patient care.

Keywords: artificial intelligence; machine learning; feature representation; classification; ensemble classifier; bioinformatics; metaverse; digital healthcare; sequence-based model; anticancer peptide

Citation: Danish, S.; Khan, A.; Dang, L.M.; Alonazi, M.; Alanazi, S.; Song, H.-K.; Moon, H. Metaverse Applications in Bioinformatics: A Machine Learning Framework for the Discrimination of Anti-Cancer Peptides. *Information* **2024**, *15*, 48. https://doi.org/10.3390/info15010048

Academic Editor: Khalid Sayood

Received: 7 December 2023
Revised: 10 January 2024
Accepted: 11 January 2024
Published: 15 January 2024

1. Introduction

Healthcare is paramount for global well-being and crucial in fostering services and contributing to economic growth [1]. Technological advancements such as blockchain, augmented reality (AR), and virtual reality (VR) have transformed patient–physician

interactions [2]. At the same time, the healthcare system in the metaverse integrates AI, telepresence, digital twinning, and blockchain to facilitate affordable treatments and improve patient outcomes [3]. Figure 1 illustrates the different applications of metaverse in healthcare. The metaverse also delivers personalized, immersive care by integrating real and virtual worlds [4]. Cancer remains a major global health threat, causing millions of deaths annually [5]. Although conventional treatments such as radiotherapy, target therapy, and chemotherapy can effectively target cancer cells, they also impose significant financial burdens and affect healthy cells [6]. To address these challenges, there is a growing need for an automated system to accurately identify ACPs from complex biological sequences, ensuring timely diagnosis and management before the condition worsens. Peptides have been recognized as effective cancer treatments due to their minimal impact on normal physiological activities and have been extensively studied in preclinical research for multiple purposes, including cardiovascular disease, diabetes, and other types of cancer [7]. In studies by Ijaz et al. [8], an early cervical cancer prediction model was developed using the CCPM. Saha et al. [9] studied different prognostic characters of GLS and GLS2 in cancer, showcasing their variance impact on clinical results across diverse cancer types. Figure 2 illustrates different cancer treatment options that involve targeting peptides.

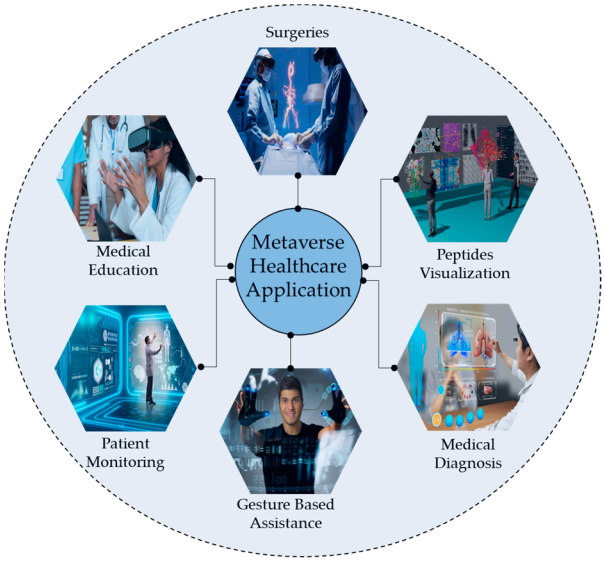


Figure 1. Several applications of the metaverse in healthcare and education for peptides visualization.

An ACP is a small sequence typically containing fewer than 50 amino acids [10]. A new direction for cancer treatment has been opened by determining ACPs because they have cationic properties that can easily target cancer-affected cells without interacting with normal cells [11]. However, ACPs have some potential drawbacks in the drug development process, i.e., high production costs and low stability [12]; despite this, they have some unique merits. ACPs are biological substances that naturally exist in a living organism; hence, they are more harmless than synthetic drugs and have better efficacy [13]. Over the last decade, peptide-based techniques have been used in numerous clinical trials for disease prevention and treatment [14], including heart disease, diabetes, and tumors. Experimental identification process of ACPs is very costly and time-consuming which is rarely utilized in clinical practice.

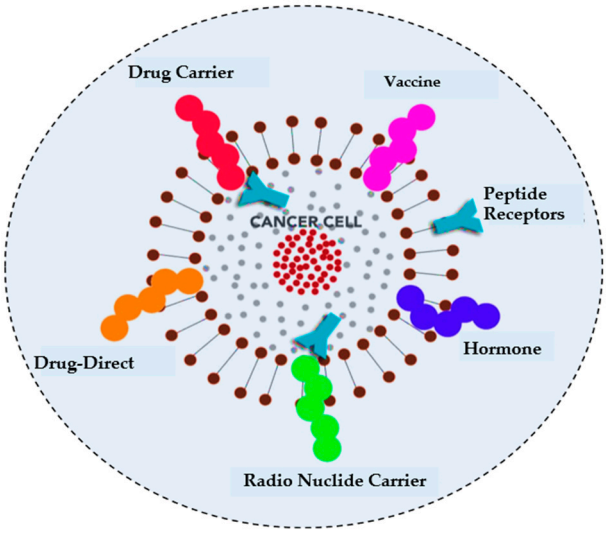


Figure 2. Peptide sequences offer a range of potential treatment options for cancer, which can be explored and evaluated for their effectiveness.

Therefore, in this study, a new computational-based framework is presented for the classification of ACPs that are mainly composed of trustworthy numerical descriptors which consists of AAOA, DOA, TOA, and EPseAAC. These descriptors are applied to extract salient and prominent patterns from peptide sequence. Additionally, to evaluate the performance of the proposed framework, various combination of machine learning techniques, such as support vector machine (SVM), naive Bayes (NB), and random forest (RF), are used for classification purpose. This study contains the following key contributions:

- A computational framework is proposed that extracts significant information from complex biological sequences, improving therapy for critical diseases. Leveraging the metaverse's immersive features, it enables precise identification of potential cancer treatments, revolutionizing disease management.
- Existing models face overfitting problems caused by class imbalances. To this end, we contribute by employing the synthetic minority oversampling technique (SMOTE) for preprocessing, reducing error scores, and fostering equitable feature learning. Additionally, we enhance overall model performance through a majority-voting ensemble decision-making technique in the final output, collectively advancing the robustness of our approach.
- We refined the existing pseudo amino acid composition (PseAAC) method for sequence classification by systematically incorporating additional physicochemical properties, addressing limitations arising from heterogeneous peptide patterns. This enhancement aims to generate context-rich features, improving the robustness and informativeness of obtained features and ultimately enhancing the methodology's effectiveness in capturing the complexity of peptide sequences.
- Comprehensive results from tests are obtained through analyses conducted on two sets of benchmark datasets, proving that the recommended trustworthy framework achieves new SOTA accuracy. Furthermore, ablation research is conducted to measure the effectiveness of each feature descriptor technique separately and evaluate the complementary strength produced from the diverse combinations of information.

The remaining paper consists of the following sections. Section 2 will explain a study of the literature on cancer peptide prediction, while the technical details of the proposed method are described in Section 3. Similarly, Section 4 contains the comprehensive experimental results with implementation details and comparisons; and in the final Section 5, our work concludes followed by future research direction.

2. Literature Review

Several intelligent models for identifying ACPs have been proposed by different researchers. For example, a silico model was proposed by [15], where AAC and binary profile-based information are used as a feature extraction method, obtaining 91.44% accuracy for binary classification. Similarly, Hajisharifi et al. [16] developed a novel approach, where PseAAC and local alignment kernel techniques are explored for feature extraction and achieved the outcomes of 89.7% and 83.82%, respectively. Next, this work is further boosted by Chen et al. [17], who introduced a sequence-based model that utilizes g-gap dipeptide composition and cross-validation techniques for superior performance; therefore, 94.77% accuracy was obtained for the discrimination of peptides. Akbar et al. [18] utilized an ensemble classifier for the discrimination of various peptides to obtain robust features: amphibolic PseAAC, reduced AAC, and g-gap DPC were considered, and a 96.45% classification score was obtained. Xu et al. [19] developed a model that considers ACP composition by incorporating binary encoding and physicochemical properties to obtain meaningful information but the performance is 91.86% which is a low discrimination score as compared to the other models. The sequence-based models have successfully identified novel ACP but due to technological advancement substantial work has been presented in the metaverse specifically for human healthcare.

Recent research has explored the metaverse's potential in healthcare, covering various applications and challenges. Bansal et al. [20] outlined its uses in immersive training experiences, telemedicine, and clinical care, emphasizing extended reality (XR) technologies and hardware. Another study [21] focused on ophthalmology, utilizing VR and digital twins (DTs) for personalized care. Next, Ali et al. [22] proposed a secure metaverse architecture with blockchain and eXplainable AI (XAI) for transparent disease prediction. Moreover, Razdan and Sharma [23] proposed a comprehensive metaverse architecture, emphasizing big data processing and security measures, while a reference architecture with distinct layers is also presented for metaverse applications, ensuring robust governance and quality of service.

Moreover, some researchers have used alternative methods such as the generalized chaos game representation [24] short-term memory models with binary profile features and k-mer sparse matrices using two novel benchmark datasets (ACP740 and ACP240) [25]. Chen et al. introduced an ACP-DA model, composed of DL and augmentation approaches, resulting 82.03% and 88.33% on ACP740 and ACP240 datasets, respectively [26]. Next, Ye et al. [27] proposed an ensemble learning model using numerous datasets [16,28] with which 95.4% and 92.4% accuracies were obtained after comprehensive experiments. Moreover, in [29] ETree classfier and AAC feature extractor are used for the discrimination of different biological sequences. Akbar et al. also explored ensemble classification for ACPs and attained 96.45% accuracy using an evolutionary genetic algorithm (iACP-GAEnsC) [18]. Shahid Akbar et al. developed cACP-DeepGram by utilizing three statistical feature representation schemes and achieved satisfactory outcomes of 96.94% accuracy [30]. Shahid Akbar et al. also introduced cACP, a discriminatory computational technique employing diverse statistical feature representation schemes, and feature selection PCA, and achieved a tremendous classification score of 96.91% [31]. Shahid Akbar et al. made a notable contribution with their proposed cACP-2LFS, a novel sequential discriminative model designed for ACP classification. In addition, their model utilizes the K-space amino acid pair (KSAAP) for extracting correlated descriptors and incorporates a two-level feature selection (2LFS) method, resulting in impressive accuracy rates of 94.11% and 93.72% using independent and LEE datasets, respectively. These achievements highlight its potential applications in the fields of medicine, proteomics, and research academia [32]. Despite these attempts, some recent studies have explored DL models for the efficient classification of ACPs from complex biological sequences. Ahmed et al. presented a novel multi-head deep CNN by extracting discriminative physicochemical properties from diverse peptides, followed by evolutionary-based features, leading 83.0%, 86.0%, and 91.0% accuracies levels on ACP-240, ACP-740, and a combination both datasets, respectively [33]. Hulam

et al. proposed DL-based method that uses dipeptide deviation from the expected mean (DDE) as a feature extractor for precise predictions of ACPs, outperforming the 85.88% and 84.8% accuracies score using ACP240 and ACP740 datasets [34]. Advanced DL methods such as MLACP 2.0 [35] incorporate seven distinctive classifiers and multiple feature encoding methods. In addition, other mainstream approaches applied neural networks and multitask learning, where they incorporate hybrid sequencing information [36,37].

Recently, ensemble classifiers have gained popularity in ACP classification, where a decision has been made based on the majority vote of multiple machine learning algorithms, such as an SVM, an RF, and AdaBoost. Ensemble classifiers leverage individual algorithms for their strengths and mitigate their weaknesses for improved performance. The results obtained with this approach are superior to those from a single algorithm [38]. Multi-algorithm ensemble clustering aggregates the results of different clustering algorithms to produce more accurate and robust consensus clustering results. Ensemble clustering has many applications, such as image segmentation and gene expression analysis, and can make the clustering process more robust and accurate by embracing ensemble methods [39]. Various integrated methods have been used to evaluate regression [40,41] and classification [42,43] problems in the literature, including weighted averaging [44], weighted voting [45], simple averaging [46], and majority voting [47]. Other techniques can also be used for protein analysis, such as combining individual sequence-based models to generate a final prediction. To predict the biological functions and properties of peptides and proteins, sequence-based statistical predictors require machine learning algorithms and feature extraction techniques. Predictors of protein–protein interactions, protein subcellular localization, and ACPs have shown promising results in several applications, and developing new therapeutics requires the development of such predictors [48].

Several computational methods have been developed to classify ACPs and non-ACPs, but there is still room for improvement. To overcome the research gap mentioned later in the paper, this study proposed a new collection of computational approaches to improve the classification performance for new biological sequences. The suggested system could lead to a better understanding of ACPs and aid in the development of new therapeutics.

- The imbalanced nature of datasets in anticancer peptide classification poses a significant hurdle for many existing machine learning methods. Biased models can emerge, favoring classes with a higher number of instances, thereby compromising the model's ability to accurately identify and classify less-represented classes. This imbalance issue is particularly critical in the context of anticancer peptides, where a thorough understanding of diverse instances is crucial for effective classification.
- Prevailing methods often lean towards simplicity, employing single-feature extractions and classifiers. While this simplicity aids in model interpretability and computational efficiency, it may fall short of capturing the intricate and nuanced patterns inherent in anticancer peptides. The complex nature of these peptides demands more sophisticated approaches that can discern subtle variations and relationships within the data, enhancing the model's discriminatory power.
- Current strategies aimed at enhancing classification accuracy often resort to fusion techniques. While these techniques offer potential improvements, they may inadvertently introduce homogeneity in the utilized information, leading to limiting the model's ability to discern diverse and subtle characteristics crucial for accurate anticancer peptide classification. Striking a balance between fusion for improved accuracy and preserving the diversity of information remains a key challenge in developing robust models.
- Some machine learning-based methods in anticancer peptide classification may exhibit a tendency to overlook the expansive landscape of feature extraction models and selection techniques. A more comprehensive exploration of this landscape is imperative to ensure that potentially more effective approaches are not neglected. The diversity among anticancer peptides demands a thorough examination of various

feature extraction methods and selection techniques to uncover the most suitable combination for accurate classification.

In this research, we experimentally verified in aspects of overcoming the limitations in the previous studies. Our approach involves the generation of noteworthy features from peptides through the incorporation of distinctive composite features, the SMOTE, PCA, and ensemble classifiers.

3. The Proposed Framework

The following section provides information regarding data acquisition and preprocessing. Next, we discuss techniques used to extract discriminative features followed by a detailed explanation of ensemble learning, where multiple classifiers are integrated to improve classification performance.

The technical details of the proposed method give a deeper understanding of the methodology used and the justification for the decision-making. The entire steps of the proposed methodology are shown in Figure 3. The complete procedure of the proposed model for peptide classification is presented in Algorithm 1.

Algorithm 1: Pseudocode for the proposed framework for peptide classification

Input: Peptide Sequences ($\check{S}_{et\_Train}$), ($\check{S}_{et\_Test}$)→($\mathcal{P}_{\check{S}eq}$)→(Å)U(∀)
//Å and ∀ represent anticancer and non-anticancer peptides
Output: Peptide Sequences ($\check{S}_{et\_Test}$) → Y*// represents the class label of the test dataset*

(1) Pre-Initialization:

($\check{S}_{et\_Train}$), ($\check{S}_{et\_Test}$)→(CD-Hit) → Ð→*Eliminate short and homogenous sequences*

(2) Feature Extraction and Training Procedure:

($\check{S}_{et\_Train}$) →M_1, M_2, M_3, M_4 *//M represents feature extraction methods*
Ð→ ($\mathcal{P}_{\check{S}eq}$)→(Å)U(∀)*//Ð represents the refined dataset*
for *i* = 1 to L-1→Ð*//L represents samples in the dataset*
Compare pattern (C)→ ($\mathcal{P}_{\check{S}eq}$)
Extract features(f) → ($\mathcal{P}_{\check{S}eq}$)
Save features→ (SF)
Repeat (Λ) → M_1, M_2, M_3, M_4 *//Λ represents repeating step 2 for all methods*
end

(3) Feature Selection Procedure:

Apply PCA→ (P)→SF
Refine features→ ($\mathcal{H}f_1$)

(4) Class Balancing Procedure:

Apply SMOTE (Š) → ($\mathcal{H}f_1$)→($\mathcal{P}_{\check{S}eq}$)→(Å)U(∀)
Equal sample → *(Æ)* → ($\mathcal{H}f_2$)

(5) Model Training Procedure:

($\check{S}_{et\_Train}$) → ($\mathcal{H}f_2$)
($\mathcal{H}f_2$) + *class label (CL)* → *ensemble classifier (EC)* →

(6) Model Testing Procedure:

Load ($\check{S}_{et\_Test}$)→($\mathcal{P}_{\check{S}eq}$)→(Å)U(∀)
for *j* = 1 to L-1→Ð*//L represents samples in the test dataset*
Repeat step 2→ feature extraction
Output: Binary classification with class label

end

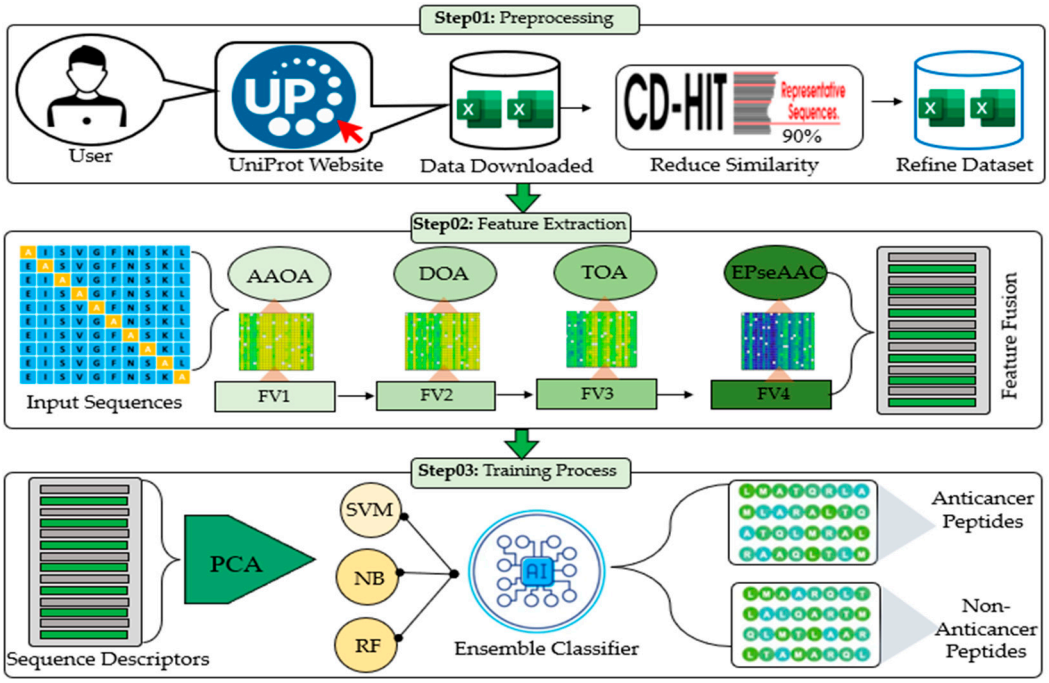


Figure 3. The proposed computational-based framework for discriminating ACPs and non-ACPs from biological sequences.

3.1. Dataset and Preprocessing

In this section, a brief description of the benchmark dataset is provided followed by the critical step for data refinement. Further information about these is given in the subsections along with the dataset statistics.

3.1.1. Dataset

Datasets of anti-cancer peptides allow new cancer treatments to be discovered. Researchers can take advantage of these datasets to gain a comprehensive and organized overview of anti-cancer peptides compiled from various sources. A systematic analysis of these datasets can identify common characteristics and features associated with specific peptides that defend against cancer, leading to the design and development of new drugs using peptides.

As a result, researchers can explore and test potential drug candidates using these datasets, thus speeding up the drug discovery process. Our understanding of cancer biology and the development of effective treatments for this devastating disease are greatly enhanced when anti-cancer peptide datasets are publicly available and analyzed. A statistical report of the dataset samples is given in Table 1.

Table 1. Statistical information of both datasets used for the experiments.

Dataset	Total Samples	Training Samples	Test Samples
Benchmark [17]	344	275	69
Independent [17]	300	240	60

Selecting or creating a valid benchmark dataset is essential for developing a computational predictor in machine learning or pattern recognition. Several datasets are available to us for this purpose, but in this study, we used a standard dataset [17]. S^+ and S^- sets were combined to create the benchmark dataset for this study. In the following Equations (1) and (2), S^+ and S^- represent ACPs and non-ACPs, respectively. Using a widely recognized and standardized dataset, our computational predictor can be validated and tested for validity

and reliability, and our approach can be compared to previous studies. The selection of common and uncommon samples is based on Equations (1) and (2).

$$S = S^{+} \cup S^{-} \tag{1}$$

It is essential to note that the intersection of the sets S^{+} and S^{-} in the dataset used in this study should be empty. In other words, no peptides should exist in both S^{+} and S^{-}. This is because any peptide that exists in both sets would be difficult to classify accurately, as it would have characteristics of both ACPs and non-ACPs. As a result, we ensured that the intersection of S^{+} and S^{-} is null to provide our dataset's integrity and reliability.

$$\phi = S^{+} \cap S^{-} \tag{2}$$

A benchmark dataset consisting of 138 anticancer peptide sequences and 206 non-anticancer peptide sequences was utilized in this study. Furthermore, there were 150 non-ACPs and 150 ACPs in the independent dataset. Both datasets were processed in the same way.

However, a cluster database at high identity with tolerance (CD-HIT) program was used to deal with any redundancies in the dataset [49]. Large datasets are frequently clustered and redundant sequences are removed using this program. CD-HIT allowed us to detect and eliminate duplicate sequences from the dataset so that we could use it for further analysis after it was cleaned with CD-HIT. As a result, any patterns or relationships identified should not be due to an overrepresentation of specific sequences but to the representation of the underlying population.

3.1.2. Preprocessing Using the SMOTE

The preprocessing step plays a significant role in various computer vision [50,51] and time series [52,53] tasks that show tremendous performance. To this end, imbalance is one main issue that degrades the entire model's performance. One of the most significant problems is that the classifier may become biased towards the majority class, resulting in deficient performance of the minority class. There are several approaches to handling imbalances in datasets. Some of these approaches involve resampling the actual dataset, while others involve giving different weights or costs to the training data. Resampling techniques involve reducing the number of examples in the majority class (under-sampling) or increasing the instances in the minority class (oversampling) to accomplish a balanced distribution of examples across all classes in the training dataset. However, oversampling with replacement has been found to have irrelevant prediction improvement in the minority class; this approach might pinpoint more specific regions within the minority class, potentially leading to overfitting of the classifier.

To address this issue, Chawla et al. [54] proposed a powerful method known as the synthetic minority oversampling technique (SMOTE) for balancing the samples of each class. In this paper, we also applied SMOTE as a solution to address the dataset's class imbalance issue. The authors demonstrated the performance of various techniques, both with and without the SMOTE, showcasing their accuracies. The authors found that, even without the SMOTE, different methods exhibited effective performance. However, the dataset's imbalance led to overfitting. Nevertheless, this problem was resolved when the SMOTE was applied to all four feature extraction techniques utilized in this paper. This approach involved concatenating the feature vectors, leading to significantly improved and consistent accuracies across all assessment methods. Overall, machine learning and deep learning applications [55] have benefited from the SMOTE algorithm as a solution to imbalanced datasets.

3.2. Computational Methods for Discriminative Features

Since machine learning algorithms only work with numerical vector data, they cannot be trained directly on biological sequences in the genomic era. No machine learning technique deals with sequence samples, as previously noted in a comprehensive review [56]. Our study utilized different protein encoding techniques in order to resolve this issue. A machine learning algorithm can process biological sequences by converting them into numerical vectors. In order to accurately classify peptides into anticancer and non-anticancer classes, we used these encoding techniques to extract meaningful features from the biological sequences. As a whole, protein encoding techniques hold great potential for improving our understanding of the underlying biology of diseases such as cancer, since they enable machine learning approaches to be applied to biological sequence data.

3.2.1. Amino Acid Occurrence Analysis (AAOA)

AAOA is a method that characterizes peptide sequences by analyzing amino acid occurrences; it is also referred to as AAC. By dividing each amino acid frequency by the length of the peptide sequence for a given peptide sequence, this method determines how frequently each amino acid occurs in each peptide sequence. AAC provides a simple and informative representation of the underlying sequence knowledge by decomposing peptide sequences into twenty-dimensional vectors. Each dimension represents how often each amino acid occurs in the sequence.

Researchers have used AAC to solve different biological problems, such as predicting mitochondrial proteins [57] and predicting subcellular localization [58]. An essential component of AAC's effectiveness is its ability to capture sequence features that are associated with peptide functions or activities. Structural or functional indications of peptide properties may be based on the frequency of specific amino acids, such as hydrophobicity or charge.

AAC is easily calculated using a mathematical formula. As $f_1, f_2, f_3, \ldots, f_n$ represents the frequency of each amino acid occurring in the sequence, we can determine how often that amino acid appears within a sequence of length L. An amino acid composition vector of the peptide sequence is generated by normalizing these frequency values to sum to 1:

$$AAOA = (f_1, f_2, f_3, \ldots, f_n)/L \quad (3)$$

This method can provide machine learning models with information about peptides' anticancer and non-anticancer properties using informative features extracted from their sequences.

3.2.2. Dipeptide Occurrence Analysis (DOA)

Among various prediction problems such as subcellular localization [58] and deoxyribonucleic acid (DNA)-binding proteins [59], dipeptide occurrence analysis, also known as dipeptide composition (DPC), is widely used. DCP assesses peptide sequences by estimating how often two adjacent amino acids occur together, versus AAC, which only calculates the frequency of each amino acid. For each peptide, this method creates a four hundred-dimensional vector whose elements represent the frequency of its dipeptides. It is advantageous to use DPC because it captures global information about the peptide sequence, whereas AAC does not. DPC primarily determines the peptide pattern. DPC is calculated by multiplying the frequency of all possible dipeptides in the sequence by the length of the peptide. Equation (4) can be used to represent DPC mathematically:

$$DOA(i,j) = f(i,j)/N \quad (4)$$

An amino acid sequence is defined by the peptide sequence length N and the frequency $f(i,j)$ of the dipeptide formed by the ith and jth amino acids. Feature vectors for DOA are 400 dimensions long, each element representing the frequency of individual dipeptides.

3.2.3. Tripeptide Occurrence Analysis (TOA)

Tripeptide occurrence analysis is also referred to as tripeptide composition (TPC). Encoding peptide sequences with numerical values captures deeper information about them. A feature vector of 8000 dimensions is generated after computing the frequency of every tripeptide in the peptide sequence. Elements of the vector represent tripeptide frequencies. The TPC analysis provides a more in-depth picture of peptide composition when compared to the AAC and DPC methods because it considers three consecutive amino acids. TPC generates new patterns based on three collective amino acids (*i*, *j*, *k*), revealing more peptide sequence information.

These studies demonstrate that TPC can be used to extract critical patterns from peptide sequences that indicate specific biological functions or characteristics. Equation (5) can be used to express TPC mathematically (5):

$$TOA(i,j,k) = f(i,j,k)/N \quad (5)$$

An amino acid sequence is defined by the peptide sequence length N and the frequency $f(i,j,k)$ of the tripeptide formed by the ith, jth, and kth amino acids. Feature vectors for TPC are 8000 dimensions long, each element representing the frequency of individual TPC.

3.2.4. Enhanced Physicochemical Property-Based Features

EPseAAC is an expansion of AAC that incorporates physicochemical properties such as side chain mass, polarity, and hydrophobicity in the feature vector. These properties can further elucidate peptide sequence–structure relationships. Among the wide variety of biological prediction tasks performed with EPseAAC are: subcellular localization of proteins [44], protein structural class prediction [45], and allergenic protein prediction [46]. As part of EPseAAC, the physicochemical properties of the amino acids in the peptide sequence are used to construct a series of descriptors. Consequently, the sequence's descriptors and amino acid frequencies are combined to produce a more informative feature vector. As EPseAAC is mathematically formulated, each amino acid is described according to a set of properties, including side chain mass, polarity, and hydrophobicity. Combining these descriptors and amino acid frequencies creates a final feature vector. As a result, EPseAAC can boost peptide classification performance by extracting powerful features. This technique captures the physicochemical properties of peptide sequences, allowing for a more precise representation and the possibility of discovering new bioactive peptides [60–63].

$$P = [a_1, a_2, \ldots a_{20}, a_{20+1}, a_{20+2}, \ldots .a_{20+\lambda}]^t (\ \lambda = 1, 2, \ldots \ldots 21) \quad (6)$$

Furthermore, EPseAAC considers the correlation factor of each amino acid in addition to the frequency of each amino acid. Several factors must be considered, including charge, irreplaceability, polarizability, polarity, surface area, flexibility, hydrophilicity, solvent accessibility, rigidity, and hydrophobicity. Feature extraction methods incorporating these properties can improve the accuracy of peptide predictions as they play a key role in peptide classification. The parameter λ in Equation (6) represents various addition techniques for the physiochemical properties. In other words, different values of λ can be used to weigh the different properties' importance when computing the feature vector for a peptide sequence. In this study, the researchers tried different λ values and realized that λ = 1 gave optimal outcomes for their classification task. EPseAAC is a powerful system for encoding peptide sequences that incorporates the amino acid frequency and its physical properties. In our study, λ represents a tuning parameter that influences the weighting of certain features or aspects within our proposed technique. Specifically, we experimented with different values of λ, including 1, 2, and 3, to assess their impact on the performance of our model. The choice of λ reflects the balance between several factors, and through empirical testing, we found that λ = 1 yielded the most favorable results.

$$\begin{cases} \phi 1 = \frac{1}{length-1}\sum_{k=1}^{length-1} I_k, k+1 \\ \phi 2 = \frac{1}{length-2}\sum_{k=1}^{length-2} I_k, k+1 \\ \phi 3 = \frac{1}{length-3}\sum_{k=1}^{length-3} I_k, k+1 \\ \phi 4 = \frac{1}{length-4}\sum_{k=1}^{length-4} I_k, k+1 \\ \phi 5 = \frac{1}{length-5}\sum_{k=1}^{length-5} I_k, k+1 \\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ \phi \lambda = \frac{1}{length-2}\sum_{k=1}^{length-2} I_k, k+1 \end{cases} \tag{7}$$

This allows for a more comprehensive and informative representation of the peptides, which can be used for various computational proteomics tasks, such as the prediction of membrane types, DNA-binding proteins, and peptide classification.

The variables $\phi 1, \phi 2,$ and $\phi \lambda$ in the EPseAAC equation correspond to the peptide $\phi \lambda$ sequence length and the ranks of the correlation factors. Specifically, $\phi 1$ is the peptide sequence length, $\phi 2$ is the first correlation factor rank, and $\phi \lambda$ is the last correlation factor rank.

3.3. Ensemble Learning for Model Training

In machine learning, bioinformatics and data mining classification play an essential role. An algorithm for classifying new instances is trained in predefined classes and then used to categorize new cases based on the training data. Classifiers can perform better in prediction by using ensemble classification. Multiple classifiers combined into an ensemble deliver a more precise prediction than a single classifier.

Several computational models use ensemble classification to decrease discrepancies due to inconsistent training sets. In addition to improving the model's generalization, combining several classifiers also improves its strength. To further enhance the performance of an ensemble classifier, the authors propose combining three classifiers, namely SVM, NB, and RF.

In many classification problems, an SVM is commonly used as a machine learning algorithm. Data with high dimensions can be handled using SVMs, as can small datasets. An RF is an algorithm based on decision trees that combines the output of several trees to create better accuracy. In naïve Bayes classification, features are assumed to be independent according to the Bayes' theorem, which is known as probabilistic technique. Large datasets can be easily processed since they are simple and efficient.

Due to each individual classifier's strengths complementing each other, SVM, RF, and NB algorithms should provide better performance in an ensemble classifier. The RF is robust to noise and outliers, while the NB effectively handles high-dimensional datasets. The SVM is efficient when handling complex datasets with non-linear decision boundaries. Therefore, ensemble classifiers could increase classification model accuracy and robustness.

$$\mathrm{E}_{\text{ensemble}} = \mathrm{SVM} \oplus \mathrm{RF} \oplus \mathrm{NB} \tag{8}$$

Multi-classifier ensembles are used to create a final classification by combining multiple classifiers. By reducing bias and variance in the classification process, ensemble classifiers aim to improve overall classification performance. Voting algorithms can be used to merge the outputs of multiple classifiers. Voting algorithms involve predicting the input classification of each classifier and determining the final classification through a majority vote among them. For example, if three classifiers predict that an input belongs to class A and two classifiers predict that it belongs to class B, the final classification would be class A. The $\oplus$ symbol in Equation (8) likely represents a margin operation, which is a way to adjust

the output of each classifier before the voting algorithm is applied. The margin operation can help to improve classification performance.

$$(C1, C2, C3)\ \mathcal{E}\ (A1, A2) \tag{9}$$

Here, $(C1, C2, C3)$ represent individual classifiers, and A1 and A2 represent predefined classes of ACPs and non-ACPs.

In the end, the outcome of the $E_{ensemble}$ using a voting algorithm is attained.

$$E_{ensemble} = \text{Maxi}(w1 \times 1, w2 \times 2, w3 \times 3) \tag{10}$$

In this equation, $E_{ensemble}$ represents the ensemble classifier with the voting algorithm, Maxi represents the maximum achievement, and $w1, w2, w3$ are the best weights of the classifiers for the class that receives a high vote.

4. Experimental Analysis

This section presents an analysis of the experimental results, a comparison of our computational-based model to SOTA techniques, and an examination of the experimental results obtained from three peptide sequence datasets. The results obtained are analyzed in depth to provide insights into the performance and effectiveness of our model.

4.1. Implementation Setup and System Specifications

The proposed peptide discrimination system has specific software and hardware requirements. TensorFlow version 2.5.0, Keres framework V2.5.0, and Python programming language V3.8.5 are required. These software components provide the necessary tools and libraries for training and developing machine learning models that are utilized in the classification system. In addition, advanced micro devices (AMDs) Ryzen 9 3900X central processing units (CPUs) with 12 cores, GeForce RTX 3090 graphics cards, and 48.0 GB RAM are required, along with Windows 10. These hardware components are essential for the machine learning models to train and run efficiently. For validation, the dataset was split between 80% training and 20% testing to prepare the data for the experiment. Machine learning models are often evaluated by leaving out a part of the training data to determine their capability to generalize to novel, untried data.

4.2. Ablation Study

In the first step, we evaluated a single extractor's performance for discriminating complex peptides. Each component was evaluated to determine whether it could identify complex peptides and differentiate them accurately. To implement the proposed strategy, the optimal components are selected based on the results from the analyses. Every component of the model must be evaluated to ensure it is working as intended and participating in the model's overall effectiveness. Additionally, we examined whether imbalanced samples affected the results. Imbalanced datasets can affect machine learning models, which is an essential part of the experiment. In Section 1, we mentioned the second contribution, which we were capable of showing by assessing the effects of imbalanced samples on the results. This means that the proposed strategy addresses the problem of imbalanced datasets. Overall, this part of the experiment highlights the importance of analyzing and evaluating the components used in a machine learning model to ensure optimal performance. You can create a more effective and accurate model by thoroughly evaluating each component and assessing its impact on the results. Additionally, evaluating the impact of imbalanced samples is critical in addressing this common issue in machine learning and improving the model's accuracy in real-world applications. In our extensive experimental analysis, we deduce that the performance of the proposed model significantly improves with the application of the SMOTE to rearrange the samples of classes within the dataset. However, it becomes evident that training the model over a variable number of samples for each class leads to convergence towards the dominant class, resulting in predominantly incorrect pre-

dictions. Acknowledging this issue, it becomes apparent that the crucial step of balancing samples is indispensable for achieving enhanced performance in the model's predictions.

4.2.1. Analysis of Benchmark Dataset

To find the optimal model, it is necessary to evaluate each component and its possible fusions. Therefore, this study applied the same strategy where various feature extractor methods integration are assessed for discriminating ACPs and non-ACPs.

As in the proposed framework, four machine learning computational sequence features are used, each with one functionality and way of extracting information from complex biological sequences. Initially, the experiments were conducted over the first five computational methods, where the best result was obtained through TOA with 95.17%, 93.05%, 97.26%, and 89.96% accuracy, sensitivity, specificity, and Matthews correlation coefficient (MCC), respectively. The reason for the best result is that it maintains a strong correlation among different peptides as compared to the other techniques. Next, various dual-feature fusions were examined, wherein diverse/informative patterns from the biological sequence are extracted, and the ensemble classifier learns more discriminative information. Moreover, like the previous empirical analysis, we also investigated different λ values to check the complementary power of the physicochemical properties and their way of addition. The dual-feature fusion mechanism boosted our model's discriminative peptide performance by using TOA + EPseAAC (λ = 3), where a high score was attained, including 97.24%, 95.83%, 98.63%, and 94.33% levels of accuracy, sensitivity, specificity, and MCC, respectively. Inspired by such feature representation fusion performance, we tried multi-feature fusion, wherein experiments are conducted through various components. After comprehensive experimental analysis, we concluded that AAOA + TOA + EPseAAC (λ = 3) achieved an accuracy of 96.55%, sensitivity of 95.83%, specificity of 97.26%, and MCC of 93.02%. Compared to the previous component collection, the discriminative performance is worsened here due to the redundant features that confuse the classifier during classification. All of the results of these various feature fusion descriptors are reported in Table 2. Furthermore, the proposed model was also deeply evaluated using three λ values and then selecting the most optimal one. The best results were achieved with the usage of PCA, which can pick the most prominent, robust, and representative features. The empirical results are given in Figure 4, while Figure 5 represents the confusion matrix of the benchmark dataset.

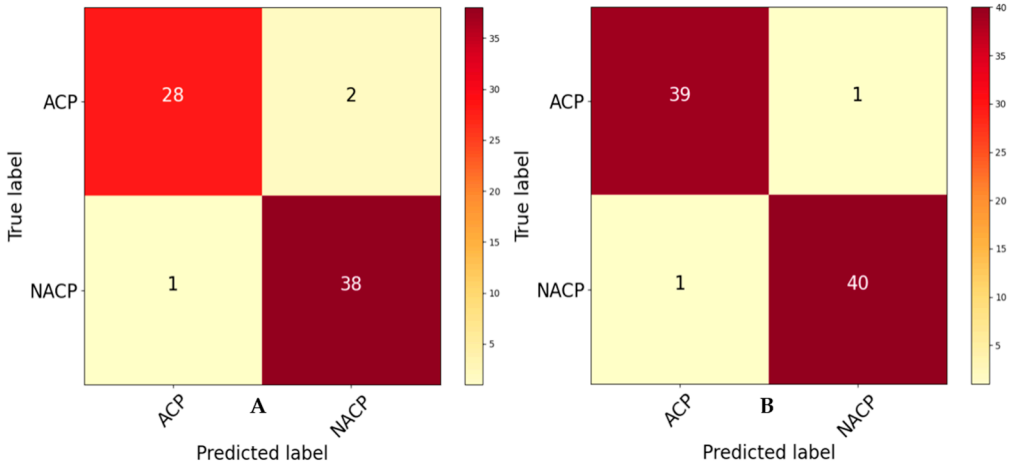


Figure 4. Confusion matrix analysis of hybrid feature extraction methods utilizing an ensemble classifier on a benchmark dataset: PCA and SMOTE. (**A**) shows the results of the benchmark dataset without the SMOTE, while (**B**) displays results with the SMOTE on the benchmark dataset.

Table 2. Performance evaluation over several feature descriptors using the SMOTE, PCA, and ensemble classifier.

Method	Accuracy	Sensitivity	Specificity	MCC
Benchmark Dataset				
AAOA	88.27	86.84	89.85	76.02
DOA	92.41	91.66	93.15	84.62
TOA	95.17	93.05	97.26	89.96
EPseAAC (λ = 1)	86.89	83.33	90.41	71.83
EPseAAC (λ = 2)	88.96	86.11	91.78	76.69
EPseAAC (λ = 3)	87.58	84.72	90.41	73.73
AAOA + DOA	93.79	91.66	95.89	87.11
AAOA + EPseAAC (λ = 1)	88.27	84.72	91.78	74.84
AAOA + EPseAAC (λ = 2)	83.57	76.92 s	86.13	62.04
AAOA + EPseAAC (λ = 3)	88.27	84.72	91.78	74.84
AAOA + TOA	90.90	96.20	83.01	79.17
DOA + EPseAAC (λ = 1)	91.72	84.72	98.63	80.33
DOA + EPseAAC (λ = 2)	94.48	90.27	98.63	87.89
DOA + EPseAAC (λ = 3)	93.83	90.27	97.29	86.70
DOA + TOA	91.03	94.44	87.67	82.45
TOA + EPseAAC (λ = 1)	96.55	95.83	97.26	93.02
TOA + EPseAAC (λ = 2)	95.86	95.83	95.89	91.71
TOA + EPseAAC (λ = 3)	97.24	95.83	98.63	94.33
AAOA + DOA + TOA	94.48	98.61	90.41	89.10
AAOA + DOA + EPseAAC (λ = 1)	93.10	88.88	97.26	84.93
AAOA + DOA + EPseAAC (λ = 2)	92.41	90.27	94.52	84.24
AAOA + DOA + EPseAAC (λ = 3)	93.79	93.05	94.52	87.43
AAOA+ TOA + EPseAAC (λ = 1)	95.86	93.05	98.63	91.22
AAOA + TOA + EPseAAC (λ = 2)	94.48	93.05	95.89	88.07
AAOA + TOA + EPseAAC (λ = 3)	96.55	95.83	97.26	93.02
DOA + TOA + EPseAAC (λ = 1)	95.17	91.66	98.63	89.59
DOA + TOA + EPseAAC (λ = 2)	95.86	94.44	97.26	91.52
DOA + TOA + EPseAAC (λ = 3)	93.79	91.66	95.89	87.11
Proposed (without SMOTE)	95.65	95.55	95.00	91.16
Proposed (with SMOTE)	97.56	97.55	97.54	95.12

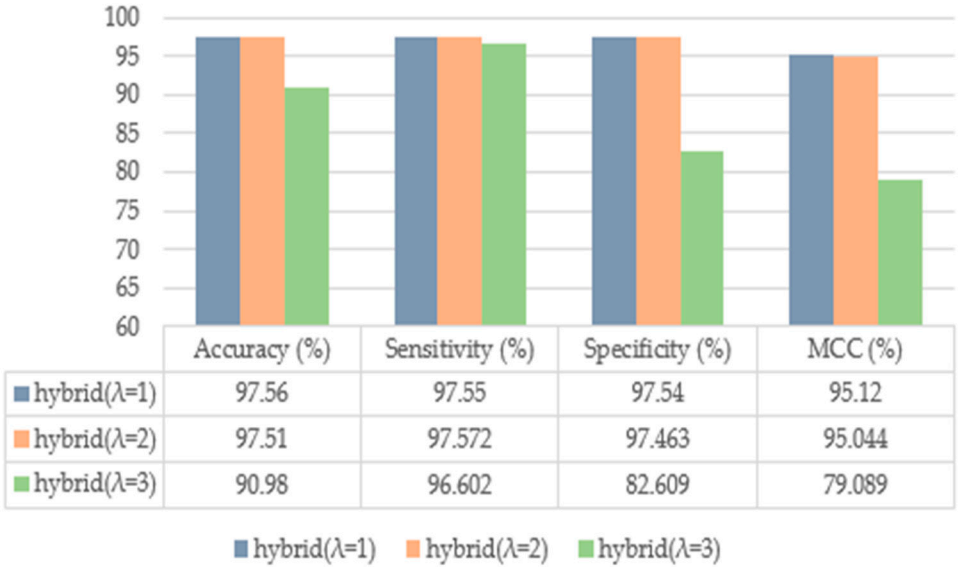
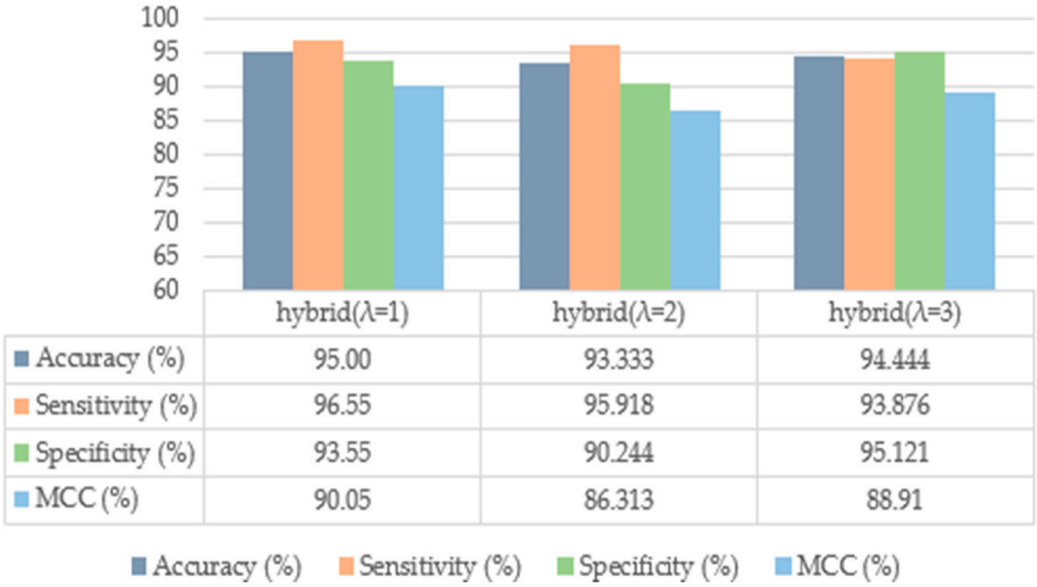


Figure 5. Graphical representation of the prediction rate of hybrid feature patterns methods using an ensemble classifier on a benchmark dataset: PCA and SMOTE.

4.2.2. Analysis with Independent Datasets

We conducted experiments on independent datasets to check the generalizability of the proposed model and different combination of feature extraction methods for classification and discrimination of ACPs and non-ACPs. Therefore, this study applied the same strategy where various feature extraction methods are integrated to utilized fairly compare outcomes with other recent approaches that have contributed a lot to the discrimination of peptides. Four feature descriptors are combined into the final hybrid feature strategy and then forwarded to the PCA to choose the most refined information. The initial experiments were conducted using the first five computational methods, and the best results were obtained through DOA, which achieved 91.11%, 95.91%, 85.36%, and 81.46% accuracy, sensitivity, specificity, and MCC, respectively. This result is because DOA considers two mutual peptides; therefore, dual-peptide patterns are more common than the others. Subsequently, various dual-feature fusion mechanisms were examined to extract diverse and informative patterns from biological sequences and enable the ensemble classifier to learn more discriminative information. After deeply analyzing the dual-feature fusion of AAOA + DOA we achieved good classification scores including 92.22%, 95.91%, 87.80%, and 83.90% for accuracy, sensitivity, specificity, and MCC, respectively. After a comprehensive experimental analysis, we concluded that the optimal combination was AAOA + DOA + TOA, which achieved an accuracy of 92.22%, sensitivity of 93.87%, specificity of 90.24%, and MCC of 84.27%. However, its discriminative performance was slightly lower than that of the previous component combination due to redundant features confusing the classifier during classification. All of the results of these various feature fusion descriptors are reported in Table 3, while the graphical outcomes of the proposed model with a diverse value of λ are represented in Figure 6, and Figure 7 shows the confusion matrix of an independent dataset.



	hybrid(λ=1)	hybrid(λ=2)	hybrid(λ=3)
Accuracy (%)	95.00	93.333	94.444
Sensitivity (%)	96.55	95.918	93.876
Specificity (%)	93.55	90.244	95.121
MCC (%)	90.05	86.313	88.91

Figure 6. A graphical representation of the prediction rate of hybrid feature extraction methods using an ensemble classifier on an independent dataset.

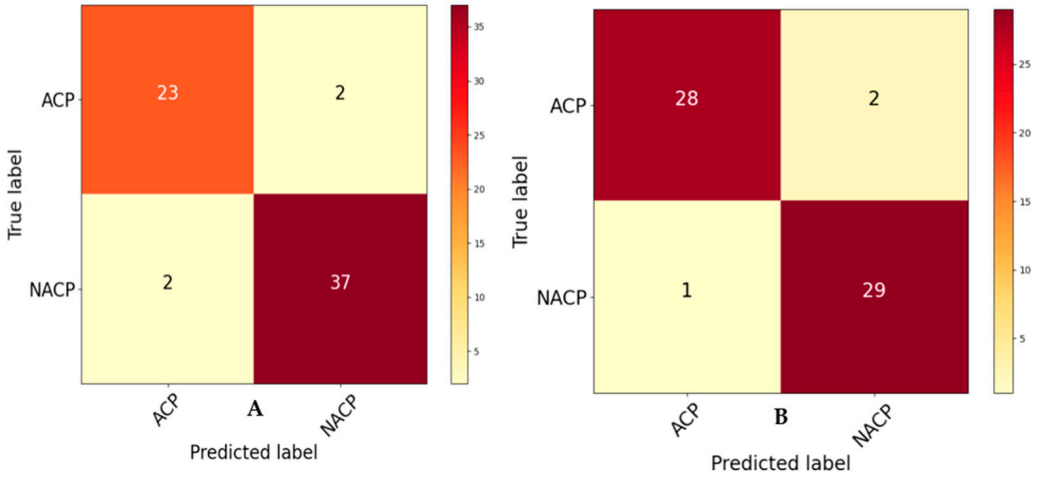


Figure 7. Confusion matrix analysis of hybrid feature extraction methods utilizing an ensemble classifier on an independent dataset: PCA and SMOTE. (**A**) shows the results of the independent dataset without the SMOTE, while (**B**) displays the results with the SMOTE on the independent dataset.

Table 3. Performance evaluation over several feature descriptors using the SMOTE, PCA, and ensemble classifier.

Method	Accuracy	Sensitivity	Specificity	MCC
Independent Dataset				
AAOA	88.88	93.87	82.92	77.14
DOA	91.11	95.91	85.36	81.46
TOA	84.44	77.55	92.68	64.55
EPseAAC (λ = 1)	89.13	92.15	85.36	77.84
EPseAAC (λ = 2)	86.66	89.79	82.92	73.13
EPseAAC (λ = 3)	86.06	89.79	82.92	73.13
AAOA + DOA	92.22	95.91	87.80	83.90
AAOA + EPseAAC (λ = 1)	90.00	93.87	85.36	79.55
AAOA + EPseAAC (λ = 2)	90.00	93.87	85.36	79.55
AAOA + EPseAAC (λ = 3)	88.88	91.83	85.36	77.53
AAOA + TOA	90.00	87.75	92.68	79.63
DOA + EPseAAC (λ = 1)	92.22	95.91	87.80	83.90
DOA + EPseAAC (λ = 2)	88.88	93.87	82.92	77.14
DOA + EPseAAC (λ = 3)	88.88	91.83	85.36	77.53
DOA + TOA	91.11	93.87	87.80	81.92
TOA + EPseAAC (λ = 1)	90.00	87.75	92.68	79.63
TOA + EPseAAC (λ = 2)	87.77	89.79	85.36	75.41
TOA + EPseAAC (λ = 3)	87.77	87.75	87.80	75.34
AAOA + DOA + TOA	92.22	93.87	90.24	84.27
AAOA + DOA + EPseAAC (λ = 1)	91.11	95.91	85.36	81.46
AAOA + DOA + EPseAAC (λ = 2)	88.88	91.83	85.36	77.53
AAOA + DOA + EPseAAC (λ = 3)	90.00	93.87	85.36	79.55
AAOA+ TOA + EPseAAC (λ = 1)	93.33	93.87	92.68	86.60
AAOA + TOA + EPseAAC (λ = 2)	88.88	91.83	85.36	77.53
AAOA + TOA + EPseAAC (λ = 3)	87.77	89.79	85.36	75.41
DOA + TOA + EPseAAC (λ = 1)	91.11	95.91	85.36	81.46
DOA + TOA + EPseAAC (λ = 2)	90.00	93.87	85.36	79.55
DOA + TOA + EPseAAC (λ = 3)	90.00	93.87	85.36	79.55
Proposed (without SMOTE)	93.75	92.00	94.87	86.87
Proposed (with SMOTE)	95.00	96.55	93.55	90.5

4.3. Comparative Analysis

This section compares the performance obtained from the proposed model on benchmark and independent datasets with current SOTA methods, which utilized the same datasets for training and testing.

4.3.1. Performance Comparison with SOTA over Benchmark Dataset

In the field of ACP and non-ACP classification, various feature extraction mechanisms and machine learning classifiers have been used by researchers. However, the performance of these models varies based on the quality of the features extracted and the machine learning algorithm utilized. For example, some prior research, such as Hajisharifi et al. [64] and Chen et al. [17], used computational-based techniques that did not sufficiently extract features, resulting in limited performance. Other studies, such as those by Hajisharifi et al. [16] and Wang [65], explored the composite peptide encoding technique but still captured redundant features, leading to some inaccuracies. Meanwhile, some researchers, such as Akbar et al. [18] and Xu et al. [19], practiced ensemble classifiers with hybrid feature spaces without optimizing the peptide information or investigating their performance. [27]. Li et al.'s lightweight model [66] minimizes the time-consuming nature of the process by taking into account low feature dimensions. Fazal et al. [67] employed a kernel sparse representation classifier for the classification of ACPs and non-ACPs.

However, They still faced a deficiency in generating more optimal features for accurate prediction in their model. In this particular scenario, researchers examined individual feature extraction components in this study. Subsequently, they assessed the hybrid models' efficacy using an ensemble classifier algorithm along with an extra optimal selection method. The proposed model showed significant improvements in classification accuracy compared to other SOTA approaches, as demonstrated in Table 4. The authors attribute this success to the deep investigation of feature extraction components, optimal selection techniques, and the use of ensemble classifiers, which helped generate more optimal features for precise prediction. The proposed model attained an amazing classification accuracy ($\lambda = 1$), representing potential for improving ACP and non-ACPs classification.

Table 4. Comparison assessing the proposed model's performance against SOTA methods on a benchmark dataset, with emphasis on highlighting the superior outcome in bold.

Model/Year	Accuracy	Sensitivity	Specificity	MCC
SPAP [64] 2013	87.00	92.00	86.00	74.0
LAK [16] 2014	92.68	89.70	85.18	78.0
iACP [17] 2016	95.06	89.86	98.54	89.0
IAP [65] 2016	93.61	89.86	96.12	86.0
iACP-GAEnsC [18] 2017	96.45	95.36	97.57	91.0
SAP [19] 2018	91.86	86.23	95.63	83.0
LDFM [66] 2020	92.73	87.70	96.10	84.0
ACP-KSRC [67] 2023	93.02	97.07	86.87	85.0
Proposed ($\lambda = 1$)	**97.56**	**97.55**	**97.54**	**95.12**

4.3.2. Performance Comparison with SOTA using Independent Datasets

We accomplished an exhaustive review of the literature and identified four research articles that assessed their models using the same dataset as ours. This allowed us to establish a fair basis for comparing our model with the existing techniques. The initial effort by Tyagi et al. [15] utilized silicon model for the classification of ACPs and non-ACPs. Ge et al. [24] proposed a novel analysis of peptide information using the chaos game representation, which generates a multidimensional feature vector that maintains the bijection property while generating a feature vector with higher dimensions. However, this technique has limitations when the sequence length is not uniform. Akbar et al. [31] proposed a cACP model to improve classification accuracy based on Geary autocorrelation, conjoint triad, and quasi-sequence alignment. Their results showed accuracy, sensitivity,

specificity, MCC, and sensitivity scores of 96.91%, 77.32%, 98.12%, and 79.0, respectively, for various classifier algorithms predicting ACPs and non-ACPs. We compared our empirical findings with those of Ahmed et al. [62], using a convolutional neural network to distinguish ACPs from non-ACPs. They did not have sufficient data to train the model, which resulted in low accuracy. In the cACP-DeepGram study, Akbar et al. [30] utilized a deep neural network (DNN) and a skip-gram-based word embedding model for ACP classification. Although their findings were efficient, additional improvements are required to improve performance.

In our paper, we introduced three models with varying λ values, and through comprehensive experiments, we found that our proposed model with $\lambda = 1$ achieved a significantly higher score than other existing approaches, as shown in Table 5. It is essential to compare our results with SOTA techniques that use similar datasets to ensure a fair comparison. Our findings highlight the effectiveness of our proposed method and demonstrate its potential for improving peptide classification accuracy.

Table 5. Comparison of the performance of our proposed model with SOTA methods using the independent dataset, highlighting the best result in bold.

Model/Year	Accuracy	Sensitivity	Specificity	MCC
NT5CT5 [15] 2013	92.65	74.67	94.44	61.0
GCGR [24] 2018	96.36	69.33	99.07	76.0
cACP [31] 2019	96.91	77.32	98.12	79.0
ACP-MHCNN [33] 2021	91.0	97.6	84.2	82.0
cACP-DeepGram [30] 2022	94.02	91.18	95.47	88.0
Proposed (λ = 1)	**95.00**	**96.55**	**93.55**	**90.05**

5. Conclusions and Future Research Direction

The metaverse exhibits significant potential for a range of medical applications, especially when combined with virtual reality (VR), augmented reality (AR), and artificial intelligence (AI). A larger number of cancer patients may benefit from this integration's potential to greatly improve medical education, advance telemedicine, encourage diversity, and boost medical literacy. This work presents a novel feature selection-based multi-voting classification algorithm specifically tailored for anti-cancer peptides. AAOA, DOA, TOA, and EPseAAC are four sequence-based feature extraction techniques that are used to identify highly discriminative characteristics that are essential for classification. The synthetic minority oversampling technique (SMOTE) algorithm is used to solve the problem of unbalanced datasets, when one class contains significantly fewer samples. By creating artificial samples for the minority class, the SMOTE balances out the unequal distribution of classes and reduces the possibility of biased categorization. An ensemble classifier is used to evaluate the extracted features, with a training set containing 80% of the data and the remaining 20% being reserved for testing. The ensemble classifier uses a voting mechanism to combine three different classifiers—support vector machine (SVM), random forest (RF), and naive Bayes (NB)—to provide the final prediction.

The suggested model is evaluated using performance metrics such as Matthews correlation coefficient (MCC), sensitivity (SN), specificity (SP), and overall accuracy (ACC). The experimental results provide a remarkable 97.56% success rate, 97.55% sensitivity, 97.54% specificity, and 95.12% MCC. These results highlight how well the suggested approach categorizes anti-cancer peptides, outperforming other methods currently used in the area. As a result, the suggested model has potential uses in academic research and fields pertaining to drugs.

There are, however, certain limitations. Due to its current focus on binary classification, the model might not function as well in situations involving multiple classifications. Furthermore, because of the limited size of the training dataset, its effectiveness may decline when faced with larger datasets. This study intends to solve these constraints in further work by concentrating on deep learning (DL) models and reviewing larger datasets. It is

expected that DL will reduce biases, improve non-expert interpretability, and guarantee the stability and dependability of machine learning-based healthcare systems.

Author Contributions: Conceptualization, S.D. and A.K.; methodology, S.D., A.K. and H.-K.S.; software, S.D.; validation, S.A., M.A. and L.M.D.; formal analysis, A.K., S.A. and M.A.; investigation, H.M. and A.K.; resources, H.M.; data curation, S.D., H.-K.S. and S.A.; writing—original draft preparation, S.D. and H.-K.S.; writing—review and editing, L.M.D., S.A., M.A. and A.K.; visualization, S.D. and H.-K.S.; supervision, H.M. and A.K.; project administration, H.M.; funding acquisition, H.M. All authors have read and agreed to the published version of the manuscript.

Funding: This work is supported by Institute of Information & communications Technology Planning & Evaluation (IITP) under the metaverse support program to nurture the best talents (IITP-2023-RS-2023-00254529) grant funded by the Korea government (MSIT) and by the Institute of Information and Communications Technology Planning and Evaluation (IITP) grant funded by the Korea government (MSIT) (No.2022-0-00106, Development of explainable AI-based diagnosis and analysis frame work using energy demand big data in multiple domains) and by Korea Institute of Planning and Evaluation for Technology in Food, Agriculture, Forestry and Fisheries (IPET) through Digital Breeding Transformation Technology Development Program, funded by Ministry of Agriculture, Food and Rural Affairs (MAFRA) (322063-03-1-SB010).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The authors have no permission to share data.

Conflicts of Interest: The authors declare no conflict of interest.

Abbreviations

AAOA	Amino acid occurrence analysis
ACC	Accuracy
ACP	Anticancer peptide
ACP-2DCNN	Anticancer peptide two-dimensional CNN
ACP-MHCNN	Anticancer peptide multi-headed CNN
AMDs	Advanced micro devices
cACP-2LFS	Classification of anticancer peptides with two-level feature selection
CCPM	Cervical cancer prediction model
CD-HIT	Cluster database at high identity with tolerance
cACP	Classifying anticancer peptides
CPUs	Central processing units
CNN	Convolutional neural network
DOA	Dipeptide occurrence analysis
DNA	Deoxyribonucleic acid
EPseAAC	Enhanced pseudo amino acid composition
MLACP	Machine learning anticancer peptide prediction
NB	Naive Bayes
NACP	Non-anticancer peptide
MCC	Mathews correlation coefficient
RF	Random forest
RTX	Ray tracing texel

References

1. Arora, T.; Grey, I. Health behaviour changes during COVID-19 and the potential consequences: A mini-review. *J. Health Psychol.* **2020**, *25*, 1155–1163. [CrossRef]
2. Maki, O.; Alshaikhli, M.; Gunduz, M.; Naji, K.K.; Abdulwahed, M. Development of digitalization road map for healthcare facility management. *IEEE Access* **2022**, *10*, 14450–14462. [CrossRef]
3. Kapoor, A.; Guha, S.; Das, M.K.; Goswami, K.C.; Yadav, R. *Digital Healthcare: The Only Solution for Better Healthcare during COVID-19 Pandemic?* Elsevier: Amsterdam, The Netherlands, 2020; Volume 72, pp. 61–64.
4. Alshamrani, M. IoT and artificial intelligence implementations for remote healthcare monitoring systems: A survey. *J. King Saud Univ. Comput. Inf. Sci.* **2022**, *34*, 4687–4701. [CrossRef]

5. Siegel, R.L.; Miller, K.D.; Wagle, N.S.; Jemal, A. Cancer statistics, 2023. *CA Cancer J. Clin.* **2023**, *73*, 17–48. [CrossRef]
6. Harris, F.; Dennison, S.R.; Singh, J.; Phoenix, D.A. On the selectivity and efficacy of defense peptides with respect to cancer cells. *Med. Res. Rev.* **2013**, *33*, 190–234. [CrossRef] [PubMed]
7. Karbalaeemohammad, S.; Naderi-Manesh, H. Two novel anticancer peptides from Aurein1. 2. *Int. J. Pept. Res. Ther.* **2011**, *17*, 159–164. [CrossRef]
8. Ijaz, M.F.; Attique, M.; Son, Y. Data-driven cervical cancer prediction model with outlier detection and over-sampling methods. *Sensors* **2020**, *20*, 2809. [CrossRef] [PubMed]
9. Saha, S.K.; Islam, S.R.; Abdullah-Al-Wadud, M.; Islam, S.; Ali, F.; Park, K.S. Multiomics analysis reveals that GLS and GLS2 differentially modulate the clinical outcomes of cancer. *J. Clin. Med.* **2019**, *8*, 355. [CrossRef]
10. Hoskin, D.W.; Ramamoorthy, A. Studies on anticancer activities of antimicrobial peptides. *Biochim. Biophys. Acta Biomembr.* **2008**, *1778*, 357–375. [CrossRef]
11. Mader, J.S.; Hoskin, D.W. Cationic antimicrobial peptides as novel cytotoxic agents for cancer treatment. *Expert Opin. Investig. Drugs* **2006**, *15*, 933–946. [CrossRef]
12. Gaspar, D.; Veiga, A.S.; Castanho, M.A. From antimicrobial to anticancer peptides. A review. *Front. Microbiol.* **2013**, *4*, 294. [CrossRef] [PubMed]
13. Huang, Y.; Feng, Q.; Yan, Q.; Hao, X.; Chen, Y. Alpha-helical cationic anticancer peptides: A promising candidate for novel anticancer drugs. *Mini Rev. Med. Chem.* **2015**, *15*, 73–81. [CrossRef] [PubMed]
14. Thundimadathil, J. Cancer treatment using peptides: Current therapies and future prospects. *J. Amino Acids* **2012**, *2012*, 967347. [CrossRef] [PubMed]
15. Tyagi, A.; Kapoor, P.; Kumar, R.; Chaudhary, K.; Gautam, A.; Raghava, G. In silico models for designing and discovering novel anticancer peptides. *Sci. Rep.* **2013**, *3*, 2984. [CrossRef] [PubMed]
16. Hajisharifi, Z.; Piryaiee, M.; Beigi, M.M.; Behbahani, M.; Mohabatkar, H. Predicting anticancer peptides with Chou′ s pseudo amino acid composition and investigating their mutagenicity via Ames test. *J. Theor. Biol.* **2014**, *341*, 34–40. [CrossRef]
17. Chen, W.; Ding, H.; Feng, P.; Lin, H.; Chou, K.-C. iACP: A sequence-based tool for identifying anticancer peptides. *Oncotarget* **2016**, *7*, 16895. [CrossRef]
18. Akbar, S.; Hayat, M.; Iqbal, M.; Jan, M.A. iACP-GAEnsC: Evolutionary genetic algorithm based ensemble classification of anticancer peptides by utilizing hybrid feature space. *Artif. Intell. Med.* **2017**, *79*, 62–70. [CrossRef]
19. Xu, L.; Liang, G.; Wang, L.; Liao, C. A novel hybrid sequence-based model for identifying anticancer peptides. *Genes* **2018**, *9*, 158. [CrossRef]
20. Bansal, G.; Rajgopal, K.; Chamola, V.; Xiong, Z.; Niyato, D. Healthcare in metaverse: A survey on current metaverse applications in healthcare. *IEEE Access* **2022**, *10*, 119914–119946. [CrossRef]
21. Tan, T.F.; Li, Y.; Lim, J.S.; Gunasekeran, D.V.; Teo, Z.L.; Ng, W.Y.; Ting, D.S. Metaverse and virtual health care in ophthalmology: Opportunities and challenges. *Asia-Pac. J. Ophthalmol.* **2022**, *11*, 237–246. [CrossRef]
22. Ali, S.; Abdullah; Armand, T.P.T.; Athar, A.; Hussain, A.; Ali, M.; Yaseen, M.; Joo, M.-I.; Kim, H.-C. Metaverse in healthcare integrated with explainable ai and blockchain: Enabling immersiveness, ensuring trust, and providing patient data security. *Sensors* **2023**, *23*, 565. [CrossRef] [PubMed]
23. Razdan, S.; Sharma, S. Internet of medical things (IoMT): Overview, emerging technologies, and case studies. *IETE Tech. Rev.* **2022**, *39*, 775–788. [CrossRef]
24. Ge, L.; Liu, J.; Zhang, Y.; Dehmer, M. Identifying anticancer peptides by using a generalized chaos game representation. *J. Math. Biol.* **2019**, *78*, 441–463. [CrossRef] [PubMed]
25. Yi, H.-C.; You, Z.-H.; Zhou, X.; Cheng, L.; Li, X.; Jiang, T.-H.; Chen, Z.-H. ACP-DL: A deep learning long short-term memory model to predict anticancer peptides using high-efficiency feature representation. *Mol. Ther. Nucleic Acids* **2019**, *17*, 1–9. [CrossRef] [PubMed]
26. Chen, X.-g.; Zhang, W.; Yang, X.; Li, C.; Chen, H. Acp-da: Improving the prediction of anticancer peptides using data augmentation. *Front. Genet.* **2021**, *12*, 698477. [CrossRef] [PubMed]
27. Ge, R.; Feng, G.; Jing, X.; Zhang, R.; Wang, P.; Wu, Q. Enacp: An ensemble learning model for identification of anticancer peptides. *Front. Genet.* **2020**, *11*, 760. [CrossRef] [PubMed]
28. Boopathi, V.; Subramaniyam, S.; Malik, A.; Lee, G.; Manavalan, B.; Yang, D.-C. mACPpred: A support vector machine-based meta-predictor for identification of anticancer peptides. *Int. J. Mol. Sci.* **2019**, *20*, 1964. [CrossRef]
29. Agrawal, P.; Bhagat, D.; Mahalwal, M.; Sharma, N.; Raghava, G.P. AntiCP 2.0: An updated model for predicting anticancer peptides. *Brief. Bioinform.* **2021**, *22*, bbaa153. [CrossRef]
30. Akbar, S.; Hayat, M.; Tahir, M.; Khan, S.; Alarfaj, F.K. cACP-DeepGram: Classification of anticancer peptides via deep neural network and skip-gram-based word embedding model. *Artif. Intell. Med.* **2022**, *131*, 102349. [CrossRef]
31. Akbar, S.; Rahman, A.U.; Hayat, M.; Sohail, M. cACP: Classifying anticancer peptides using discriminative intelligent model via Chou's 5-step rules and general pseudo components. *Chemom. Intell. Lab. Syst.* **2020**, *196*, 103912. [CrossRef]
32. Akbar, S.; Hayat, M.; Tahir, M.; Chong, K.T. cACP-2LFS: Classification of anticancer peptides using sequential discriminative model of KSAAP and two-level feature selection approach. *IEEE Access* **2020**, *8*, 131939–131948. [CrossRef]
33. Ahmed, S.; Muhammod, R.; Khan, Z.H.; Adilina, S.; Sharma, A.; Shatabda, S.; Dehzangi, A. ACP-MHCNN: An accurate multi-headed deep-convolutional neural network to predict anticancer peptides. *Sci. Rep.* **2021**, *11*, 23676. [CrossRef]

34. Ghulam, A.; Ali, F.; Sikander, R.; Ahmad, A.; Ahmed, A.; Patil, S. ACP-2DCNN: Deep learning-based model for improving prediction of anticancer peptides using two-dimensional convolutional neural network. *Chemom. Intell. Lab. Syst.* **2022**, *226*, 104589. [CrossRef]
35. Park, H.W.; Pitti, T.; Madhavan, T.; Jeon, Y.-J.; Manavalan, B. MLACP 2.0: An updated machine learning tool for anticancer peptide prediction. *Comput. Struct. Biotechnol. J.* **2022**, *20*, 4473–4480.
36. Chen, J.; Cheong, H.H.; Siu, S.W. xDeep-AcPEP: Deep learning method for anticancer peptide activity prediction based on convolutional neural network and multitask learning. *J. Chem. Inf. Model.* **2021**, *61*, 3789–3803. [CrossRef]
37. Sun, M.; Yang, S.; Hu, X.; Zhou, Y. ACPNet: A deep learning network to identify anticancer peptides by hybrid sequence information. *Molecules* **2022**, *27*, 1544. [CrossRef] [PubMed]
38. Onan, A.; Korukoğlu, S.; Bulut, H. A multiobjective weighted voting ensemble classifier based on differential evolution algorithm for text sentiment classification. *Expert Syst. Appl.* **2016**, *62*, 1–16. [CrossRef]
39. Dimitriadou, E.; Weingessel, A.; Hornik, K. A cluster ensembles framework. In *Design and Application of Hybrid Intelligent Systems*; IOS Press: Amsterdam, The Netherlands, 2003.
40. Khan, S.U.; Haq, I.U.; Khan, Z.A.; Khan, N.; Lee, M.Y.; Baik, S.W. Atrous Convolutions and Residual GRU Based Architecture for Matching Power Demand with Supply. *Sensors* **2021**, *21*, 7191. [CrossRef]
41. Khan, S.U.; Khan, N.; Ullah, F.U.M.; Kim, M.J.; Lee, M.Y.; Baik, S.W. Towards intelligent building energy management: AI-based framework for power consumption and generation forecasting. *Energy Build.* **2023**, *279*, 112705. [CrossRef]
42. Hussain, A.; Khan, S.U.; Rida, I.; Khan, N.; Baik, S.W. Human Centric Attention with Deep Multiscale Feature Fusion Framework for Activity Recognition in Internet of Medical Things. *Inf. Fusion* **2023**, 102211. [CrossRef]
43. Hussain, A.; Khan, S.U.; Khan, N.; Shabaz, M.; Baik, S.W. AI-driven behavior biometrics framework for robust human activity recognition in surveillance systems. *Eng. Appl. Artif. Intell.* **2024**, *127*, 107218. [CrossRef]
44. Hussain, A.; Amin, S.U.; Lee, H.; Khan, A.; Khan, N.F.; Seo, S. An Automated Chest X-Ray Image Analysis for Covid-19 and Pneumonia Diagnosis using Deep Ensemble Strategy. *IEEE Access* **2023**, *11*, 97207–97220. [CrossRef]
45. Ekbal, A.; Saha, S. Weighted vote based classifier ensemble selection using genetic algorithm for named entity recognition. In Proceedings of the International Conference on Application of Natural Language to Information Systems, Cardiff, UK, 23–25 June 2010; pp. 256–267.
46. Opitz, D.; Maclin, R. Popular ensemble methods: An empirical study. *J. Artif. Intell. Res.* **1999**, *11*, 169–198. [CrossRef]
47. Hayat, M.; Khan, A.; Yeasin, M. Prediction of membrane proteins using split amino acid and ensemble classification. *Amino Acids* **2012**, *42*, 2447–2460. [CrossRef]
48. Chen, W.; Ding, H.; Zhou, X.; Lin, H.; Chou, K.-C. iRNA (m6A)-PseDNC: Identifying N6-methyladenosine sites using pseudo dinucleotide composition. *Anal. Biochem.* **2018**, *561*, 59–65. [CrossRef] [PubMed]
49. Fu, L.; Niu, B.; Zhu, Z.; Wu, S.; Li, W. CD-HIT: Accelerated for clustering the next-generation sequencing data. *Bioinformatics* **2012**, *28*, 3150–3152. [CrossRef]
50. Khan, S.U.; Khan, N.; Hussain, T.; Baik, S.W. An intelligent correlation learning system for person Re-identification. *Eng. Appl. Artif. Intell.* **2024**, *128*, 107213. [CrossRef]
51. Dilshad, N.; Khan, S.U.; Alghamdi, N.S.; Taleb, T.; Song, J. Towards Efficient Fire Detection in IoT Environment: A Modified Attention Network and Large-Scale Dataset. *IEEE Internet Things J.* **2023**. [CrossRef]
52. Nguyen, T.N.; Nguyen-Xuan, H.; Lee, J. A novel data-driven nonlinear solver for solid mechanics using time series forecasting. *Finite Elem. Anal. Des.* **2020**, *171*, 103377. [CrossRef]
53. Nguyen, T.N.; Lee, S.; Nguyen, P.-C.; Nguyen-Xuan, H.; Lee, J. Geometrically nonlinear postbuckling behavior of imperfect FG-CNTRC shells under axial compression using isogeometric analysis. *Eur. J. Mech. A/Solids* **2020**, *84*, 104066. [CrossRef]
54. Chawla, N.V.; Bowyer, K.W.; Hall, L.O.; Kegelmeyer, W.P. SMOTE: Synthetic minority over-sampling technique. *J. Artif. Intell. Res.* **2002**, *16*, 321–357. [CrossRef]
55. Khan, S.U.; Khan, N.; Hussain, T.; Muhammad, K.; Hijji, M.; Del Ser, J.; Baik, S.W. Visual Appearance and Soft Biometrics Fusion for Person Re-identification using Deep Learning. *IEEE J. Sel. Top. Signal Process.* **2023**, *17*, 575–586. [CrossRef]
56. Chou, K.-C. Impacts of bioinformatics to medicinal chemistry. *Med. Chem.* **2015**, *11*, 218–234. [CrossRef] [PubMed]
57. Khan, S.U.; Baik, R. MPPIF-Net: Identification of Plasmodium Falciparum Parasite Mitochondrial Proteins Using Deep Features with Multilayer Bi-directional LSTM. *Processes* **2020**, *8*, 725. [CrossRef]
58. Cheng, X.; Xiao, X.; Chou, K.-C. pLoc-mGneg: Predict subcellular localization of Gram-negative bacterial proteins by deep gene ontology learning via general PseAAC. *Genomics* **2018**, *110*, 231–239. [CrossRef]
59. Waris, M.; Ahmad, K.; Kabir, M.; Hayat, M. Identification of DNA binding proteins using evolutionary profiles position specific scoring matrix. *Neurocomputing* **2016**, *199*, 154–162. [CrossRef]
60. Huang, T.; Chen, L.; Cai, Y.-D.; Chou, K.-C. Classification and analysis of regulatory pathways using graph property, biochemical and physicochemical property, and functional property. *PLoS ONE* **2011**, *6*, e25297. [CrossRef]
61. Behbahani, M.; Mohabatkar, H.; Nosrati, M. Analysis and comparison of lignin peroxidases between fungi and bacteria using three different modes of Chou's general pseudo amino acid composition. *J. Theor. Biol.* **2016**, *411*, 1–5. [CrossRef]
62. Meher, P.K.; Sahu, T.K.; Saini, V.; Rao, A.R. Predicting antimicrobial peptides with improved accuracy by incorporating the compositional, physico-chemical and structural features into Chou's general PseAAC. *Sci. Rep.* **2017**, *7*, 42362. [CrossRef]
63. Chou, K.-C. An unprecedented revolution in medicinal chemistry driven by the progress of biological science. *Curr. Top. Med. Chem.* **2017**, *17*, 2337–2358. [CrossRef]

64. Hajisharifi, Z.; Mohabatkar, H. In silico prediction of anticancer peptides by TRAINER tool. *Mol. Biol. Res. Commun.* **2013**, *2*, 39–45.
65. Li, F.-M.; Wang, X.-Q. Identifying anticancer peptides by using improved hybrid compositions. *Sci. Rep.* **2016**, *6*, 33910. [CrossRef] [PubMed]
66. Li, Q.; Zhou, W.; Wang, D.; Wang, S.; Li, Q. Prediction of anticancer peptides using a low-dimensional feature model. *Front. Bioeng. Biotechnol.* **2020**, *8*, 892. [CrossRef] [PubMed]
67. Fazal, E.; Ibrahim, M.S.; Park, S.; Naseem, I.; Wahab, A. Anticancer Peptides Classification Using Kernel Sparse Representation Classifier. *IEEE Access* **2023**, *11*, 17626–17637. [CrossRef]