**KANampyl: A Novel Kolmogorov-Arnold Network Framework for Accurate Prediction of AMPylation Sites in Fic Domain Proteins**

Tasmin Karim1, †, Md. Shazzad Hossain shaon1, †, Md. Fahim Sultan1, Sayed Mehedi Azim2, Iman Dehzangi2,3, Md. Zahid Hasan4,, Mst Shapna Akter1, Ahmed Moustafa6,7

1. Department of Computer Science and Engineering, Oakland University, Rochester, MI 48309, USA.
2. Center for Computational and Integrative Biology, Rutgers University, Camden, NJ 08102, USA.
3. Department of Computer Science, Rutgers University, Camden, NJ, 08102, USA
4. Health Informatics Research Lab, Department of Computer Science and Engineering, Daffodil International University, Daffodil Smart City, Birulia, Dhaka-1314, Bangladesh
5. Department of Human Anatomy and Physiology, The Faculty of Health Sciences, University of Johannesburg, Johannesburg, South Africa.
6. School of Psychology, Centre for Data Analytics, Bond University, Gold Coast, QLD, Australia.

†. This author contributed equally to the work and are co-first authors.

**Email Addresses:**

[tasminkarim@oakland.edu](mailto:tasminkarim@oakland.edu),

[shaon@oakland.edu](mailto:shaon@oakland.edu),

[mdfahimsultan@oakland.edu](mailto:mdfahimsultan@oakland.edu),

[sayedmehedi.azim@rutgers.edu](mailto:sayedmehedi.azim@rutgers.edu),

[i.dehzangi@rutgers.edu](mailto:i.dehzangi@rutgers.edu) ,

[zahid.cse@diu.edu.bd](mailto:zahid.cse@diu.edu.bd),

[akter@oakland.edu](mailto:akter@oakland.edu),

[ahmedhalimo@gmail.com](mailto:ahmedhalimo@gmail.com)

**Abstract:**

Post-translational modifications are critical for regulating cellular processes and protein functionality. AMPylation is a newly identified post-translational modification that involves attaching adenosine monophosphate to proteins, potentially regulating their activity. Recent studies reveal that this post-translational modification plays a direct role in regulating neurodevelopment, neurodegeneration, and various physiological processes. Despite its biological importance, computational approaches of AMPylation site prediction have faced challenges due to suboptimal feature extraction and validation techniques. In this study we introduce a novel computational framework, Kolmogorov-Arnold Network Framework for Accurate Prediction of AMPylation( KANampyl), employing the Kolmogorov-Arnold Network model for AMPylation site prediction. KANampyl leverages diverse feature extraction techniques, including Z-Scale index (Z-scale), Amino Acid Composition (AAC), BLOSUM matrix types feature encoding (BLOSUM62), Overlapping Property Features - 10 bit (OPF\_10bit Auto-Encoded Secondary Structure Neural Network 3 (AESNN3), and merged features, with 10-fold cross-validation to ensure prediction robustness. Our comprehensive analysis reveals that Z-scale features consistently outperformed other feature sets, achieving 92.33% accuracy, 86.04% sensitivity, 95.92% specificity, and an AUC score of 90.98%, significantly surpassing existing state-of-the-art models for this task. To promote accessibility and facilitate further research, we made KANampyl publicly available and can be accessed at: [**https://github.com/Shazzad-Shaon3404/kan**](https://github.com/Shazzad-Shaon3404/kan)

**Keywords:** Post-translational modifications; AMPylation, Kolmogorov-Arnold Network; Amino Acid Composition.

**Introduction**

Post-translational modification (PTM) is the catalytic or pharmacological modification of an amino acid while it has been produced on the ribosome. These modifications are made by removing protein categories, irreversibly combining distinct chemical pairs, as well degrading changed peptides [1,2]. PTMs are essential for comprehending cellular activities and biological processes, especially cellular interactions and stiffness [3]. PTMs serve as crucial processes for increasing proteome diversity and represent a major part of functional genomics by regulating protein operation, translation, and relationships with other parts of the cell such as amino acids, lipids, nucleic acids, and cofactors [4]. These alterations can modify the structure, ionization potential, and dealings of proteins, adjusting retracting, binding to specific cellular compartments, ligand or protein interactions, and changes in operational states such as signaling or catalysis [5]. Several PTMs have been discovered and identified, including phosphorylation, glycosylation, ubiquitination, nitrosylation, methylation, acetylation, lipidation, and proteolysis, all of which influence practically every aspect of cell biology and pathology [6]. AMPylation is a recently identified post-translational modification aided by a bacterial virulence factor that triggers the exchange of adenosine monophosphate (AMP) from adenosine triphosphate (ATP) to the hydroxyl group of a threonine residue on eukaryotic targeted proteins [7,8]. AMPylation involves the covalent addition of an AMP molecule to a protein [9]. AMPylation has been explored especially in combination with Fic domain peptides, which have been conserved through evolution and found in several organisms, from microorganisms to mammals. These enzymes operate in both bacterial pathogenicity and eukaryotic signaling pathways through the attachment of AMP to Rho-family GTPases. [10,11]. Identifying ampylated proteins is crucial for improving the comprehension of cellular physiology, and disease processes. Experimental approaches for detecting PTM sites are frequently difficult, costly, and time-consuming [13-23].

* 1. **Prior Models and algorithms and limitations**

As a result, various research has proposed computer algorithms as faster and less expensive options for predicting PTM sites. In 2015, Khater, et al. applied Support Vector Machine (SVM) and Hidden Markov Model (HMM) approaches to build a computational framework for detecting AMPylation domains. Furthermore, they divided these domains into functional subfamilies, including those that catalyze AMPylation, deAMPylation, phosphorylation, and phosphocholine transfer. However, their method did not predict whether a particular peptide displayed AMPylation or non-AMPylation properties [24]. In 2022, Azim et al. implemented the DeepAMP framework, that incorporates Convolutional Neural Network (CNN) approaches to improve AMPylation detection. Despite their endeavors, the dataset used for this analysis reveals that there is potential to improve AMPylation detection by exploring other feature extraction approaches [3]. Swati et al. developed the XGboost-Ampy model using machine learning techniques in the same year, 2022. Their methodology included 5-fold cross-validation and only used two feature extraction methods. However, considering the variety of feature extraction options available and the demand for robust validation, a more thorough examination of feature extraction methods, together with the use of 10-fold cross-validation, would most certainly improve the model's performance [25]. In 2024, Prabhu et al. employed machine learning techniques to identify AMPylation using several feature extraction methods. Their Random Forest (RF)-based model reached 80% accuracy. Despite this development, the model's performance still has a lot of room for improvement [26].

**1.2 The current study**

To the best of our knowledge, few computational approaches have been proposed for detecting AMPylation sites, particularly within Fic domain proteins. Most studies have not achieved sufficient performance in this area, largely due to the limited range of feature extraction methods employed. There is a clear need for more comprehensive and robust methodologies to improve the accuracy and reliability of AMPylation site predictions. Therefore, this study aims to enhance performance on similar datasets by employing a novel Kolmogorov-Arnold Network (KAN) approach called the “KANampyl” framework to more accurately detect AMPylation. We utilized various feature extraction methods to achieve this goal, aiming to improve the accuracy and reliability of AMPylation site predictions. In this work, we have explored the KAN model as compared to other neural network-based models which deal with fixed approximating functions, it establishes continuous functions through sums of univariate functions, which is expected to make them more accurate. This approach obtained 92.33% accuracy, 86.04% sensitivity, 95.92% specificity, and an AUC score of 90.98%. These results outperform the currently available state-of-the-art models where no other predictors are capable of more than 90% accuracy with harmonious specificity and sensitivity. We are optimistic that this article would enormously facilitate researchers to overcome the current research deficiencies in this area of study. Our presented dataset and the standalone predictor, KANampyl, are now accessible to the public, offering valuable resources for further research and development. In the following sections, we describe the overall working flows of the study for the identification of AMPylation.

**2. Methods and materials**

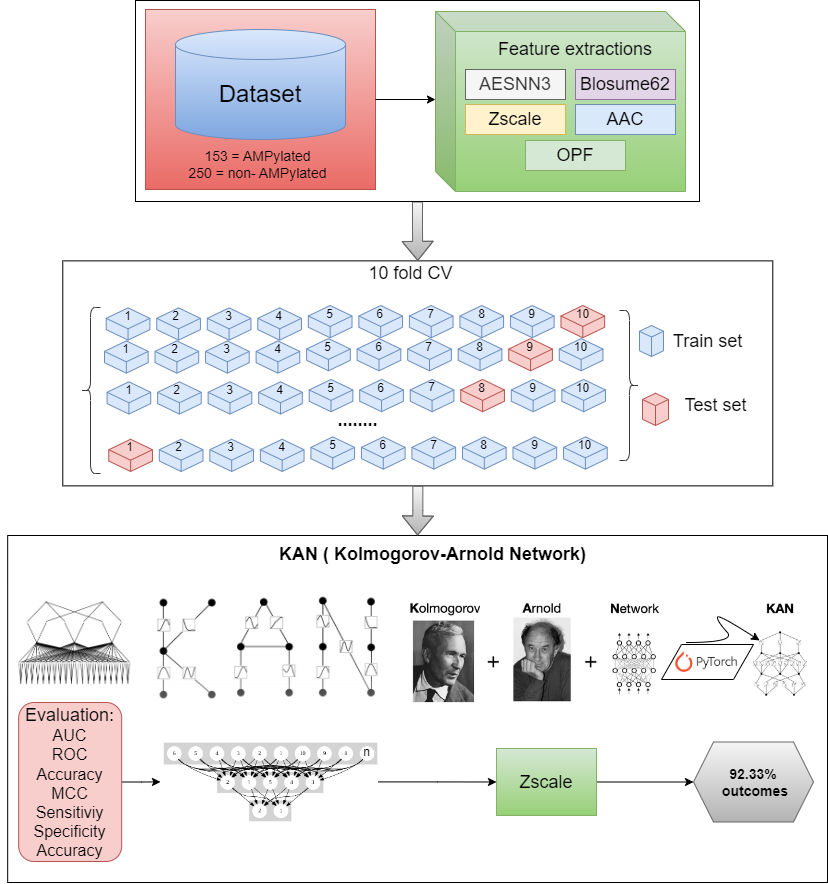
This section provides a thorough overview of the experimental framework, including the study's analysis, data-gathering methods, experimental strategy, and feature optimization. It describes the analytical procedures utilized to perform the research, providing an in-depth understanding of the methodology used to acquire the results discussed in the research project.

**2.1 Dataset collection**

The dataset used in this analysis is similar to that used by Azim et al. [3], which was derived from Kielkowski et al.'s [10] experiments. The authors of the current article gathered the data by finding AMPylation in intact cancer cells using LC-MS/MS and imaging methods. Using a pronucleotide probe, scientists measured protein AMPylation in livinge cells, treating different cell lines such as HeLa and SH-SY5Y to find sites. The found AMPylated proteins have involvement in a variety of metabolic processes, including glycolysis regulation, proteolysis, PTM modulation, and UPR. A total of 162 protein sequences implicated in this alteration were found, with the dataset containing around 133 distinct protein sequences from the UniProt database. Azim et al. [3] employed the Cluster Database at High Identity with Tolerance (CD-Hit) for excluding proteins with more than 40% consecutive similarities to minimize repetition from the set of proteins. It is a widely used tool for clustering tremendous biological sequence datasets to prevent duplication and still preserve sequences that indicate different clusters [27]. As a result, they found 153 AMPylated and 250 non-AMPylated sites. Following this precedence, the current study used 153 positive and 250 negative samples to ensure similarity in dataset foundation.

**2.2 Overview of the study**

The study focuses on developing an effective approach for finding AMPylation sites using a variety of feature extraction techniques. **Figure 1** shows the analysis's strategy, which includes applying five different types of feature extraction methods and exploring them separately in addition to working together. The merging of features is intended for determining the most ideal feature collection. All feature extraction techniques are plugged into the KAN with 10-fold cross-validation, which has been selected because of its capacity for adaptation to smaller collections of data. Validation measures have been employed to acquire the best feature encoded form, which brings about better performance in detecting AMPylation sites.



**Figure 1.** Represents the study's procedure: datasets were obtained from existing materials feature extraction methods were utilized, and 10-fold cross-validation was executed on each feature set and the amalgamated properties. An experimental approach, KAN, was used for accurate prediction, and feature engineering techniques were explored to improve AMPylation prediction.

**2.3. Feature engineering**

Biological sequences might be extremely lengthy, generating high-dimensional data. Feature encoding approaches prevent dimensionality, making statistics less difficult to organize and algorithms more computationally beneficial. Therefore, this study used five distinct features encoding methods: Residue composition features, Auto-Encoded Secondary Structure Neural Network 3 (AESNN3), and Overlapping Property Features - 10 bit (OPF\_10bit) [28-30]. BLOSUM matrix types feature encoding: BLOSUM62 [31]. Z-Scale index-based Z-Scale features [32], and Amino acid composition-based features called Amino Acid Composition (AAC) for the detection of AMPylation [33]. These features can speed up the development in bioinformatics and computational biology, for analyzing and predicting various biochemical properties, including ampylation, a post-translational modification involving the transfer of an AMP group to a protein [34, 35].

**2.3.1 Auto-Encoded Secondary Structure Neural Network 3(AESNN3)**

AESNN3 is a feature representation approach that uses sequence alignments to learn about protein evolution and structure. It uses autoencoders and neural networks to extract useful patterns from multiple sequence alignments (MSA) [28,29]. The general equation for AESNN3 feature extraction is expressed as follows:

(1)

Here, x is the input sequence alignment features, W means weight matrix and b is the bias, and A is the activation function, which helps encode relevant structural properties.

**2.3.2 Overlapping Property Features - 10 bit (OPF\_10bit)**

OPF\_10bit is a feature encoding technique used in bioinformatics to describe peptide or protein sequences using overlapping physicochemical patterns. This method splits a sequence into overlapping pieces of 10 residues and encodes each fragment with numerical property values such as hydrophobicity, charge, or structural inclinations [30]. It can be mathematically expressed as follows:

(2)

Where, represents the physicochemical properties at position *i*. The addition of incorporates the data across a 10-residue window to capture local sequence information.

**2.3.3** **Blosume Matrix (BLOSUM62)**

The BLOSUM62 matrix encodes protein sequences by displaying each amino acid as a 21-dimensional vector (20 amino acids plus 1 terminal signal). A sequence window of size (*n*) results in a (*21 times n* ) matrix. Each row represents a normalized BLOSUM62 substitution score, which captures evolutionary connects and assists machine learning models classify peptides and anticipate their functions [31].

**2.3.4 Z-Scale features (Z-Scale)**

Z-Scale Features are based on principal component analysis (PCA) of several molecular characteristics, including hydrophobicity, steric bulk, and electronic effects. Each amino acid is represented by a five-dimensional *(Z1-Z5)* vector, which captures important biological features [32].

**2.3.5 Amino Acid Composition (AAC)**

The AAC determines the standardized amounts for each amino acid sequence. It summarizes the proportions of every peptide [33]. The calculation is as follows:

(3)

Here, k denotes the amino acids, is the length of the sequences and number pf the amino acids.

**2.4 Proposed model construction procedures**

**2.4.1 KANampyl model development**

In 2024, Liu et al. introduced a model known as KAN [36]. The distinctive structure provides an improved method for function approximation, driven according to the Kolmogorov-Arnold representation theorem. KAN has trainable activation functions on edges, fundamentally altering the neural network's structure. This particular feature eliminates linear weight matrices, replacing them with trainable 1D spline functions. That deviation from common patterns allows KAN to successfully incorporate the benefits of splines [37]. Current research on AMPylation detection has demonstrated poor efficiency. This work overcomes this challenge by using PyTorch and a hyperparameter adjusting method [38-42]. Our specific technique outperforms previous endeavors. In bellowing there is a brief explanation of how the KANampyl model handled our data with forward and backward passes. The working principles of the KANampyl model’s mathematical formulation can be stated as:

(4)

Initially, the dataset is read into a data frame and converted features and labels into PyTorch tensors, then we create data loader objects to handle batching and shuffling. denoted the input data, and represents the target classes. is the shape of the datasets, where *b* is the samples in the batch and *n* is the number of features per sample and *B* is the data loader function. These features were fed into the models for the detection of the classes.

(5)

From the equation 2, the model was done with a forward pass for moving input data through the network, is the overall framework, *a* is the input vector, represents as the transformation matrix function, where the matrix layer denotes as *p*. are non-linier transformation matrices.

(6)

From equation 3, here, is the learnable function, is the sigmoid function, where the outcome would be 0 or 1, is another learnable function, where *S* denotes a trainable *B-spline* function, where, are represents the place of the control points.

Following that, to improve the performances of the KAN-based models, the current study employed the Adam optimizer, which maintains a per-parameter learning rate that changes as learning progresses [43-46]. Accordingly, to reduce the gap between the anticipated class probabilities and the actual class labels, we used the loss function as cross-entropy [47-50]. Then backward pass propagation for compute the gradients of loss values for the model parameter [51]. The equation can be expressed as:

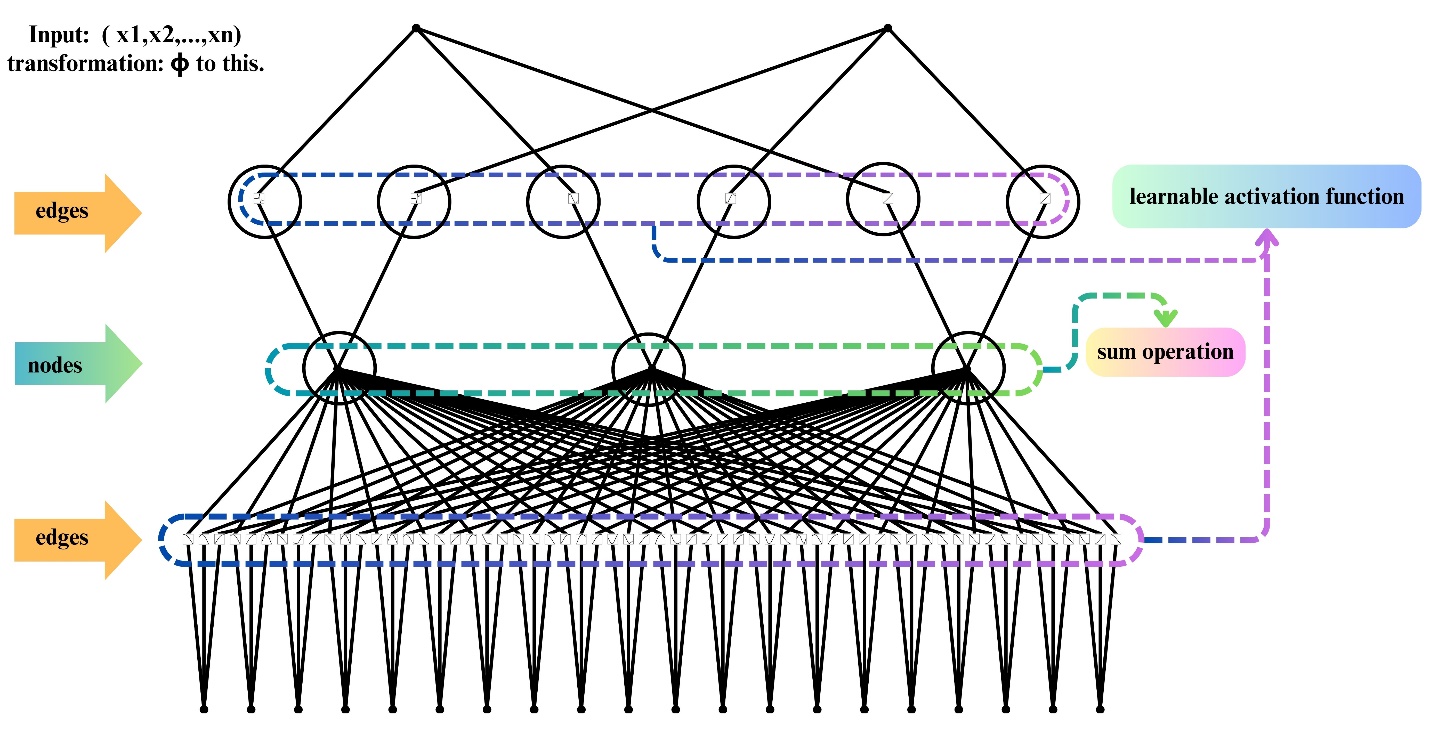
(7)

(8)

Where, formula 4, denotes the Adam optimization. Here, is the parameters of the model at the time step is *t*. denotes as the learning rate, represents the first-moment estimation bias-correctness.is the second-moment estimation bias-correctness. denotes a small constant that prevents division by zero. Accordingly, equation 5 refers to the loss function, where, *c* is the number of classes, represents the true label (1 if the samples belong to class *i*, otherwise 0), and denotes the probability values for class *i*.

**2.4.2. Overall outcome procedure of KANampyl**

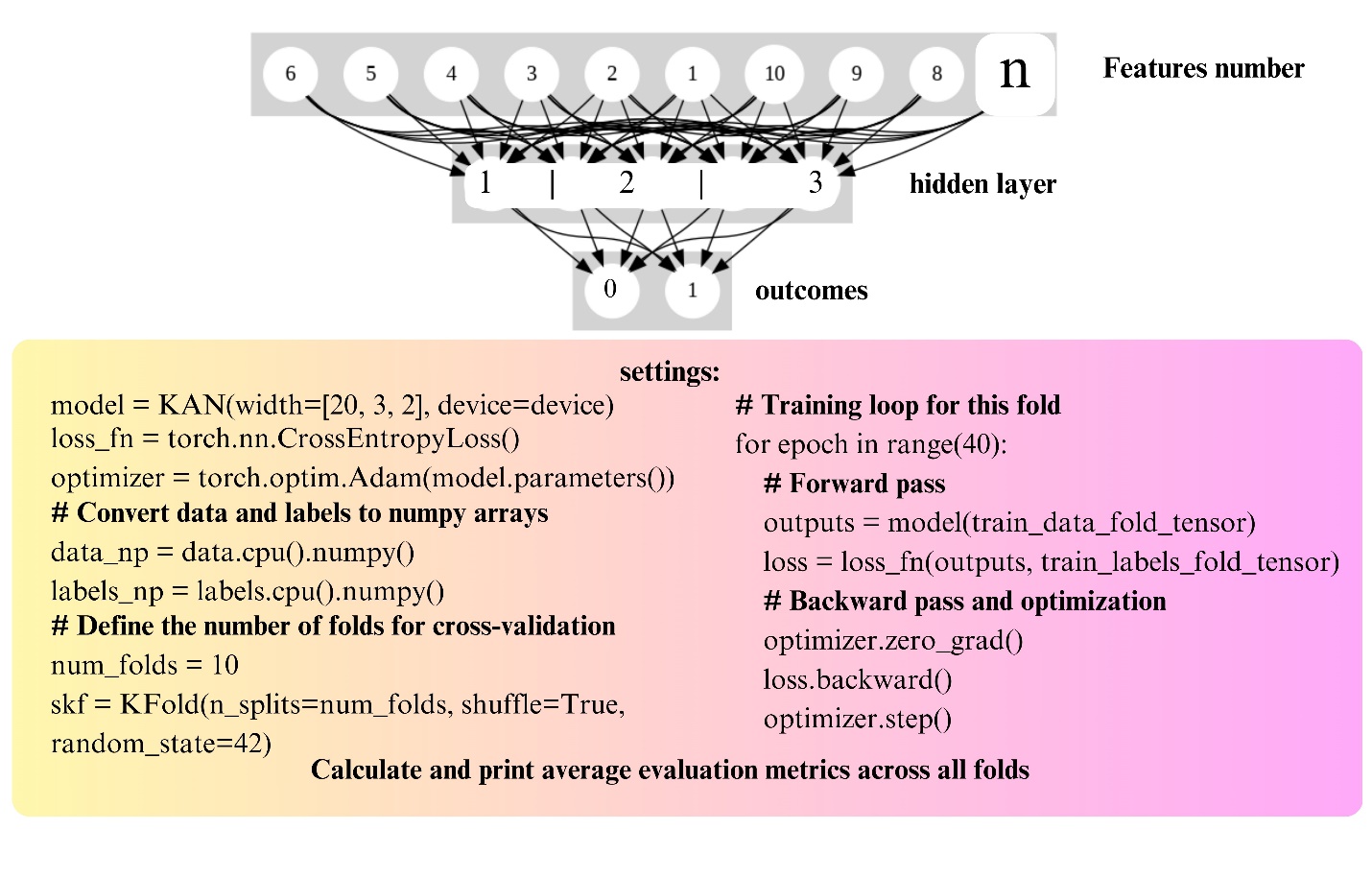
Figure 2 highlights the finalized pattern of the KANampyl model, which is based on the KAN theorem [36, 37]. In contrast to standard neural networks, the KAN model places activation functions on the edges, which correspond to the weights, while all operations take place at the nodes. Therefore, KAN has been projected to be more reliable and understandable with fewer variables. According to the structure, a single KAN layer takes input (*x1, x2, ⋯, xn*) and implements a modification *ϕ* to output.



**Figure 2**: illustrates the general arrangement of the approach to KAN, with all edges having an adaptable activation parameter while all nodes execute the averaging procedure.

**2.4.3. Summary of the KANampyl framework**

Figure 3 demonstrates the study's overall layer structure and parameter settings. In this configuration, we build the model utilizing the KAN architecture. The width of the model is defined by the number of features, hidden layers, and output layers. For instance, widths of 20, 3, and 2 correspond to 20 features, 3 hidden layers, and 2 output layers, with 0 indicating a negative outcome and 1 indicating an acceptable outcome. As we have already stated we used a loss function and an optimizer, which were set based on the figures. The research then splits the dataset into ten equal pieces (folds). In every session, one-fold is utilized as a validation set, with the remaining nine folds used for training. This operation is repeated ten times, with each fold used precisely once as the validation set. It ensures a robust assessment of the model by decreasing the variance associated with a single train-test split. It also gives a thorough knowledge of model performance across diverse data subsets. In the cross-validation, we used the shuffling function that the data before splitting guarantees that each fold represents the whole dataset, eliminating bias resulting from data order. Accordingly, we set a random seed number is 42 for ensure that the data splits remain the same. Furthermore, each epoch represents one entirely operated through the whole training dataset. Training for 40 epochs suggests the model observes every data point forty times. The forward and backward passes then execute appropriate optimization environments, ensuring that the model parameters are successfully informed which results in convergent around a minimal loss. Lastly, the study employed a variety of evaluation criteria to conduct the analyses.



**Figure 3**: Overall summary of the KANampyl approach, where the model was conducted with features number, 3 hidden layers, and outcome layers.

**2.5. Evaluation metrics settings**

To evaluate the model’s effectiveness, we employed several key metrics: Accuracy, Sensitivity (Sn), Specificity (Sp), Matthews Correlation Coefficient (MCC), Area under curve (AUC), Recall (Rec) F1 Score (F1), and Precision (Pre). These metrics provided a comprehensive quantitative assessment of the model’s performance. In this context, TP, TN, FP, and FN represent true positives, true negatives, false positives, and false negatives, respectively [52,53]. The mathematical formulas for these metrics are as follows:

(9)

**3. Experimental results**

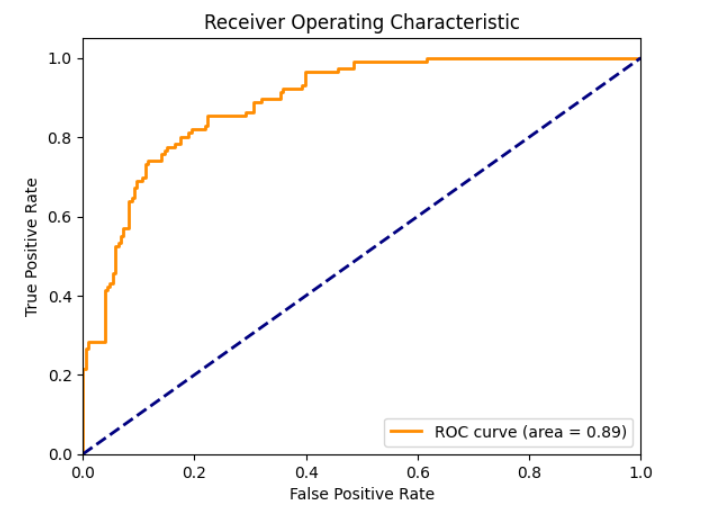
The study performed an extensive study to determine AMPylation from the PTM sequences. Table 1 represents the comprehensive exploratory findings from the current study, which were performed using the 10-fold cross-validation approach.

**Table 1.** Overall performance of the KANampyl model for recognizing AMPylation sites.

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Features** | **Mode** | **Accuracy** | **Sn** | **Sp** | **AUC** | **MCC** | **Pre** | **F1** | **Rec** |
| AAC | Training | 0.8881 | 0.8706 | 0.8980 | 0.9514 | 0.7607 | 0.8278 | 0.8487 | 0.8706 |
| Testing | 0.8395 | 0.7567 | 0.9090 | 0.8808 | 0.6730 | 0.8750 | 0.8115 | 0.7567 |
| **Z-scale** | **Training** | **0.9460** | **0.8934** | **0.9756** | **0.9345** | **0.8786** | **0.9464** | **0.9323** | **0.9345** |
| **Testing** | **0.9233** | **0.8604** | **0.9592** | **0.9098** | **0.8240** | **0.9152** | **0.9075** | **0.9098** |
| Blosume 62 | Training | 0.9150 | 0.9269 | 0.9083 | 0.9176 | 0.8224 | 0.9071 | 0.9021 | 0.9176 |
| Testing | 0.8678 | 0.8209 | 0.8974 | 0.8591 | 0.7096 | 0.8535 | 0.8385 | 0.8591 |
| OPF | Training | 0.8266 | 0.4157 | 0.9989 | 0.7568 | 0.6130 | 0.8932 | 0.7490 | 0.7568 |
| Testing | 0.7590 | 0.4594 | 0.9471 | 0.7032 | 0.4559 | 0.7719 | 0.6803 | 0.7032 |
| AESNN3 | Training | 0.9118 | 0.7703 | 0.9903 | 0.8803 | 0.7849 | 0.8946 | 0.8803 | 0.8692 |
| Testing | 0.8368 | 0.6335 | 0.9645 | 0.7990 | 0.6145 | 0.7774 | 0.7759 | 0.7990 |
| Merged features | Training | 0.7283 | 0.2461 | 1.00 | 0.6230 | 0.4079 | 0.8515 | 0.6094 | 0.6230 |
| Testing | 0.6988 | 0.1975 | 0.9858 | 0.5917 | 0.2840 | 0.7258 | 0.5535 | 0.5917 |

Table 1 shows how each feature set performed on AAC, Z-scale, OPF, AESNN3, and merged features throughout both the training and testing stages. Overall, the study concluded that Z-scale features are the most appropriate and best-fitting feature embedding strategy for the KANampyl model. Other feature extractions are also performed splendidly according to the feature structure, where most of the features have the ability the detect AMPylation more than 80%, especially AAC features acquired 83.95%, Blosume62 derived 86.78%, and AESNN3 features produced 83.68% on the testing purpose respectively. Considering OPF and merged features the applied models generated insufficient results with 75.90% and 93.88% accuracy. Throughout the training and testing stages, the Z-scale features consistently delivered the best results across all criteria. This higher performance is due to the Z-scale's capacity of amino acids, which are essential for detecting AMPylation sites. The Z-scale features most likely provide a richer and more discriminative representation of the sequences, resulting in improved accuracy (0.9460 in training and 0.9233 in testing) and resilience (high AUC of 0.9345 and 0.9098, respectively). The AAC and Blosum62 features performed well, although they were not as effective as the Z-scale features. The OPF and AESNN3 features performed acceptably, however, had reduced sensitivity, indicating that they might be missing a considerable percentage of genuine positives. The merged features strategy did not outperform separate features, implying that the combination could have contributed noise or duplicated information, resulting in worse performance measures. Therefore, the study concluded that the Z-scale features are the most appropriate of the KANampyl model.

Figure 5 shows the Receiver Operating Characteristic (ROC) curve for KANampyl models with Z-scale features. The True Positive Rate (TPR) is represented on the Y axis, and the False Positive Rate (FPR) with the X axis. Our analysis of the KANampyl model implies phenomenal results on every evaluation criterion. The model's probabilistic outputs performed satisfactorily during testing, as shown in the ROC curves. These graphs demonstrate that the model properly identifies more than 85% of the actual classes, with a low percentage of false positive predictions. The high AUC score, as well as the model's ability to hold onto a high TPR and a low FPR, indicates that the KANampyl model based on Zscscale features not only delivers high accuracy but also efficiently differentiates between positive and negative classes. This performance demonstrates the model's resilience and dependability in prediction operations, which means that it works effectively in a variety of configurations and information.



**Figure 5:** KANampyl model’s ROC curves visualization based on Z-scale features.

**4. Discussion**

Detecting AMPylation using computational approaches is vital considering its fundamental significance in regulating different cellular processes and its relationship with a variety of diseases. The major benefit of computational detection is its capacity to quickly and reliably anticipate AMPylation sites, which is necessary when investigating protein function and interaction. The high accuracy of these predictions facilitates the unambiguous determination of actual AMPylation sites, lowering false positives and negatives, making them crucial for downstream experimental validation. Furthermore, identifying important variables that contribute to AMPylation improves our understanding of the underlying biological mechanisms, allowing for the creation of targeted therapeutics and diagnostics. Overall, computational approaches are a cost-effective, efficient, and exact approach to researching AMPylation, which advances both fundamental and practical biomedical research. As a result, concentrating on the benefits of computational approaches, this study developed the KANampyl model utilizing the KAN method. KAN is an innovative technique, and we are the first to investigate its applicability to biological genome sequences for AMPylation detection. Previous research on KAN has proved its potential to successfully gather data points; based on this validation, we combined KAN with specified parameters and a robust feature extraction approach for AMPylation detection. According to our findings, this strategy outperforms modern technologies in terms of accuracy and dependability. This highlights the KAN method's potential to considerably enhance computer detection tools in scientific study. In Table 2, we compared our model with existing predictors.

**Table 2.** Comparison of the state-of-the-art with the KANampyl model

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Applied Models** | **Year** | **Features** | **Effective Features** | **Accuracy** | **MCC** | **AUC** | **Sn** | **Sp** |
| ANN [26] | 2024 | mat,counts, Pro2Vec, tfeat | mat | 0.80 | 0.58 | 0.85 | - | - |
| RF [26] | 2024 | mat,counts, Pro2Vec, tfeat | mat, counts | 0.80 | 0.58 | 0.88 | - | - |
| DeepAMP [3] | 2022 | Binary profile | Binary profile | 0.77 | 0.55 | 0.85 | 0.79 | 0.76 |
| XGboost-Ampy [25] | 2022 | PAAC, Blosume62, PseAAC | Blosume62 | 0.86 | 0.75 | 0.88 | 0.76 | 0.97 |
| XGboost [25] | 2022 | PAAC, Blosume62, PseAAC | PAAC | 0.86 | 0.73 | 0.87 | 0.75 | 0.97 |
| **KANampyl**  [**this study**] | 2**025** | AAC, Z-scale, Blosume62, AESNN3, OPF, Merged features | **Z-scale** | **0.92** | **0.82** | **0.90** | **0.86** | **0.95** |

According to previous work, there have been limited computer algorithms indicated for AMPylation detection, and the majority have not performed satisfactorily on the same datasets. For developing the KANampyl model, we used a variety of feature extraction approaches and analyzed the datasets successfully. Therefore, we received an accuracy of 0.92, MCC of 0.82, AUC score of 0.90, and sensitivity and specificity (Sn and Sp) of 0.86 and 0.95. In 2024, ANN and RF methods [26] utilizing mat and counts-based feature representations attained an overall accuracy of 0.80. Our model outperformed theirs by increasing accuracy by 0.12 and boosting MCC by more than 20%, demonstrating that the KANampyl model more accurately captured and predicted the data. In 2022, the DeepAMP technique [3] achieved an accuracy of 0.77 and an MCC of 0.55. Our analysis found that employing Z-scale features increased accuracy by 15% and MCC by more than 25%. Furthermore, our model demonstrated balanced sensitivity and specificity, with 5% and 20% increases, significantly. In 2022, the XGboost-Ampy approach [25] reached 0.86 accuracy with Blosum62 features. When evaluated with Blosum62 attributes, our study produced a similar 0.86 accuracy rate. However, using Z-scale features, our experiment enhanced accuracy by 6%. Although XGboost-Ampy claimed a specificity of 0.97, the result could have been overfitted due to the use of 5-fold validation, whereas this study employed 10-fold cross-validation and obtained a specificity of 0.95, demonstrating excellent performance. Thus, we observe that the KANampyl model based on Z-scale attributes has more effective results than the other proposed approaches. The robustness and accuracy of this model ensures its potential for widespread use in various biological and medical fields.

**Conclusion**

The KANampyl framework presented in this study represents a significant advancement in computational detection of AMPylation sites. By leveraging the Kolmogorov-Arnold Network and robust feature extraction methods, particularly Z-scale features, the framework achieves unprecedented performance, outperforming the state-of-the-art methods by 6%, 11%, 9% and 3% in terms of accuracy, sensitivity, MCC, and AUC scores, respectively. These improvements address previous limitations in feature extraction and validation. Despite the success, achieving perfect accuracy remains challenging due to data limitations. In the future iterations of our work, we aim to mitigate this limitation by incorporating additional AMPylation sites into the dataset and further optimizing the model. The findings of this study underscore the value of computational approaches in PTM research, highlighting their efficiency, cost-effectiveness, and potential for experimental validation. We believe the publicly available KANampyl predictor will serve as invaluable resources for researchers, driving advancements in AMPylation site prediction and broader PTM studies.

**Author Contributions**

**Conceptualization,** T. Karim, M.S.H. Shaon, M.F. Sultan, Azim, S.M, M.S. Akter ; **Data curation,** **Formal analysis, Investigation,** T. Karim, M.S.H. Shaon, M.F. Sultan, Azim, S.M; **Methodology,** T. Karim, M.S.H. Shaon, M.F. Sultan, Azim, S.M; **Project administration,** M.S. Akter, Dehzangi, I, Hasan, M.Z; **Resources, Software,** T. Karim, M.S.H. Shaon, M.F. Sultan, M.S. Akter; **Supervision, Validation,** T. Karim, M.S.H. Shaon, Azim, S.M, Dehzangi, I, M.S. Akter; **Visualization,** T. Karim, M.S.H. Shaon; **Funding,** Moustafa, A**; Writing - original draft, Writing - review editing,** T. Karim, M.S.H. Shaon, M.F. Sultan, M.S. Akter, S.M. Azim, I. Dehzangi; All authors have read and approved the final version of the manuscript.

**References**

1. Jensen, O. N. Modification-specific proteomics: Characterization of post-translational modifications by mass spectrometry. Curr. Opin. Chem. Biol. 8, 33–41 (2004).
2. Kia-Ki, H. & Martinage, A. Post-translational chemical modification (s) of proteins. Int. J. Biochem. 24, 19–28 (1992).
3. Azim, S.M., Sharma, A., Noshadi, I. et al. A convolutional neural network based tool for predicting protein AMPylation sites from binary profile representation. Sci Rep 12, 11451
4. Lin, J., Liang, H., Yan, J. & Luo, L. The molecular mechanism and post-transcriptional regulation characteristic of Tetragenococcus halophilus acclimation to osmotic stress revealed by quantitative proteomics. J. Proteomics 168, 1–14 (2017).
5. Kristiansen, K. Molecular mechanisms of ligand binding, signaling, and regulation within the superfamily of g-protein-coupled receptors: Molecular modeling and mutagenesis approaches to receptor structure and function. Pharmacol. Therap. 103, 21–80 (2004).
6. Chauhan, M., Tarique, M., Sourabh, S. & Tuteja, R. Overview of posttranslational modifications of biochemically characterized plasmodium falciparum helicases. In Helicases from All Domains of Life, 113–124 (Elsevier, 2019).
7. Yarbrough, M. L. et al. Ampylation of rho gtpases by vibrio vops disrupts effector binding and downstream signaling. Science 323,
8. 9–272 (2009).
9. Yarbrough, M. L. & Orth, K. Ampylation is a new post-translational modification. Nat. Chem. Biol. 5, 378–379 (2009).
10. Casey, A. K. & Orth, K. Enzymes involved in ampylation and deampylation. Chem. Rev. 118, 1199–1215 (2018).
11. Kielkowski, P. et al. Ficd activity and ampylation remodelling modulate human neurogenesis. Nat. Commun. 11, 1–13 (2020).
12. Mullard, A. Examining the fic domain. Nat. Rev. Microbiol. 7, 405 (2009).
13. Itzen, A., Blankenfeldt, W. & Goody, R. S. Adenylylation: Renaissance of a forgotten post-translational modification. Trends Biochem. Sci. 36, 221–228 (2011).
14. Choudhary, C., Weinert, B. T., Nishida, Y., Verdin, E. & Mann, M. The growing landscape of lysine acetylation links metabolism and cell signalling. Nat. Rev. Mol. Cell Biol. 15, 536–550 (2014).
15. Nishida, Y. et al. Sirt5 regulates both cytosolic and mitochondrial protein malonylation with glycolysis as a major target. Mol. Cell 59, 321–332 (2015).
16. Xie, Z. et al. Lysine succinylation and lysine malonylation in histones. Mol. Cell. Proteomics 11, 100–107 (2012).
17. Olsen, C. A. Expansion of the lysine acylation landscape. Angew. Chem. Int. Ed. 51, 3755–3756 (2012).
18. Peng, C. et al. The first identification of lysine malonylation substrates and its regulatory enzyme. Mol. Cell. Proteomics 10, M111 (2011).
19. Hirschey, M. D. & Zhao, Y. Metabolic regulation by lysine malonylation, succinylation, and glutarylation. Mol. Cell. Proteomics 14, 2308–2315 (2015).
20. Reddy, H. M. et al. Glystruct: Glycation prediction using structural properties of amino acid residues. BMC Bioinform. 19, 55–64 (2019).
21. Dipta, S. R. et al. Semal: Accurate protein malonylation site predictor using structural and evolutionary information. Comput. Biol. Med. 125, 104022 (2020).
22. Chandra, A. et al. Phoglystruct: Prediction of phosphoglycerylated lysine residues using structural properties of amino acids. Sci. Rep. 8, 1–11 (2018).
23. Uddin, M. R. et al. Evostruct-sub: An accurate gram-positive protein subcellular localization predictor using evolutionary and structural features. J. Theor. Biol. 443, 138–146 (2018).
24. Taherzadeh, G., Dehzangi, A., Golchin, M., Zhou, Y. & Campbell, M. P. Sprint-gly: Predicting n-and o-linked glycosylation sites of human and mouse proteins by using sequence and predicted structural properties. Bioinformatics 35, 4140–4146 (2019).
25. Khater, S. & Mohanty, D. In silico identification of ampylating enzymes and study of their divergent evolution. Sci. Rep. 5, 1–17 (2015).
26. Swati, Z.N.K., Ghulam, A., Sohail, M., Arshed, J.U., Sikander, R. and Malik, M.S., 2022. XGboost-Ampy: Identification of AMPylation Protein Function Prediction Using Machine Learning. VAWKUM Transactions on Computer Sciences, 10(2), pp.83-95.
27. Prabhu, H., Bhosale, H., Sane, A. et al. Protein feature engineering framework for AMPylation site prediction. Sci Rep 14, 8695 (2024).
28. Huang, Y., Niu, B., Gao, Y., Fu, L. & Li, W. Cd-hit suite: A web server for clustering and comparing biological sequences. Bioinformatics 26, 680–682 (2010).
29. Liu B., Gao X., Zhang H. BioSeq-Analysis2.0: an updated platform for analyzing DNA, RNA and protein sequences at sequence level and residue level based on machine learning approaches. Nucleic Acids Res. 2019; 47:e127.
30. Lin K., May A.C., Taylor W.R. Amino acid encoding schemes from protein structure alignments: multi-dimensional vectors to describe residue types. J. Theor. Biol. 2002; 216:361–365.
31. Wei L., Zhou C., Chen H., Song J., Su R. ACPred-FL: a sequence-based predictor using effective feature representation to improve the prediction of anti-cancer peptides. Bioinformatics. 2018; 34:4007–4016.
32. Lee T.Y., Chen S.A., Hung H.Y., Ou Y.Y. Incorporating distant sequence features and radial basis function networks to identify ubiquitin conjugation sites. PLoS One. 2011; 6:e17331.
33. Chen Y.Z., Chen Z., Gong Y.A., Ying G. SUMOhydro: a novel method for the prediction of sumoylation sites based on hydrophobic properties. PLoS One. 2012; 7:e39195.
34. Bhasin M., Raghava G.P. Classification of nuclear receptors based on amino acid composition and dipeptide composition. J. Biol. Chem. 2004; 279:23262–23266.
35. M. Idrees, A.R. Mohammad, N. Karodia, A. Rahman, Multimodal role of amino acids in microbial control and drug development, Antibiotics 9 (6) (2020) 330.
36. Karim, T., Shaon, M.S.H., Sultan, M.F., Hasan, M.Z. and Kafy, A.A., 2024. ANNprob-ACPs: A novel anticancer peptide identifier based on probabilistic feature fusion approach. Computers in Biology and Medicine, 169, p.107915.
37. Liu, Z., Wang, Y., Vaidya, S., Ruehle, F., Halverson, J., Soljačić, M., Hou, T.Y. and Tegmark, M., 2024. Kan: Kolmogorov-arnold networks. arXiv preprint arXiv:2404.19756.
38. Kolmogorov-Arnold Networks: A Comprehensive Guide to Neural Network Advancement

Accessible link: <https://medium.com/@saadsalmanakram/kolmogorov-arnold-networks-a-comprehensive-guide-to-neural-network-advancement-5919fc8f81b1> ( 16, May, 2024)

1. Qiu, Q., Zhu, T., Gong, H., Chen, L. and Ning, H., 2024. ReLU-KAN: New Kolmogorov-Arnold Networks that Only Need Matrix Addition, Dot Multiplication, and ReLU. arXiv preprint arXiv:2406.02075.
2. Bozorgasl, Z. and Chen, H., 2024. Wav-kan: Wavelet kolmogorov-arnold networks. arXiv preprint arXiv:2405.12832.
3. Imambi, S., Prakash, K.B. and Kanagachidambaresan, G.R., 2021. PyTorch. Programming with TensorFlow: Solution for Edge Computing Applications, pp.87-104.
4. Pytorch, A.D.I., 2018. Pytorch.
5. Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L. and Desmaison, A., 2019. Pytorch: An imperative style, high-performance deep learning library. Advances in neural information processing systems, 32.
6. Zhang, Z., 2018, June. Improved adam optimizer for deep neural networks. In 2018 IEEE/ACM 26th international symposium on quality of service (IWQoS) (pp. 1-2). Ieee.
7. Tato, A. and Nkambou, R., 2018. Improving adam optimizer.
8. Bock, S., Goppold, J. and Weiß, M., 2018. An improvement of the convergence proof of the ADAM-Optimizer. arXiv preprint arXiv:1804.10587.
9. Bae, K., Ryu, H. and Shin, H., 2019. Does Adam optimizer keep close to the optimal point?. arXiv preprint arXiv:1911.00289.
10. Li, L., Doroslovački, M. and Loew, M.H., 2020. Approximating the gradient of cross-entropy loss function. IEEE access, 8, pp.111626-111635.
11. Mao, A., Mohri, M. and Zhong, Y., 2023, July. Cross-entropy loss functions: Theoretical analysis and applications. In International Conference on Machine Learning (pp. 23803-23828). PMLR.
12. Ho, Y. and Wookey, S., 2019. The real-world-weight cross-entropy loss function: Modeling the costs of mislabeling. IEEE access, 8, pp.4806-4813.
13. Feng, L., Shu, S., Lin, Z., Lv, F., Li, L. and An, B., 2021, January. Can cross entropy loss be robust to label noise?. In Proceedings of the twenty-ninth international conference on international joint conferences on artificial intelligence (pp. 2206-2212).
14. Goldfarb, G. and Li, G., 2009. Efficient backward-propagation using wavelet-based filtering for fiber backward-propagation. Optics Express, 17(11), pp.8815-8821.
15. Oostwal, E., Straat, M. & Biehl, M. Hidden unit specialization in layered neural networks: ReLU vs. sigmoidal activation. Phys. A Stat. Mech. Appl. 15(564), 125517.
16. Umakantha, N. A new approach to probability theory with reference to statistics and statistical physics. J. Mod. Phys. 7(09), 989.