Journal of Electronics, Electromedical Engineering, and Medical Informatics
Multidisciplinary: Rapid Review: Open Access Journal **Vol. 6, No. 4, October 2024, pp: 405-414; eISSN: 2656-8632**

RESEARCH ARTICLE **OPEN ACCESS**

Manuscript received May 2, 2024; revised May 23, 2024; accepted May 27, 2024; date of publication October 2, 2024
Digital Object Identifier (**DOI**): https://doi.org/10.35882/jeeemi.v6i4.458

How to cite: Md. Firoz Ahmed, and M. Hasnat Kabir, "A Circular Ring Patch Antenna for Breast Cancer Detection Based on Return Loss and Voltage Standing Wave Ratio", Journal of Electronics, Electromedical Engineering, and Medical Informatics, vol. 6, no. 4, pp. 405-414, October 2024.

1D and 2D Feature Extraction Based on AAC and DC Protein Descriptors for Classification of Acetylation in Lysine Proteins using Convolutional Neural Network

Mohammad Reza Faisal[1*], Laila Adawiyah[1], Triando Hamonangan Saragih[1], Dwi Kartini[1], Rudy Herteno[1], Favorisen Rosyking Lumbanraja[2], Lilies Handayani[3,4], and Siti Aisyah Solechah[5]

[1]Department of Computer Science, Lambung Mangkurat University, Banjarbaru, Indonesia
[2]Department of Computer Science, University of Lampung, Lampung, Indonesia
[3]Graduate School of Natural Science and Technology, Kanazawa University, Kanazawa, Japan
[4]Department of Statistics, Tadulako University, Palu, Indonesia
[5]Departement of Public Health, Lambung Mangkurat University, Banjarbaru, Indonesia

Corresponding author: Mohammad Reza Faisal (e-mail: reza.faisal@ulm.ac.id).

ABSTRACT Post-Translational Modification (PTM) denotes a biochemical alteration observed in an amino acid, playing crucial roles in protein activity, functionality, and the regulation of protein structure. The recognition of associated PTMs serves as a fundamental basis for understanding biological processes, therapeutic interventions for diseases, and the development of pharmaceutical agents. Using computational approaches (in silico) offers an efficient and cost-effective means to identify PTM sites swiftly. The exploration of protein classification commences with extracting protein sequence features that are subsequently transformed into numerical features for utilization in classification algorithms. Feature extraction methodologies involve using protein descriptors like Amino Acid Composition (AAC) and Dipeptide Composition (DC). Yet, these approaches exhibit a limitation by neglecting crucial amino acid sequence details. Moreover, both descriptor techniques generate a limited number of 1-dimensional (1D) features, which may not be ideal for processing through the Convolutional Neural Network (CNN) classification method. This investigation presents a novel approach to enhance feature diversity through protein sequence segmentation techniques, employing adjacent and overlapping segment strategies. Furthermore, the study illustrates the organization of features into 1D and 2D formats to facilitate processing through 1D CNN and 2D CNN classification methodologies. The findings of this research endeavour highlight the potential for enhancing the accuracy of acetylation classification in lysine proteins through the multiplication of protein sequence segments in a 2D configuration. The highest accuracy achieved for AAC and DC-based feature extraction methods is 77.39% and 76.75%, respectively. The findings of this research demonstrate the capability of extracting 2D protein characteristics to enhance the overall efficiency of protein categorization, particularly in the context of identifying acetylation in lysine proteins.

INDEX TERMS classification of acetylation, lysine proteins, protein segmentation, protein descriptor, convolutional neural network.

I. INTRODUCTION

Post-Translational Modification (PTM) constitutes a crucial mechanism essential for protein constituents. After the translation process, a chemical alteration occurs in the protein, thereby diversifying a finite pool of amino acids through PTM. This expands 20 amino acids to an infinite array of potential residues. Modifications are necessary to facilitate cell growth, transcription regulation, and metabolic activities vital for daily sustenance. [1], [2], [3]. Among the indispensable PTMs is the process of acetylation, acknowledged as one of the most significant post-translational protein modifications, exerting a pivotal influence on a myriad of cellular functions [4].

Typically observed at lysine residues, acetylation proves beneficial in aiding calcium absorption, hormone synthesis, collagen formation, and antibody production, contributing to repair of DNA damage, transcription, and gene expression.

Predicting PTM within protein sequences through in vitro experiments demands considerable time and effort, mainly when identifying extensive data [5], [6]. A feasible approach to streamline this process involves the utilization of in silico methods, leveraging algorithms and computational tools [7], [8], [9]. In the realm of in silico research on classification scenarios, the methodology generally encompasses two primary phases: feature extraction and classification.

In the initial phase, the focus lies on feature extraction, transforming unstructured data into structured data comprising numerical values, thereby rendering it amenable for processing by classification algorithms. Unstructured data, such as audio[10], images [11], [12], text [13], [14], and others, necessitates passage through the feature extraction process.

Protein sequence data resembles text data, where each instance comprises a sequence of characters. In text data, the configuration of characters culminates in the construction of words and sentences in natural human language. In contrast, protein sequence data embodies an assemblage of characters symbolizing the sequence of amino acids. Currently, the prevalent technique for feature extraction in text data involves word-based embedding methods [13], [14]. Analogously, word-based embedding strategies find application in processing protein sequence data by segmenting a sequence and subjecting each segment to computation using the word embedding model [15], [16].

Another widely employed approach for extracting features from protein sequence data involves the utilization of protein descriptor-based techniques. Over time, various protein descriptors have been devised employing diverse calculation methodologies. Noteworthy among these are descriptors, such as AAC, DC, and TC [17] , based on composition calculations. AAC yields 20 features, with each feature value computed by comparing the frequency of occurrence of an amino acid against the total amino acids in the sequence. DC is predicated on dipeptide ratios within the sequence, while TC hinges on tripeptide comparisons in the sequence. These two descriptors generate more features than AAC, amounting to 400 and 8000, respectively. Feature extraction based on AAC and DC and four machine learning algorithms, including Support Vector Machines (SVM), K-Nearest Neighbor, Random Forest, and Naïve Bayes, were applied to discriminate psychrophilic enzymes [18]. Results indicated the superior performance of AAC over DC and the combined AAC and DC features. In the context of acetylation classification in lysine proteins, AAC and DC-based feature extraction were employed in conjunction with the SVM classification algorithm [3]. Findings from this investigation demonstrated the superior performance of DC over AAC. Despite variations in protein sequence lengths, all three descriptors yield a fixed number of features and uniformly produce 1-dimensional (1D) structured data. However, the limitation of the protein descriptors' output lies in its absence of amino acid sequence information regarding the sequence.

Other protein descriptors, such as Autocorrelation Descriptors [19], [20], Composition/Transition/Distribution (CTD) Descriptor [19], [21] and Quasi-sequence-order descriptors [22]. The quantity of structured data features each protein descriptor produces is detailed in
TABLE 1.

TABLE 1
Commonly Used Descriptors

Descriptor Name	Features
Amino Acid Composition	20
Dipeptide Composition	400
Tripeptide Composition	8000
Normalised Moreau-Broto Autocorrelation	240
Moran Autocorrelation	240
Geary Autocorrelation	240
CTD	147
Conjoint Triad	343
Sequence-Order-Coupling Number	60
Quasi-Sequence-Order Descriptors	100
Pseudo-Amino Acid Composition	50
Amphiphilic Pseudo-Amino Acid Composition	80

Another protein descriptor identified is the Position-Specific Scoring Matrix (PSSM) profile, which exhibits distinct characteristics compared to previously mentioned descriptors [17], [23], [24]. The PSSM profile generates 2-dimensional (2D) data represented as an L x 20 matrix, where L denotes the sequence length, producing varied outputs for sequences of different lengths, necessitating dimension equalization through zero-vector padding techniques [25], [26].

The subsequent stage involves processing structured data utilizing classification algorithms. Machine learning algorithms like Support Vector Machine (SVM), Naïve Bayes, K-Nearest Neighbors (KNN), and Random Forest are utilized for 1D data classification, while deep learning-based algorithms like Deep Neural Network (DNN), Convolutional Neural Network (CNN), and Long Short Term Memory (LSTM) are gaining popularity for outperforming machine learning algorithms [17], [27], [28]. A study [17] compared classification performance between models constructed using the DNN algorithm and result-structured data from composition-based protein descriptors with PSSM profiles, indicating that models incorporating composition-based extraction features like AAC and DC exhibit superior DNN classification performance compared to those utilizing structured data from PSSM Profile based feature extraction. Although other deep learning algorithms such as CNN and LSTM were explored, input from composition-based protein descriptors was not used due to limited features in the descriptor structured data, rendering it suboptimal for processing by both algorithms.

Building upon the rationale above, this study undertook a distinct case classification analysis focusing on acetylation classification in lysine proteins employing composition-based feature extraction and a CNN classification algorithm. The

investigation centred on two protein descriptors, AAC and DC. However, structured data derived from AAC based feature extraction yielded insufficient structured data, hindering CNN from producing an optimal classification model. To address this limitation, the study suggests employing feature extraction techniques like adjacent and overlapped segmentation methods to augment the feature count. This segmentation approach not only overcomes the constraints of composition-based protein descriptors in obtaining amino acid sequence information but also expands the features for input to CNN, an algorithm capable of accommodating inputs beyond 1D, yet unexplored in protein sequence classification studies utilizing composition-based descriptors [3], [29].

The primary objective of this study was to develop an optimal classification model for acetylation in lysine proteins. To achieve this objective, a series of experiments were conducted to address the following research questions:

1. What is the accuracy of the 1D CNN classification model using the original input?
2. What is the accuracy of the 1D CNN classification model with 1D structured data input sequence segmentation results?
3. What is the accuracy of the 2D CNN classification model with 2D structured data input sequence segmentation results?

The study findings contributed to:

1. Establishment of 2D structured data through sequence segmentation techniques and composition-based protein descriptors.
2. Using the CNN algorithm, developing an optimal classification model for acetylation in lysine proteins.

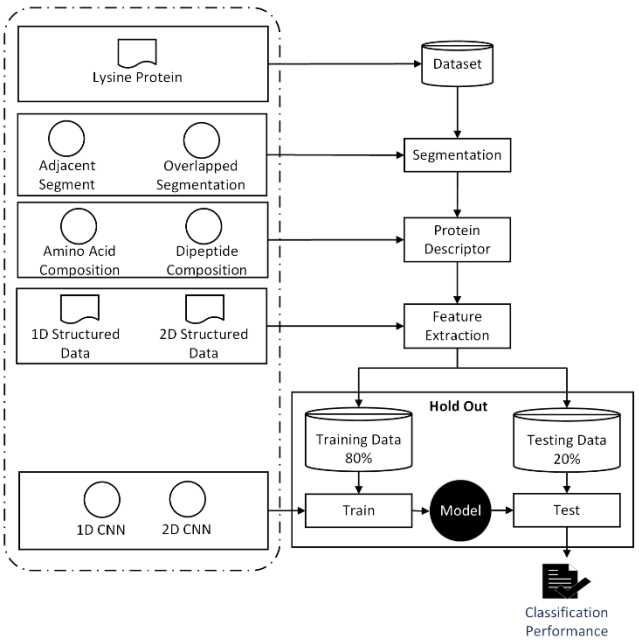

FIGURE 1. Research flow.

II. MATERIAL AND METHODS

The research flow of this research can be seen in FIGURE 1.

A. DATASET

The datasets utilized in this investigation consist of acetylated (Positive) and unacetylated (Negative) lysine proteins, which are displayed in TABLE 2 [3].

TABLE 2
Dataset

Positive	Negative	Total
8701	8701	17402

TABLE 2 presents a dataset that includes two label classes, namely positive and negative. The positive class comprises 8701 protein sequences, while the negative class also contains 8701 protein sequences. The total number of protein sequences amounts to 17402. This data reveals that the quantity of both classes is equal, indicating that the classification scenario addressed in this study is a balanced data classification. This dataset encompasses protein sequences, some examples of which are visible in TABLE 3. The initial column displays the sequential number of the protein sequence; this column was excluded during the development of the classification model. The subsequent column presents the textual representation of the protein sequence. Lastly, the third column indicates the label class corresponding to each record.

TABLE 3
Protein Sequence

No	Protein sequence	Label
1	RKDAAEHTLTAYKAAQDIANS	Negative
2	DKIVVCCVTGSTTAGILAGMA	Negative
…	…	…
8701	EQPVVLHTWTKESAHNYENNC	Negative
8702	FFDIDTKYYTKELHKAAFVLP	Positive
…	…	…
17401	RKDAAEHTLTAYKAAQDIANS	Positive
17402	MLTCNKAGSRMVVDAANSNGP	Positive

In the second column of TABLE 3, each protein sequence exhibits a uniform character count, specifically comprising 21 characters. Each character serves as a representation of a distinct amino acid. The elucidation of each amino acid symbol employed within the protein sequence is detailed in the accompanying description in TABLE 4. The second column of TABLE 4 presents the amino acid symbols, which consist of an abbreviation represented in character form alongside a corresponding word. The third column delineates the full nomenclature of the amino acid.

TABLE 4
Amino Acid

No	Symbol	Amino Acid	
1	L	Leu	Leucine
2	A	Ala	Alanine
3	P	Pro	Proline
4	V	Val	Valine
5	G	Gly	Glycine
6	Y	Tyr	Tyrosine
7	I	Ile	Isoleucine
8	M	Met	Methionine
9	F	Phe	Phenylalanine
10	W	Trp	Tryptophan
11	S	Ser	Serine
12	T	Thr	Threonine
13	C	Cys	Cysteine
14	N	Asn	Asparagine
15	Q	Gln	Glutamine

No	Symbol		Amino Acid
16	D	Asp	Aspartate
17	K	Lys	Lysine
18	H	His	Histidine
19	R	Arg	Arginine
20	E	Glu	Glutamate

B. SEGMENTATION

The primary stage of the research workflow involves segmentation. This segmentation methodology is designed to partition protein sequences into *k* segments of identical length to extract novel insights. Two categories of segments exist, namely adjacent and overlapped segments [3], [29].

Adjacent segments divide the protein sequence length into k sections, where k = 3. For example, in a sequence like "DKIVVCCVTGSTTAGILAGMA" with a length of 21, the first segment is computed with a length of 7 from the sequence start, yielding "DKIVVCC" as the first segment, "VTGSTTA" as the second segment, and "GILAGMA" as the third segment calculated from the end of the initial sequence. An illustrative depiction of creating adjacent segments can be observed in FIGURE 2.

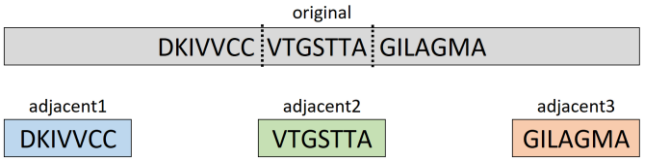

FIGURE 2. Adjacent segment.

Overlapped segments combine half of the last segment with half of the initial segment. They introduce a new sequence feature by combining the latter half of the first segment with the former half of the second segment and the latter half of the second segment with the initial half of the third segment resulting in "VCCVTGS" and "TTAGILA" as overlapped segments in the sequence. Features derived from overlapped segments encompass the amalgamation of the original sequence, adjacent segments 1, 2, 3, overlapped segments 1, and 2. An illustration of generating overlapped segments can be seen in FIGURE 3.

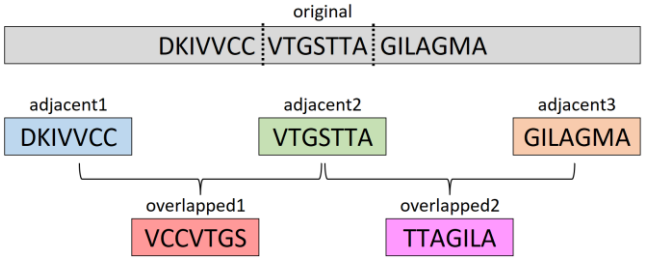

FIGURE 3. Overlapped segment.

C. PROTEIN DESCRIPTOR

A protein descriptor is a technique to convert the textual representation of a protein sequence into a numerical value. Various methods can be employed for this conversion process, including Amino Acid Composition (AAC) and Dipeptide Composition (DC) techniques.

Amino Acid Composition (AAC) quantifies the proportion of each amino acid type in a protein sequence. The fraction of all 20 amino acids is shown in Eq. (1)[30] .

$$f_{(r)} = \frac{N_r}{N} \quad r = 1,2,3 \ldots 20 \qquad (1)$$

Where *Nr* represents the number of amino acid types r, and N is the sequence length.

Dipeptide Composition (DC) furnishes a 400-feature descriptor, defined in Eq. (2) [30].

$$f_{(r,s)} = \frac{N_{rs}}{N-1} \quad r,s = 1,2,3 \ldots 20 \qquad (2)$$

Where N_{rs} denotes the number of dipeptides formed by *r* type dan tipe *s* type.

The implementation of the protein descriptor in this study utilizes the protr package developed in the R programming language [31]. The extracTAAC() function is used for transforming protein sequence text with AAC, while the extracDC() function is employed for DC.

D. FEATURE EXTRACTION

Feature extraction in this study leverages segmentation techniques and protein descriptors as previously elucidated. Two categories of structured data, namely 1D and 2D, were generated.

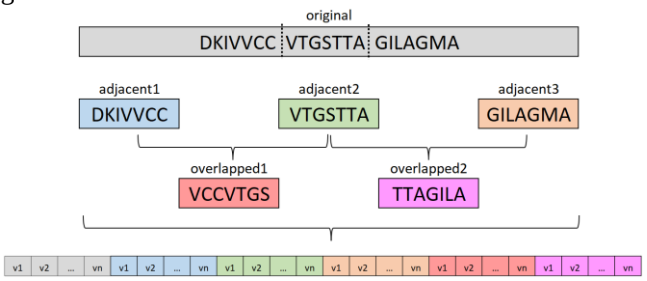

FIGURE 4. Feature extraction generates 1D structured data.

The formation process of 1D structured data based on prior experiments [29], [3], is illustrated in FIGURE 4. The original sequence undergoes processing by the protein descriptor, resulting in 1D structured data with a specific number of corresponding features. Subsequently, the protein descriptor processes each adjacent and overlapped segment derived from the segmentation process. Each 1D structured data is sequentially merged as depicted in FIGURE 4.

For 1D data, the features outlined in Eq. (3) are produced. Where $N_{feature}$ represents the resultant feature, k signifies the number of adjacent segments, and n denotes the number of features generated by the protein descriptor.

$$N_{feature} = 2k \times n \qquad (3)$$

The 2D structured data formation proposed in our study follows a sequential process, but the data combination occurs concurrently, as illustrated in FIGURE 5. The resultant 2D data forms a matrix with $m \times n$ dimensions where m represents the row and n signifies the column. The values $m = 2k$, k denote the number of adjacent segments, while n indicates the quantity of features produced by the protein descriptor.

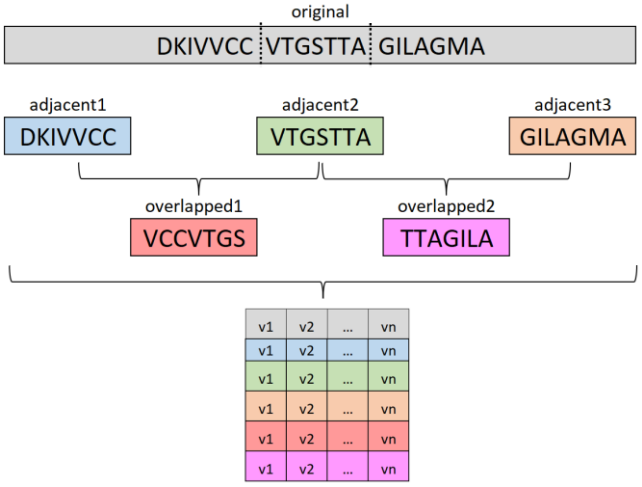

FIGURE 5. Feature extraction generates 2D structured data.

E. DATA SPLIT

Developing a classification model involves data training and testing using the holdout method. The dataset is split into an 80:20 ratio, where 80% is allocated for training and 20% for testing. Data sharing entails random sampling based on the distribution of label classes within the dataset.

F. TRAIN AND TESTING

The learning process entails constructing a classification model utilizing a specific algorithm, such as the Convolutional Neural Network (CNN) used in this research. CNN is an artificial neural network designed to analyze grid-format data, comprising key layers like the convolution layer for feature learning, the pooling layer for dimension reduction, and the fully connected (dense) layer for final classification output [32], [33]. CNN accommodates data of varying dimensions, including 1D, 2D, and 3D [34].

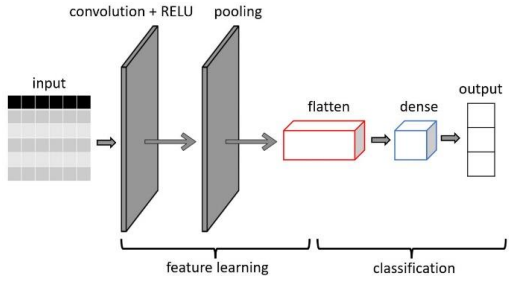

FIGURE 6. Common CNN architecture

FIGURE 6 shows the general CNN architecture. CNN is divided into two parts, namely feature learning and classifier. Feature learning is the ability of a model to extract important features from input automatically. This extraction operation begins with a convolution operation carried out by the convolution layer. The result of the convolutional operation is a feature map. The ReLU activation function is applied to each value in the feature maps. Next, the pooling layer reduces the spatial dimensions of the feature map. The classification begins by converting the feature map into a one-dimensional vector by a flatten layer. Then, the vector is received by the dense layer to be processed with linear and non-linear operations. The output layer produces predictions [35], [36].

1D CNN algorithms are applicable in diverse classifications like machine crack signals, cardiac electrical signals, and text categorization [37], [38], [39]. Additionally, 1D CNN is relevant for protein classification tasks [40]. The architectural configuration of 1D CNN utilized in this study is detailed in TABLE 5 and the architecture is shown in FIGURE 7.

TABLE 5
1D CNN Architecture

Layer (type)	Output Shape	Number of Param
conv1d (Convd1D	(None, 118, 64)	256
dropout (Dropout)	(None, 118, 64)	0
conv1d_1 (Convd1D	(None, 118, 128)	24704
dropout_1 (Dropout)	(None, 118, 128	0
max_pooling1d (Maxpooling1D)	(None, 59, 128)	0
flatten (Flatten)	(None, 7552)	0
dense (Dense)	(None, 768)	5800704
dropout_2 (Dropout)	(None, 768)	0
dense_1 (Dense)	(None, 256)	196864
dropout_3 (Dropout)	(None, 256)	0
dense_2 (Dense)	(None, 1)	257
Total params: 6,022,785 Trainable params: 6,022,785 Non-trainable params: 0		

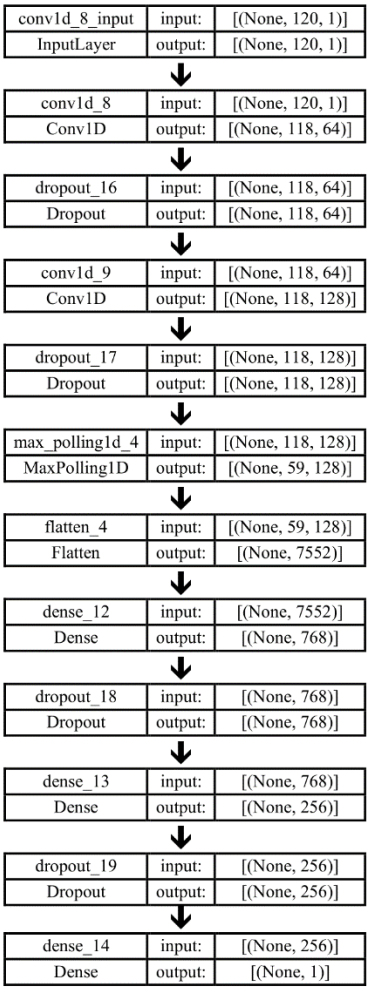

FIGURE 7. Plot model for 1D CNN architecture.

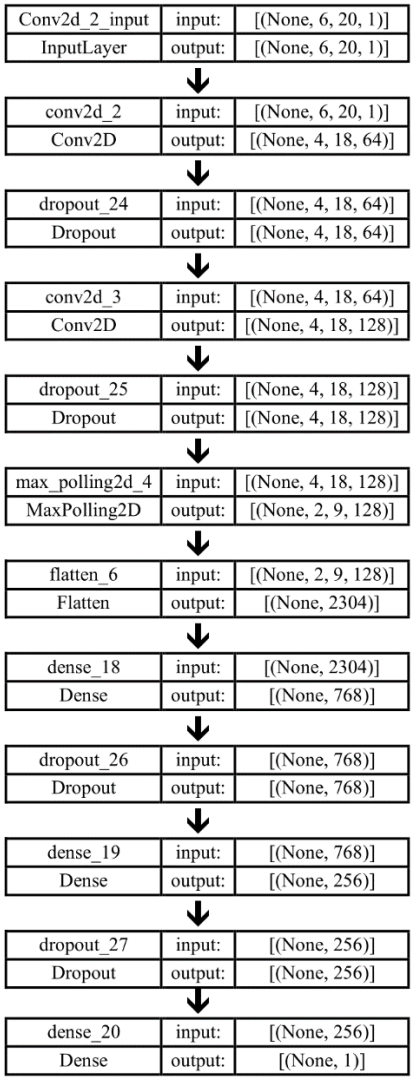

FIGURE 7. Plot model for 2D CNN architecture

While 2D CNN is commonly utilized for image classification purposes [41], it can also be deployed for text and protein classification tasks [42], [43]. The input for 2D CNN comprises 2D data presented in matrix form with rows and columns. The 2D CNN architecture utilized in this study is outlined in TABLE 6 and the architecture is shown in FIGURE 8. The classification algorithm processes the training data to establish a classification model, which is then employed in the testing stage to predict class labels for the testing data. The classification model's performance is evaluated by comparing these predictions with the actual class labels, utilizing metrics like accuracy, sensitivity, and specificity [10], [44].

TABLE 6
2D CNN Architecture

Layer (type)	Output Shape	Number of Param
conv2d (Convd1D	(None, 4, 18, 64)	640
dropout (Dropout)	(None, 4, 18, 64)	0
conv2d_1 (Convd1D	(None, 4, 18, 128)	73856
dropout_1 (Dropout)	(None, 4, 18, 128	0
max_pooling2d (Maxpooling2D)	(None, 2, 59, 128)	0
flatten (Flatten)	(None, 2304)	0
dense (Dense)	(None, 768)	1770240
dropout_2 (Dropout)	(None, 768)	0
dense_1 (Dense)	(None, 256)	196864
dropout_3 (Dropout)	(None, 256)	0
dense_2 (Dense)	(None, 1)	257
Total params: 6,022,785 Trainable params: 6,022,785 Non-trainable params: 0		

III. RESULTS

A. FEATURE EXTRACTION RESULTS

In this investigation, feature extraction was conducted utilizing a segmentation methodology with varying values of k, specifically 3, 4, and 5, and subsequently analyzed with descriptor proteins AAC and DC. The outcomes of this feature extraction are presented in TABLE 7. Within the table, the initial column provides data regarding the dimensions, each associated with two protein descriptors showcased in the second column. The quantity of secured segmentations is detailed in the third column. Subsequent columns, namely the fourth, fifth, and sixth, sequentially display the counts of adjacent segments, overlapped segments, and the total number of features. For instance, the first row illustrates the creation of 1D structured data employing the protein descriptor AAC and a segment number of $k = 3$. This process yields six sequences comprising three adjacent segments, two overlapped segments, and an original segment. Following the processing and combination of each sequence with AAC, a total of 120 features are generated, calculated as $6 \times 20 = 120$ features.

TABLE 7
Feature Extraction Results

Data Dimension	Protein Descriptor	k	# Adjacent	# Overlapped	# Features
1D	AAC	3	3	2	120
		4	4	3	160
		5	5	4	200
	DC	3	3	2	2400
		4	4	3	3200
		5	5	4	4000
2D	AAC	3	3	2	6×20
		4	4	3	8×20
		5	5	4	10×20
	DC	3	3	2	6×400
		4	4	3	6×400
		5	5	4	6×400

The feature extraction procedure in this investigation yielded 12 structured data, encompassing six 1D data and six 2D data. These data sets were processed utilizing a CNN algorithm that generated 12 classification models.

B. EVALUATION RESULTS

Any residual data generated was utilized to construct a classification model and assess its performance. Initially, the execution of the CNN algorithm in this study involved specific parameters that are batch size: 256, Epoch: 20, and Learning rate: 0.001.

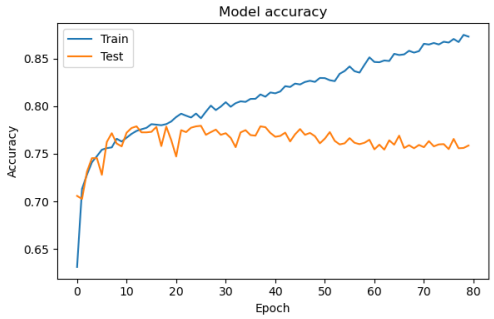

FIGURE 8. Model accuracy.

The selection of epoch values was based on preliminary experiments, setting epoch values at 80. The outcomes are depicted in FIGURE 9, illustrating that the model's accuracy remains constant or does not improve beyond 20 epochs. TABLE 8 showcases the classification performance using 1D structured data processed by 1D CNN, with the highest accuracy recorded at 77.16% from models utilizing inputs from AAC-based feature extraction with segmentation using $k = 5$.

TABLE 8
Performance of 1D CNN

Protein Descriptor	k	Accuracy (%)	Sensitivity (%)	Specificity (%)
AAC	3	76.18	83.50	80.70
	4	77.04	78.76	78.25
	5	**77.16**	79.97	79.03
DC	3	75.61	79.94	78.05
	4	75.84	79.08	77.64
	5	76.50	81.78	79.64

TABLE 9 exhibits the classification performance utilizing 2D structured data processed by 2D CNNs, with the highest accuracy achieved at 77.39% from models employing inputs from AAC-based feature extraction with segmentation using $k = 5$.

TABLE 9
Performance of 2D CNN

Protein Descriptor	k	Accuracy (%)	Sensitivity (%)	Specificity (%)
AAC	3	77.27	81.14	79.57
	4	76.98	83.33	81.15
	5	**77.39**	80.67	79.55
DC	3	76.27	79.94	78.36
	4	75.26	79.65	77.70
	5	76.75	81.89	79.83

IV. DISCUSSION

The classification performance results are compared based on the protein descriptor group and the classification algorithm employed.

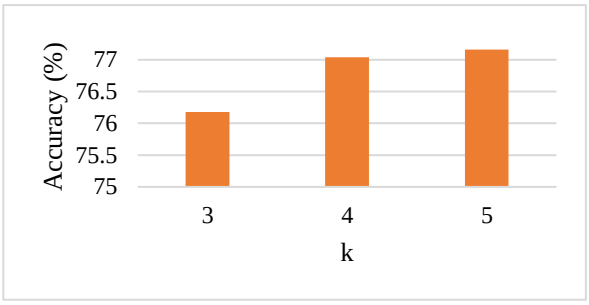

FIGURE 9. Classification performance of AAC and 1D CNN model

FIGURE 9 and FIGURE 12 indicate a consistent increase in model performance as the value of k rises in the 1D CNN algorithm. In contrast, FIGURE 11 and FIGURE 13 display a similar performance enhancement in models constructed with the 2D CNN algorithm as k increases, except for a decline in performance at $k = 2$.

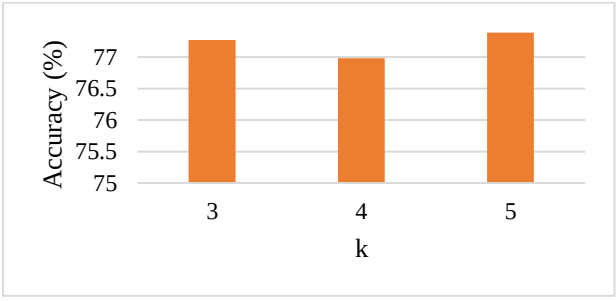

FIGURE 10. Classification performance of AAC and 2D CNN model

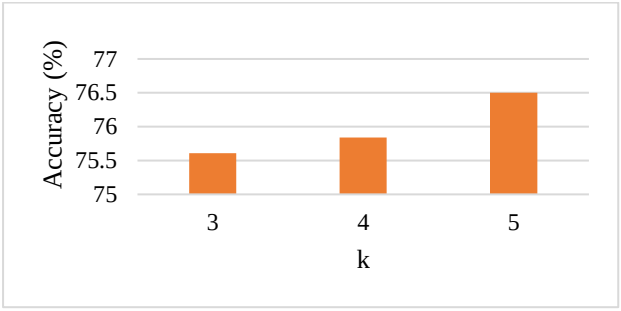

FIGURE 11. Classification performance of DC and 1D CNN model

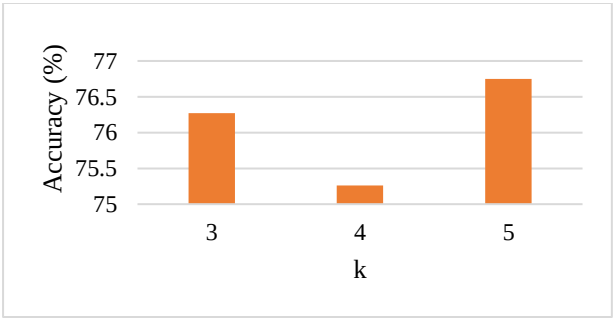

FIGURE 12. Classification performance of DC and 2D CNN model

The proportional enhancement in performance with increasing k values suggests that the CNN algorithm can excel in data with numerous features and can select crucial features to enhance classification performance. Despite a decrease in accuracy observed in models utilizing segmentation inputs in

the 2D CNN algorithm, this algorithm can generate superior models compared to those created with 1D CNN, as evidenced in TABLE 10.

TABLE 10
Performance Comparison Base on Dimension

Protein Descriptor	k	Accuracy (%)	
		1D CNN	2D CNN
AAC	5	77.16	**77.39**
DC	5	76.50	**76.75**

Upon comparing the values between the 1D CNN and 2D CNN columns, it is apparent that 2D CNN outperforms. For inputs based on AAC descriptor proteins, the performance enhancement was 0.23%, and for DC, it was 0.25%. Although the increment is modest, statistical testing using paired t-test yielded a two-tailed P-value of 0.0265, indicating statistical significance by conventional criteria.

TABLE 11
Performance Comparison Base On Segmentation

	Accuracy (%)	
Protein Descriptor	Original	Segmentation
AAC	49.64	**77.39**
DC	73.88	**76.75**

TABLE 11 compares the performance of classification models utilizing feature extraction outcomes from the original sequence alone and the feature extraction results employing the original sequence segmentation method. These findings successfully exhibit a rise of 27.75% and 2.87% for AAC and DC-based feature extraction. The outcomes of this evaluation indicate that the suggested segmentation method effectively enhances the performance of the CNN algorithm.

TABLE 10 reveals that models constructed utilizing input from protein descriptor AAC-based feature extraction can exhibit superior performance even with fewer characteristics than DC. This issue arises due to using short-sized sequences in the dataset, while DC-based feature extraction generates 400 attributes, resulting in sparse data. This situation worsens when the segmentation operation shortens the sequence, increasing zero values within the resultant structured data. Sparse data, which includes zero values, can influence the effectiveness of classification methodologies [45], [46]. According to these findings, it is evident that the segmentation method employed could reduce classification effectiveness when a protein descriptor is compositionally based on features commonly utilized for processing short sequences.

The aforementioned elucidation delineates the potential limitations of inadequate feature extraction methods when dealing with concise protein sequences. This limitation arises from the constraint on the number of possible segmentations. The outcomes of subpar feature extraction undeniably impact the precision of the categorization process, resulting in an accuracy rate below 80%. Consequently, there exists a significant margin for enhancing the classification accuracy. The efficacy of this extraction method necessitates evaluation, especially in the context of protein classification involving lengthy sequence dimensions.

The findings of this investigation propose that amalgamating protein sequence segmentation with composition-centric protein descriptors could be leveraged to generate numerous features amenable to processing through sophisticated algorithms like CNN. Furthermore, this study presents a methodology for configuring 2D features by organizing the protein descriptor outputs from individual segments into a two-dimensional array.

V. CONCLUSION

The outcomes of this investigation indicate that the precision of the classification model utilizing the original sequence with AAC and DC feature extraction and 1D CNN classification approach was 49.64% and 73.88%, respectively. The precision of the 1D CNN classification model was enhanced to 77.16% and 76.50% after conducting segmentation with three segments on AAC and DC-based feature extraction. The study also showcased the effectiveness of the proposed feature extraction method in creating 2D structured data handled by the 2D CNN algorithm, resulting in optimal accuracies of 77.39% and 76.75%.

Nevertheless, our proposed feature extraction technique based on sequence segmentation and protein descriptor composition may have disadvantages when employing protein descriptors that yield high-dimensional features like DCs. In this research, a DC descriptor protein with a concise input sequence yields sparse data with numerous zero values, reducing classification efficacy. Moreover, the highest accuracy achieved in this study remains below 80%, indicating opportunities for further investigations to enhance accuracy.

Considering the constraints of this study, future research endeavors aimed at enhancing accuracy could involve implementing 1D and 2D segmentation-based feature extraction on alternative types of descriptor proteins. Additional studies could explore the utilization of other deep learning algorithms such as LSTM and hybrid CNN LSTM, which exhibit proficiency in processing sequential data information.

REFERENCES

[1] K.-C. Chou, "Progresses in predicting post-translational modification," *International Journal of Peptide Research and Therapeutics*, vol. 26, no. 2, pp. 873–888, 2020.
[2] A. H. Shukri, V. Lukinović, F. Charih, and K. K. Biggar, "Unraveling the battle for lysine: A review of the competition among post-translational modifications," *Biochimica et Biophysica Acta (BBA)-Gene Regulatory Mechanisms*, p. 194990, 2023.
[3] A. Rizqiana, M. R. Faisal, and F. R. Lumbanraja, "Implementation protein sequence segmentation in AAC and DC as protein descriptors for improving a classification performance of acetylation prediction," in *Journal of Physics: Conference Series*, 2021, vol. 1751, no. 1, p. 12031.
[4] V. Vaghasia, K. S. Lata, S. Patel, and J. Das, "Deciphering the lysine acetylation pattern of leptospiral strains by in silico approach," *Network Modeling Analysis in Health Informatics and Bioinformatics*, vol. 12, no. 1, p. 15, 2023.
[5] J. Chen and Y.-H. Tsai, "Applications of genetic code expansion in studying protein post-translational modification," *Journal of Molecular Biology*, vol. 434, no. 8, p. 167424, 2022.
[6] B. Abapihi *et al.*, "Parameter estimation for high dimensional classification model on colon cancer microarray dataset," *Journal of Physics: Conference Series*, vol. 1899, no. 1, p. 12113, May 2021, doi: 10.1088/1742-6596/1899/1/012113.
[7] J. P. Utami, N. Kurnianingsih, and M. R. Faisal, "An in silico study of the Cathepsin L inhibitory activity of bioactive compounds in

Stachytarpheta jamaicensis as a COVID-19 drug therapy," *Makara Journal of Science*, vol. 26, no. 1, p. 3, 2022.

[8] M. D. Darma, M. Reza Faisal, I. Budiman, R. Herteno, J. P. Utami, and B. Abapihi, "In Silico Prediction of Indonesian Herbs Compounds as Covid-19 Supportive Therapy using Support Vector Machine," in *2021 4th International Conference of Computer and Informatics Engineering (IC2IE)*, Sep. 2021, pp. 62–67. doi: 10.1109/IC2IE53219.2021.9649383.

[9] F. Indriani, K. R. Mahmudah, B. Purnama, and K. Satou, "Prottrans-glutar: Incorporating features from pre-trained transformer-based models for predicting glutarylation sites," *Frontiers in Genetics*, vol. 13, p. 885929, 2022.

[10] P. A. Riadi, M. R. Faisal, D. Kartini, R. A. Nugroho, D. T. Nugrahadi, and D. B. Magfira, "A Comparative Study of Machine Learning Methods for Baby Cry Detection Using MFCC Features," *Journal of Electronics, Electromedical Engineering, and Medical Informatics*, vol. 6, no. 1, pp. 73–83, 2024.

[11] R. A. Rahma, R. A. Nugroho, D. Kartini, M. R. Faisal, and F. Abadi, "Combination of texture feature extraction and forward selection for one-class support vector machine improvement in self-portrait classification," *International Journal of Electrical and Computer Engineering*, vol. 13, no. 1, pp. 425–434, 2023, doi: 10.11591/ijece.v13i1.pp425-434.

[12] K. A. Putri and W. F. Al Maki, "Enhancing Pneumonia Disease Classification using Genetic Algorithm-Tuned DCGANs and VGG-16 Integration," *Journal of Electronics, Electromedical Engineering, and Medical Informatics*, vol. 6, no. 1, pp. 11–22, 2024.

[13] M. R. Faisal, I. Budiman, F. Abadi, D. T. Nugrahadi, M. Haekal, and I. Sutedja, "Applying Features Based on Word Embedding Techniques to 1D CNN for Natural Disaster Messages Classification," *2022 5th International Conference on Computer and Informatics Engineering, IC2IE 2022*, no. December, pp. 192–197, 2022, doi: 10.1109/IC2IE56416.2022.9970188.

[14] M. R. Faisal, R. A. Nugroho, R. Ramadhani, F. Abadi, R. Herteno, and T. H. Saragih, "Natural Disaster on Twitter: Role of Feature Extraction Method of Word2Vec and Lexicon Based for Determining Direct Eyewitness," *Trends in Sciences*, vol. 18, no. 23, p. 680, 2021.

[15] H. Zulfiqar *et al.*, "Deep-STP: A deep learning-based approach to predict snake toxin proteins by using word embeddings," *Frontiers in Medicine*, vol. 10, p. 1291352, 2024.

[16] P. Pratyush, S. Pokharel, H. Saigo, and D. B. Kc, "pLMSNOSite: an ensemble-based approach for predicting protein S-nitrosylation sites by integrating supervised word embedding and embedding from pre-trained protein language model," *BMC bioinformatics*, vol. 24, no. 1, p. 41, 2023.

[17] L. Yu, L. Xue, F. Liu, Y. Li, R. Jing, and J. Luo, "The applications of deep learning algorithms on in silico druggable proteins identification," *Journal of Advanced Research*, vol. 41, pp. 219–231, 2022.

[18] A. Huang, F. Lu, and F. Liu, "Discrimination of psychrophilic enzymes using machine learning algorithms with amino acid composition descriptor," *Frontiers in Microbiology*, vol. 14, 2023, doi: 10.3389/fmicb.2023.1130594.

[19] A. Mckenna and S. Dubey, "Machine learning based predictive model for the analysis of sequence activity relationships using protein spectra and protein descriptors," *Journal of Biomedical Informatics*, vol. 128, p. 104016, 2022.

[20] F. Zandi, P. Mansouri, and M. Goodarzi, "Global protein-protein interaction networks in yeast saccharomyces cerevisiae and helicobacter pylori," *Talanta*, vol. 265, p. 124836, 2023.

[21] S. Charles, A. Subeesh, and J. Natarajan, "Tree based models for classification of membrane and secreted proteins in heart," *Journal of Proteins and Proteomics*, pp. 1–11, 2024.

[22] L. Wang and L. Hu, "A deep learning algorithm for predicting protein-protein interactions with nonnegative latent factorization," in *2021 International Conference on Cyber-Physical Social Intelligence (ICCSI)*, 2021, pp. 1–6.

[23] Q.-H. Kha, Q.-T. Ho, and N. Q. K. Le, "Identifying SNARE proteins using an alignment-free method based on multiscan convolutional neural network and PSSM profiles," *Journal of Chemical Information and Modeling*, vol. 62, no. 19, pp. 4820–4826, 2022.

[24] W. Gao, D. Xu, H. Li, J. Du, G. Wang, and D. Li, "Identification of adaptor proteins by incorporating deep learning and PSSM profiles," *Methods*, vol. 209, pp. 10–17, 2023.

[25] S. Chauhan and S. Ahmad, "Enabling full-length evolutionary profiles based deep convolutional neural network for predicting DNA-binding proteins from sequence," *Proteins: Structure, Function, and Bioinformatics*, vol. 88, no. 1, pp. 15–30, 2020.

[26] Y. He and S. Wang, "SE-BLTCNN: A channel attention adapted deep learning model based on PSSM for membrane protein classification," *Computational biology and chemistry*, vol. 98, p. 107680, 2022.

[27] P. Wang, E. Fan, and P. Wang, "Comparative analysis of image classification algorithms based on traditional machine learning and deep learning," *Pattern Recognition Letters*, vol. 141, pp. 61–67, 2021.

[28] S. Huang, I. Arpaci, M. Al-Emran, S. K\il\içarslan, and M. A. Al-Sharafi, "A comparative analysis of classical machine learning and deep learning techniques for predicting lung cancer survivability," *Multimedia Tools and Applications*, vol. 82, no. 22, pp. 34183–34198, 2023.

[29] M. R. Faisal *et al.*, "Improving Protein Sequence Classification Performance Using Adjacent and Overlapped Segments on Existing Protein Descriptors," *Journal of Biomedical Science and Engineering*, vol. 11, no. 06, pp. 126–143, 2018, doi: 10.4236/jbise.2018.116012.

[30] N. Xiao, D.-S. Cao, M.-F. Zhu, and Q.-S. Xu, "protr: R package for generating various numerical representation schemes of protein sequences," 2017. https://cran.r-project.org/web/packages/protr/vignettes/protr.html (accessed Dec. 20, 2017).

[31] N. Xiao, D.-S. Cao, M.-F. Zhu, and Q.-S. Xu, "protr/ProtrWeb: R package and web server for generating various numerical representation schemes of protein sequences," *Bioinformatics*, vol. 31, no. 11, pp. 1857–1859, 2015, doi: 10.1093/bioinformatics/btv042.

[32] I. Budiman, M. R. Faisal, D. T. Nugrahadi, M. K. Delimayanti, S. E. Prastya, and others, "Harvesting Natural Disaster Reports from Social Media with 1D Convolutional Neural Network and Long Short-Term Memory," in *2023 Eighth International Conference on Informatics and Computing (ICIC)*, 2023, pp. 1–6.

[33] S. Alsaadi, T. J. Anande, and M. S. Leeson, "Comparative Analysis of 1D-CNN and 2D-CNN for Network Intrusion Detection in Software Defined Networks," in *International Conference on Emerging Internet, Data \& Web Technologies*, 2024, pp. 480–491.

[34] X. Ma *et al.*, "Urban feature extraction within a complex urban area with an improved 3D-CNN using airborne hyperspectral data," *Remote Sensing*, vol. 15, no. 4, p. 992, 2023.

[35] A. Zhang, Z. C. Lipton, M. Li, and A. J. Smola, *Dive into deep learning*. Cambridge University Press, 2023.

[36] S. Kiranyaz, O. Avci, O. Abdeljaber, T. Ince, M. Gabbouj, and D. J. Inman, "1D convolutional neural networks and applications: A survey," *Mechanical systems and signal processing*, vol. 151, p. 107398, 2021.

[37] R. F. R. Junior, I. A. dos Santos Areias, M. M. Campos, C. E. Teixeira, L. E. B. da Silva, and G. F. Gomes, "Fault detection and diagnosis in electric motors using 1d convolutional neural networks with multi-channel vibration signals," *Measurement*, vol. 190, p. 110759, 2022.

[38] M. Khairie, M. R. Faisal, R. Herteno, I. Budiman, F. Abadi, and M. I. Mazdadi, "The Effect of Channel Size on Performance of 1D CNN Architecture for Automatic Detection of Self-Reported COVID-19 Symptoms on Twitter," in *2023 International Seminar on Intelligent Technology and Its Applications (ISITIA)*, 2023, pp. 621–625.

[39] S. Sattar *et al.*, "Cardiac Arrhythmia Classification Using Advanced Deep Learning Techniques on Digitized ECG Datasets," *Sensors*, vol. 24, no. 8, p. 2484, 2024.

[40] A. Kumar, D. Singh, S. Singh, and S. Sharma, "Multiview learning with shallow 1D-CNN for anticancer activity classification of therapeutic peptides," in *Deep Learning Applications in Translational Bioinformatics*, Elsevier, 2024, pp. 79–95.

[41] H. M. Rai and K. Chatterjee, "2D MRI image analysis and brain tumor detection using deep learning CNN model LeU-Net," *Multimedia Tools and Applications*, vol. 80, no. 28, pp. 36111–36141, 2021.

[42] M. R. Faisal *et al.*, "A Social Community Sensor for Natural Disaster Monitoring in Indonesia Using Hybrid 2D CNN LSTM," in

Proceedings of the 8th International Conference on Sustainable Information Engineering and Technology, 2023, pp. 250–258.
[43] N. Q. K. Le, T. T. Huynh, E. K. Y. Yapp, and H. Y. Yeh, "Identification of clathrin proteins by incorporating hyperparameter optimization in deep learning and PSSM profiles," *Computer Methods and Programs in Biomedicine*, vol. 177, pp. 81–88, 2019, doi: 10.1016/j.cmpb.2019.05.016.
[44] N. H. Arif, M. R. Faisal, A. Farmadi, D. Nugrahadi, F. Abadi, and U. A. Ahmad, "An Approach to ECG-based Gender Recognition Using Random Forest Algorithm," *Journal of Electronics, Electromedical Engineering, and Medical Informatics*, vol. 6, no. 2, pp. 107–115, 2024.
[45] J. Bissmark and O. Wärnling, "The sparse data problem within classification algorithms: The effect of sparse data on the na{\"\i}ve Bayes algorithm." 2017.
[46] K. Poulinakis, D. Drikakis, I. W. Kokkinakis, and S. M. Spottswood, "Machine-learning methods on noisy and sparse data," *Mathematics*, vol. 11, no. 1, p. 236, 2023.

AUTHOR BIOGRAPHY

Mohammad Reza Faisal was born in Banjarmasin. Following his graduation from high school, he pursued his undergraduate studies in the Informatics department at Pasundan University in 1995 and later majored in Physics at Bandung Institute of Technology in 1997. After completing his bachelor's program, he gained experience as a training trainer in the field of information technology and software development. Since 2008, he has been a lecturer in computer science at Universitas Lambung Mangkurat while also pursuing his master's program in Informatics at Bandung Institute of Technology in 2010. In 2015, he furthered his education by pursuing a doctoral degree in Bioinformatics at Kanazawa University, Japan. To this day, he continues his work as a lecturer in Computer Science at Universitas Lambung Mangkurat. His research interests encompass Data Science, Software Engineering, and Bioinformatics.

Laila Adawiyah originated in Jelapat, Barito Kuala, South Kalimantan. Since 2018, she has pursued her academic endeavors as a student of Computer Science Department at Lambung Mangkurat University. Her current area of research lies within the realm of data science. Additionally, her final project entailed conducting research that centered around the bioinformatics and protein sequence classification.

Dwi Kartini received her bachelor's and master's degrees in computer science from the Faculty of Computer Science, Putra Indonesia "YPTK" Padang, Indonesia. She is also a lecturer in the Department of Computer Science. She instructs in various subjects such as linear algebra, discrete mathematics, research methods, and others. Her research interests include the applications of Artificial Intelligence and Data Mining. She is an assistant professor in the Department of Computer Science, Faculty of Mathematics and Natural Sciences, Lambung Mangkurat University in Banjarbaru, Indonesia.

Rudy Herteno is currently a lecturer in the Faculty of Mathematics and Natural Science, Lambung Mangkurat University. He received his bachelor's degree in Computer Science from Lambung Mangkurat University and a master's degree in Informatics from STMIK Amikom University. His research interests include software engineering, software defect prediction and deep learning.

Favorisen Rosyking Lumbanraja was born in Lampung, Indonesia. He completed his bachelor's degree in 2007 in Department of Computer Science at IPB University, Bogor, Indonesia, and hist master's degree in Department of Computer Science at IPB University, Bogor, Indonesia, in 2011. In 2014, he furthered his education by pursuing a doctoral degree in Bioinformatics at Kanazawa University, Japan. To this day, he continues his work as a lecturer in Computer Science at Lampung University.

Lilies Handayani was born in Palu, Indonesia. She completed her bachelor's degree in 2011 in Department of Statistics at Hasanuddin University, Makassar, Indonesia, and her master's degree in Department of Statistics at IPB University, Bogor, Indonesia, in 2014. Since 2015, she has been an Assistant Professor in Department of Statistics at Tadulako University, Palu, Indonesia. In 2022, she started her doctoral's degree in Bioinformatics at Kanazawa University, Kanazawa, Japan until now.

Siti Aisyah Solechah was born in Banjarmasin. She received her bachelor's degree from the Faculty of Dentistry, University of Indonesia in 2007. In 2012, he pursued his master's program in the Department of Community Nutrition, Faculty of Human Ecology, IPB University, Indonesia, and graduated in 2014. After completing his master's program, she worked as a freelance translator in the field of Nutrition and Health. From 2020 to 2023, he worked as a lecturer in nutrition science at Husada Borneo Institute of Health Sciences, Indonesia. Since 2023, she has been a lecturer in the Public Health Study Program, Faculty of Medicine and Health Sciences at Lambung Mangkurat University. His research interests encompass Nutrition, especially Community Nutrition.

Triando Hamonangan Saragih is a lecturer in the Department of Computer Science, Lambung Mangkurat University. His research interest is centered on Data Science and Computer Networking. He completed his bachelor's degree in Informatics at Brawijaya University, Malang in 2016. After that, he pursued a master's degree in Computer Science at Brawijaya University, Malang in 2018. The research field he is involved in is Data Science.