Improving Bi-LSTM for High Accuracy Protein Sequence Family Classifier

Roslidar Roslidar[1], Novia Brilianty[2], Muhammad Jurej Alhamdi[3], Cut Nanda Nurbadriani[4], Essy Harnelly[5], and Zulkarnain Zulkarnain[6]
[1,2,3,4]Department of Electrical and Computer Engineering, Universitas Syiah Kuala, Indonesia
[5]Department of Biology, Faculty of Mathematics and Natural Sciences, Universitas Syiah Kuala, Indonesia,
[6]Faculty of Medicine, Universitas Syiah Kuala, Indonesia

Article Info

Article history:

Received Apr 20, 2023
Revised Nov 25, 2023
Accepted Jan 14, 2024

Keywords:

Protein Sequence
Family Classification
Bi-Lstm
Deep Learning

ABSTRACT

The primary nutrient that is crucial for identifying biochemical processes and biological norms in living cells is protein. Proteins are usually centered around one or a few functions which are defined by their family type. Hence, identification and classification are needed to separate the proteins according to their structure and families. In this work, we built a model to classify families of protein sequences. We used the protein sequences dataset consists of various macromolecules of biological significance. The classifier is built up using deep learning of Bi-LSTM. We began the research by collecting the dataset from the Protein Data Bank of the Research Collaboratory for Structural Bioinformatics, pre-processing the data using tokenizing, and modeling the classifier based on deep learning network of Bi-LSTM. As we get the best accuracy rate of the trained model, we figure out the model performance using the evaluation metrics of learning curve, accuracy rate, and loss. The results show that Deep Bi-LSTM provides excellent performance with fit learning curve, 99% accuracy rate, and 0.042 loss.

Corresponding Author:

Roslidar Roslidar,
Department of Electrical and Computer Engineering, Universitas Syiah Kuala
Jln. Teuku Nyak Arief, Darussalam, Banda Aceh, Aceh, 23111, Indonesia
Email: roslidar@usk.ac.id

1. INTRODUCTION

Proteins play important roles in the biology of living things. They are made up of chains of 20 different types of amino acids that are interconnected with each other. Each protein has a varied amino acid makeup, which can be changed through modifications [1, 2]. This sequence is important for determining the function of a protein. One of the earliest ways to figure out what a new protein does is to compare its sequence to the sequences of proteins with known functions. If two proteins have similar sequences, they will likely have similar structures, which can help predict their functions [3].

Proteins have specific tasks, which are determined by their family type. For example, Hydrolase group proteins are responsible for breaking down bonds by adding water, helping to break down chains of proteins or other molecules. On the other hand, transport proteins help other substances such as sugar, fructose, or water enter and exit cells. Unrecognized proteins comprise an important subset of proteins implicated in many bioinformatics research areas [4]. The determination of protein function assays may be one of the primary bioinformatics bottlenecks because of the enormous and continuously growing volume of data from genome sequencing findings. Practically speaking, it is hard to run functional testing on all the uncharacterized proteins received from several ongoing genome sequencing studies [5]. Therefore, a fast and precise computational system is needed in the classification process. To date, deep learning shows the best artificial neural network (ANN) in recognizing the input and performing classification tasks. ANN is flexible and can adapt to solve complex problems not clearly described by mathematical models, such as pattern recognition and classification, function, and control approaches [6, 7].

Nevertheless, studies on developing protein sequence classifiers based on deep learning still expose low performance of learning and accuracy [8-11]. Thus, work on improving the protein classifier model is needed. Our work applied a deep learning network called Recurrent Neural Network (RNN) to develop and process sequential data. We propose a variety RNN of Bidirectional Long Short-Term Memory (Bi-LSTM). LSTMs outperform other types of artificial neural networks in modeling sequences with long-term dependencies because they are better at capturing the long-term temporal structure of the input sequence and Bi-LSTM works better at categorizing sequences data because it has forward and backward layers that help the model learn and improve the classification accuracy. Here, the input sequence includes protein sequences from different important biological molecules.

The main contributions of our work are:

1. Providing a review of existing research in developing a protein sequence classifier.
2. Proposing a framework for preparing the protein sequences dataset for family classification.
3. Providing a step-by-step learning process applying the state-of-the-art Bi-LSTM, introducing the data pre-processing and the embedding layer
4. Investigating of the best deep network model for a high accuracy rate of protein sequence family classification system based on existing RNN models.

The rest of this article consists of related work in Section Two, followed by methods in Section Three that present the data preparation and model development. Section four discusses the simulation results and discussion. Finally, Section five concludes the findings.

2. RELATED WORK

2.1. Protein Sequence

Proteins comprise hundreds or thousands of smaller units known as amino acids. Twenty different kinds of amino acids are linked together by a peptide bond to make a protein molecule [1]. A protein molecule can then be parsed into small fragments and used as fundamental determinants of biological structure and function. Proteins in each organism can have similarities and differences, and these similarities or differences can be used as a matrix to see the relationship between proteins. In order to find areas of similarity that could result from functional, structural, or evolutionary links between the sequences, the main sequences of a protein are arranged in a manner known as sequence alignment [12]. Examples of protein sequence alignment are displayed in Table 1. Data on protein sequences is kept in the protein database for scientific research. The Protein database is a compilation of sequences from several sources, including as records from SwissProt, PIR, PRF, and annotated coding sections in GenBank, RefSeq, and TPA [13].

Table 1. Protein sequence alignments

family_id	sequence_name	family_accession	aligned_sequence	sequence
MORN_2	Q8EI47_SHEON/428-449	PF07661.13	LHGEFRNQTSSGQLLELI.NFNH	LHGEFRNQTSSGQLLELINFNH
Plexin_cytopl	H2TB23_TAKRU/1240-1793	PF08337.12	.MPFLDYKTYTDCNFFLPSKDGAND......AMITRKLQIPE.......	MPFLDYKTYTDCNFFLPSKDGANDAMITRKLQIPEARRAIVAQALN...
RT_RNaseH	H3H8E9_PHYRM/405-501	PF17917.1	DYSRRFHVFADAS.GH.QIGGVIVQ.......................	DYSRRFHVFADASGHQIGGVIVQGRRILACFSRSMTDTQKKYSTME...
Transposase_20	Q981X5_RHILO/224-313	PF02371.16	VEAYQAMRGASFLVAVIFAAEI.GDV.RR.FDTPPQLMAFLGLVPG...	VEAYQAMRGASFLVAVIFAAEIGDVRRFDTPPQLMAFLGLVPGERS...
Mycobact_memb	MMPS4_MYCLE/16-154	PF05423.13	LSRIWIPLVILVVLVVGGFVVYRVHSYFASEKRESYADSNLGSSKP...	LSRIWIPLVILVVLVVGGFVVYRVHSYFASEKRESYADSNLGSSKP...

2.2. Recurrent Neural Networks

In this study, we applied the RNN network of an improved Bidirectional LSTM (Bi-LSTM) and other algorithms of LSTM, Bi-GRU, GRU, and Transformer as a comparison. Bi-LSTM works better at categorizing protein sequences because it has forward and backward layers that help the model learn and improve the classification accuracy. As described by Graves et al. [14], the Bi-LSTM network can access both past (via forward states) and future input features (via backward states) for a given time. Bi-LSTM has been applied in protein sequence identification, but the accuracy rate is minimal [15-18].

Meanwhile, LSTM is the same as RNN, except that the hidden layer updates are replaced by purpose-built memory cells. As a result, LSTM may be better at finding and exploiting long-range dependencies in the data [19]. This network has also been applied in the prediction of epitope regions and shows better performance compared to conventional methods [20]. Other RNN variants for the protein sequence tagging task of the Gated Recurrent Unit (GRU) [21] have also been applied by Li et al. to build Gonet to annotate proteins [22]. While a bidirectional GRU, or BiGRU, is a sequence processing model that consists of two GRU, it was applied by Sharma et al. [23] and Wang et al. [24] using hybrid methods. Transformer [25], the sequence-to-sequence tasks while handling long-range dependencies with ease, is the current state-of-the-art technique in the Natural language processing (NLP) field. This network applied by Cao et al. [26] and Clauwaert et al. [27] for protein sequence labeling. The performance of the networks have motivated us to implement them in this study.

2.3. Existing Protein Sequence Classifier based on ANN

Individual proteins of known sequence and structure present challenges to the understanding of their function. Many function prediction methods rely on identifying sequence and/or structure similarities between a protein of unknown function and one or more well-understood proteins. Alternative methods include inferring conservation patterns in members of a functionally uncharacterized family for which many sequences and structures are known. However, these inferences are tenuous. Such methods provide reasonable guesses at function but are far from foolproof. In Table 2, some recently proposed protein sequence-based artificial neural networks (ANN) were introduced, leaving some future work to improve the accuracy rate.

Table 2. Some studies on protein sequence indetification using ANN

Work	Method	Dataset	Result
Hu et al., 2019 [15]	Bi-LSTM based ensemble model dual loss function	Training data: non-homologous protein sequences extracted from the Protein Data Bank (PDB) database Testing data: 1199, 513 protein Cuff & Barton set (CB513) and 203 proteins from Critical Appraisals Skills Programme (CASP203))	the ensemble model achieved 84.3% in Q3 accuracy and 81.9% in segment overlap measure (SOV) score by using 10-fold cross validation
Bihter DAŞ and Suat Toraman, 2020 [8]	ResNet	Protein sequences were taken from the NCBI dataset. The dataset, which belongs to the supervised dataset, categorizes three proteins such as superfamilies globin 395, trypsin 254, and ras 337	the accuracy is 95.03%
Villegas-Morcillo et al. 2021 [28]	Combination of 1D-convolutional layers with gated recurrent unit (GRU) layers.	uniprot20_2016_02 HHM database	The evaluation results over the well-known LINDAHL and SCOP_TEST sets, along with a proposed LINDAHL test set updated to SCOP 1.75,
Pande and Roy, 2022 [9]	1D Convolutional Neural Network (CNN)	Structural Bioinformatics (RCSB), Protein Data Bank (PDB) dataset	85% with the proposed 1D CNN and 92% after increasing the filter size.
T. Sudha Rani, et al. 2022 [29]	Fast Convolution Neural Network (FCNN)	The 1672 proteins of 25% identified sequence are comprised in 25PDB; 640 proteins of about 25% identified sequence are comprised in 640 PDB; 513 proteins of below 25% identified sequence are comprised in CB513	Accuracy 75%, precision 85%, recall 85%, F1-Score 65%
Aymen Qabel, et al. 2022 [30]	Antibiotic Resistant Genes (ARG) - Graph Neural Network (GNN)	Antibiotic Resistant Genes (ARGs) over 18 antibiotics resistance categories demonstrate	Accuracy 72.90%, F1-Score 63.78%
Nadav Brandes, et al. 2022 [10]	ProteinBERT	8943 most frequent GO annotations that occurred at least 100 times in UniRef90.	Accuracy 74%
Isam Abu-Qasmieh, et al. 2023 [11]	Convolutional Neural Network (CNN)	381 proteins sequences from each family	92.4% accuracy, 94.3% precision, and 91.1% recall.

3. RESEARCH METHOD

The proposed method is based on three major steps for classifying protein sequence families shown in Figure 1. The work started with dataset preparation by collecting and pre-processing the protein sequences dataset using tokenizing. At this stage, a tokenizer is carried out to convert protein sequences into numeric data. Then, the size of the protein sequence was normalized. Next, We developed the model using Bidirectional Long Short-Term Memory (Bi-LSTM). Before inputting the dataset into the network, embedding is proceeded so that each sequence has an embedding space. Finally, the model was evaluated by observing the learning curve and the loss and accuracy rates.

Dataset Preparation
Model Development
Model Evaluation

Figure 1. Research methods

3.1. Dataset Preparation

In this study, we use the Structural Protein Sequences dataset that was downloaded from the Protein Data Bank of the Research Collaboratory for Structural Bioinformatics (RCSB) (PDB). The PDB dataset is a database of protein data that aids in pinpointing the locations of all atoms connected to molecules. It is utilized in techniques like cryo-electron microscopy and NMR spectroscop [31]. The website of PDB dataset is available on the Kaggle website. We used 1.339.083 protein sequences and separated them into three partitions, which were training, validation, and testing. The dataset for training has 1,086,741 sequences, and validation and testing have the same amount of data, which is 126,171 sequences.

In Figure 2, we presented sample distributions of protein sequence classes. PF13649.6 is the protein family with the largest sample spread used in this study. In addition, we provided the distribution of each protein molecule in Figure 3. We can see on distribution of each protein molecule, that L is the amino acid with the highest occurrence rate in this dataset, whether it is in training data, validation or testing.

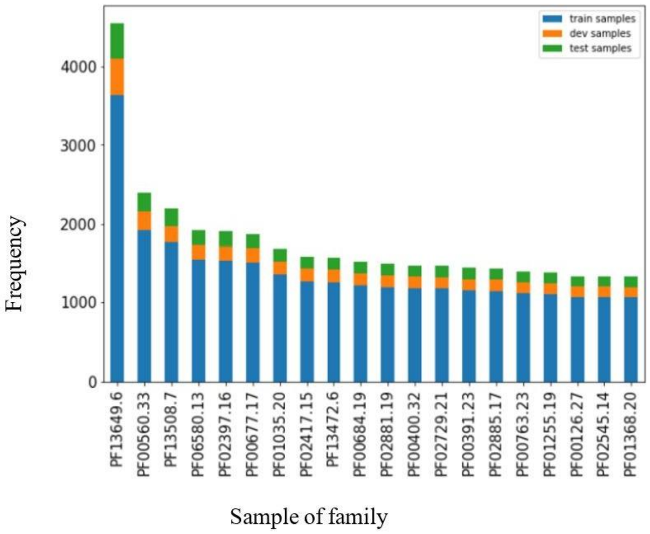
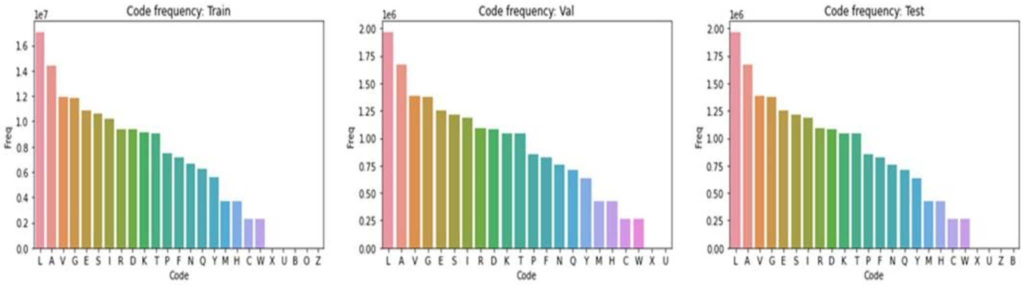


Figure 2. Data distribution protein family accession



(a) (b) (c)

Figure 3. Distribution each protein molecule (a) Frequency distribution of data training, (b) Frequency distribution of data validation, (c) Frequency distribution of data testing.

This value will be added to the total unique code shown in Table 3. Because the ANN does not understand the textual data, it is necessary to convert the total unique code into numerical data. Therefore, each unique code is then entered into the integer encoding stage. Integer encoding means that each unique category value is assigned an integer value (tokenized). For example, ”A” is 1, ”C” is 2, and ”D” is 3, and sequentially until the last unique code. At this stage, data is ready to be fed into the model [31].

Table 3. Integer encoding each unique code

A	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y
1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20

3.2. Model Development

In this study, a classification process of protein sequences is carried out in stages as shown in Figure 4 We proposed several layers to enhance the result. The pre-processed data is conducting using integer encoding, while the model development following several stages of implementing embedding layers, Bi-LSTM, batch normalization, and multilayer perceptron (MLP).

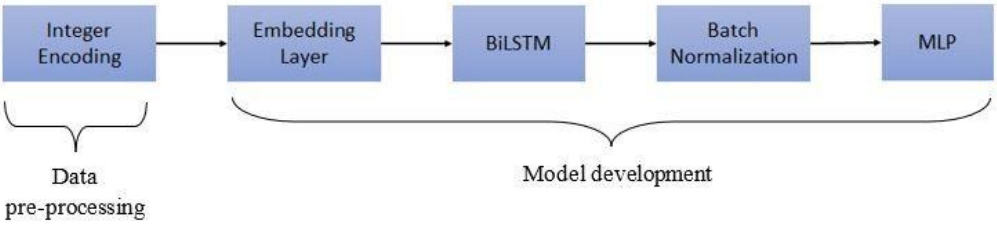


Figure 4. Model development workflow

3.2.1. Embedding Layer

A word embedding is a vector of real numbers that, depending on the context of use, represents a single word. The ability to map every word in a vocabulary to a point in a vector space is made possible by the numerical representation of words [12]. The vector is not generated through mathematical calculations, instead each input integer serves as an index to access a table that holds all the potential vectors. The use of an embedding layer results in smaller input sizes and decreased computational complexity, leading to a faster training process. As seen in Figure 5, according to the ”distributional theory,” words that exist in a related context have related meanings. Therefore, it is anticipated that, compared to unrelated words, the embeddings for semantically or syntactically similar words will be positioned closer to each other in the vector space. The level of relatedness is solely based on the text data, also known as corpus, used to generate the embeddings.

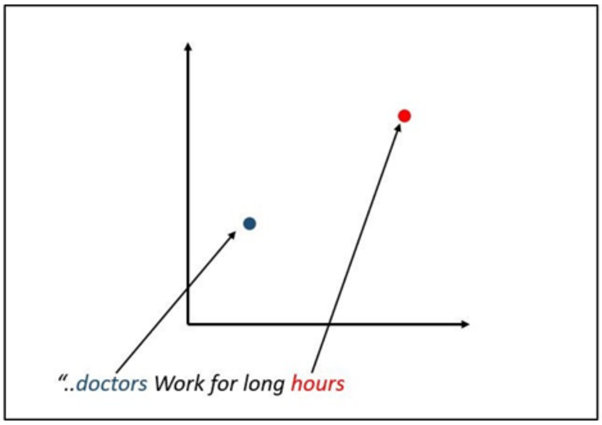


Figure 5. Visualization of word embedding in 2D space [32]

3.2.2. Classification using Bi-LSTM

An improved variant of LSTM is called Bi-LSTM. A Bi-LSTM is created by superimposing two LSTMs. To fully understand the information of the entire input sequence in Bi-LSTM, not simply the information prior to the present time, two LSTM networks that input are forward and backward of input sequences, respectively must be trained. Memory blocks are used in the Bi-LSTM network in place of the summation units found in the traditional RNN’s hidden layer. The Bi-LSTM architecture is made up of a collectionof memory blocks, also known as recurrently linked subnets, as shown in Figure 6. Three multiplicative units (the input, output, and forget gates) offer continuous analogs of write, read, and reset operations for the cells, and each block comprises one or more self-connected memory cells. The cell is activated by means of multiplications by the three gates, which are nonlinearsummation units that gather activations from both within and outside the block (small black circles). Bi-LSTM memory cells can retain and retrieve information for extended periods of time because of the multiplicative gates, which alleviates the vanishing gradient issue [13].

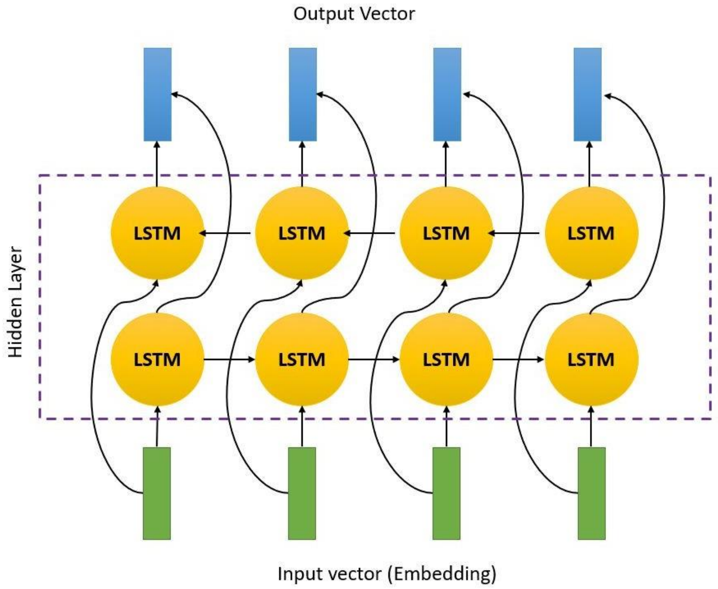


Figure 6. Bi-LSTM network

In both the forward and backward propagation directions, the Bi-LSTM spans two parallel LSTM layers, as illustrated in Figure 5. Internal state information is kept in $h_{f(t)}$ from previous time series values in the forward direction and from future sequence values in the backward direction. To produce the final output, the two distinct hidden states at time step t, $h_{f(t)}$ and $h_{b(t)}$, are connected one after the other. Following are the two directional layers' recurrent states for the Bi-LSTM network at time step t.

$$h_{f(t)} = \varphi(W_{fh}\, xt + W_{fhh}\, h_{f(t-1)} + b_{fb}) \quad (1)$$

$$h_{b(t)} = \varphi(W_{bh}\, xt + W_{bhh}\, h_{b(t+1)} + b_{b}) \quad (2)$$

where weight matrices W_{fh} and W_{bh} denote the forward and backward weights from input to recurrent units, respectively. Concurrently, the terms b_{fb} and b_b denote bias signals in both directions, respectively. The symbol ϕ denotesthe recurrent layer activation function and is set to tanh in this paper. The Bi-LSTM generates a final output vector which is displayed as follows:

$$Y_t = \sigma(W_{fhy}\, h_{f(t)} + W_{bhy}\, h_{bt} + b_y) \quad (3)$$

where W_{fhy} and W_{bhy} denote the forward and backward weights from the internal unit to the output, respectively, and σ denotes the output layer acti-vation function, which is set to sigmoid or linear functions, and by represents the bias vector of the output layer [13].

After defining the model, we proceed to create the Bi-LSTM workflow as illustrated in the Figure 7 with architecture of the proposed method shown on Figure 8. In this work, we use 2-layer of Bi-LSTM, to make model more complex and make it easy to recognize the class of protein sequences. On each layer Bi-LSTM we use batch normalization to deal with the convergence speed by improving the stability and consistency of activations during train-ing and to reduce exploding gradient.

In addition, we used simple Multilayer Perceptron (MLP) for predicting the class of protein sequences by using probability each node [14]. MLP is a type of artificial neural network that is designed to solve supervised learning problems. It is a feed-forward network, which means that the information flows in a single direction, from input to output, without looping back. The network is composed of multiple layers of artificial neurons, each of which is connected to the next. For learning,it makes use of the backpropagation method. It consists of neurons that serve as receivers in the input layer, one or more hidden layers of neurons that process the data iteratively, and neurons that anticipate the outcome inthe output layer [36]. At least three layers make up MLP: an input layer, a hidden layer or layers, an output layer, and an input layer [37]. In this work,we use 2 layers which are the layer to receive the output of Bi-LSTM andthe layer to classify the protein sequence classes.

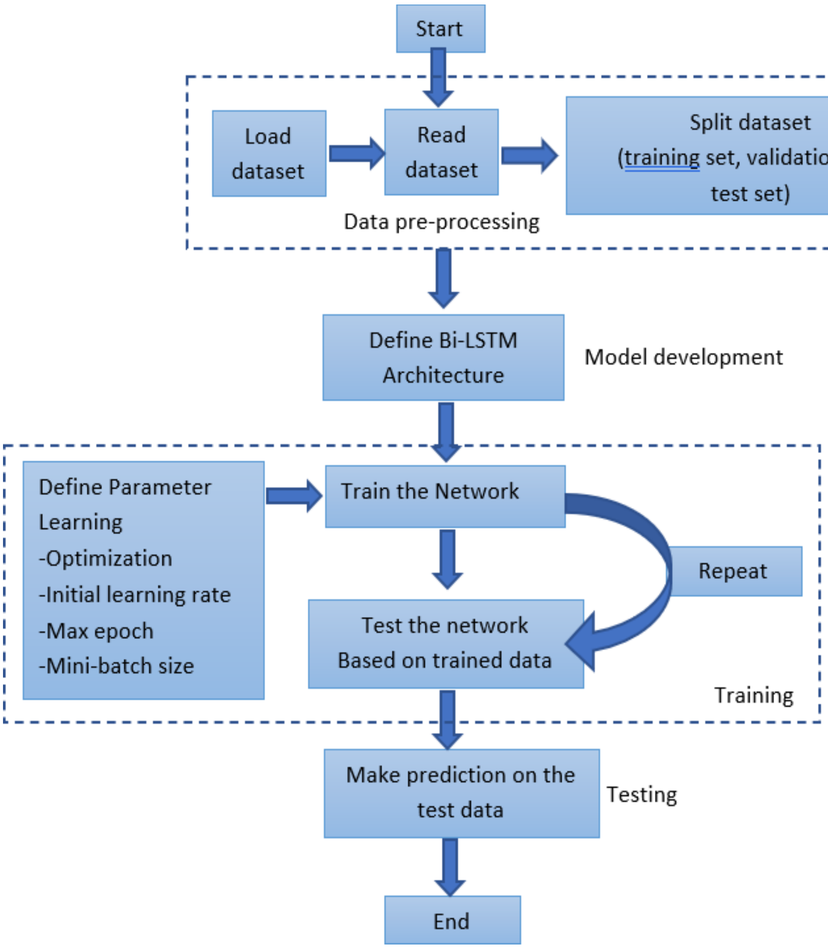

Start
Load dataset
Read dataset
Split dataset (training set, validation set, test set)
Data pre-processing
Define Bi-LSTM Architecture
Model development
Define Parameter Learning
-Optimization
-Initial learning rate
-Max epoch
-Mini-batch size
Train the Network
Repeat
Test the network Based on trained data
Training
Make prediction on the test data
Testing
End

Figure 7. Model development workflow [35]

3.2.3. Training Process

Training is used so that the system learns and builds its path from existingpossibilities. Usually, the training data process is done by 80% of the existingdataset. The training and validation processes are carried out sequentially ineach epoch or iteration, and after each training, it is followed by a validation process.

The training deep Bi-LSTM model is based on the label set Y_{train} and the training set X_{train} as equations 4 and 5, respectively. M and N are thelengths of the training and test portions of the protein sequences, respectively, and M and N are the numbers of the training and test samples, respectively[15]:

$$X_{train} = [t_1, t_2, \ldots, t_m]^T \tag{4}$$

$$Y_{train} = [p_1, p_2, \ldots, p_m]^T \tag{5}$$

For network training, we use Adam optimizer with learning rate of 0.01, loss function of crossentropy, with 100 epoch and 250 batch size. The epoch was performed 100 times. Epoch determines how often the deep learning algorithm works through the entire forward and backward dataset. Then, we can see the progress and, at the same time, make the machine learn new possibilities from the given iteration limit. Limits are also used so that the computational process does not take too long.

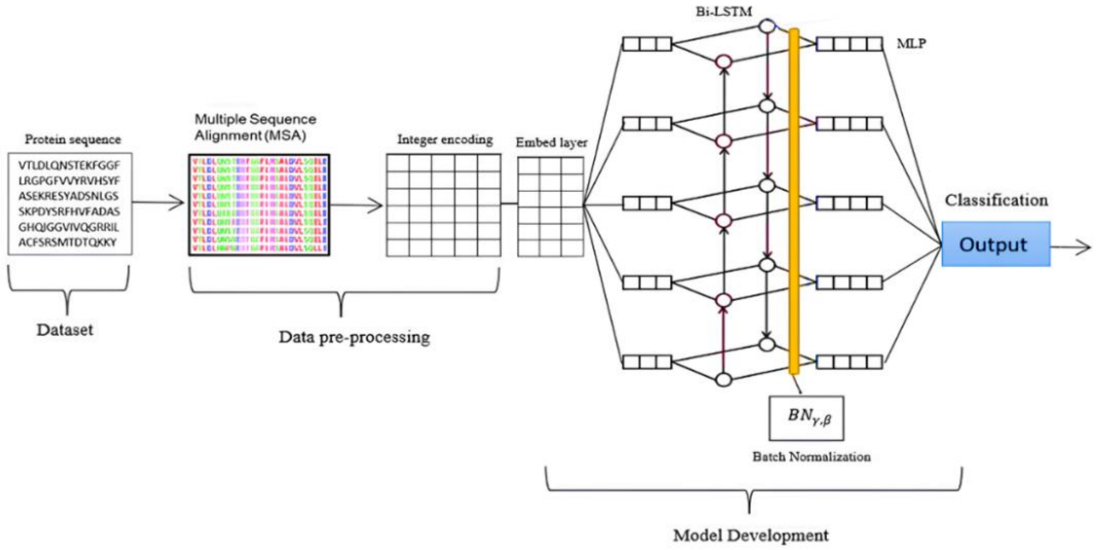


Figure 8. The architecture of the proposed network

3.3. Model Evaluation

After we get the training result, we recorded the learning curve to see the convergence and tested the model. Testing is used to figure out the model accuracy rate in recognizing new data. The testing dataset used in this study was 20% of the existing dataset. The testing set and label set are:

$$X_{test} = [t_{m+1}, t_{m+2}, \ldots, t_{m+n}]^T \tag{6}$$

$$Y_{test} = [p_{m+1}, p_{m+2}, \ldots, p_{m+n}]^T \tag{7}$$

The test accuracy rate shows the capability of the model to predict a dataset that model never seen before. We compared the performance of improved Bi-LSTM to other RNN algorithms of LSTM, GRU, Bi-GRU, and Transformer. Finally, we also compared our proposed model with other works on the similar task.

4. RESULTS AND DISCUSSION

The classification model of the protein sequence family built in this study is a good fit, as shown in Figure 9. It shows that our improved Bi-LSTM model is not overfitting or underfitting, as indicated by the accuracy of training and validation accuracy. It shows that the value continues to increase consistently and does not decrease. Similarly, losses in training and validation consistently decrease, approaching zero. As we increased the number of epochs, we could improve the model performance and stability. In terms of accuracy rate, the accuracy of training, validation, and testing is 0.9987, 0.9968, and 0.9913, respectively. Meanwhile, the values of training, validation, and testing loss are 0.0373, 0.042, and 0.044, respectively. Table 4 shows the accuracy and loss rates of the model.

The model can successfully predict the protein sequences and their classes or families from the given sequence data. The model can predict the family accession of protein sequences with a high accuracy rate. To evaluate the correctness, we also span the table on each prediction with boolean parameters, ”true” or ”false”. The results show that the model can predict most protein classes with a ”true” status, as shown in Table 5. However, more work is needed to deal with the “false” protein prediction.

Table 4. Model Performance

	Loss	Accuracy
Training	0.037	0.9987
Validation	0.042	0.9968
Test	0.044	0.9913

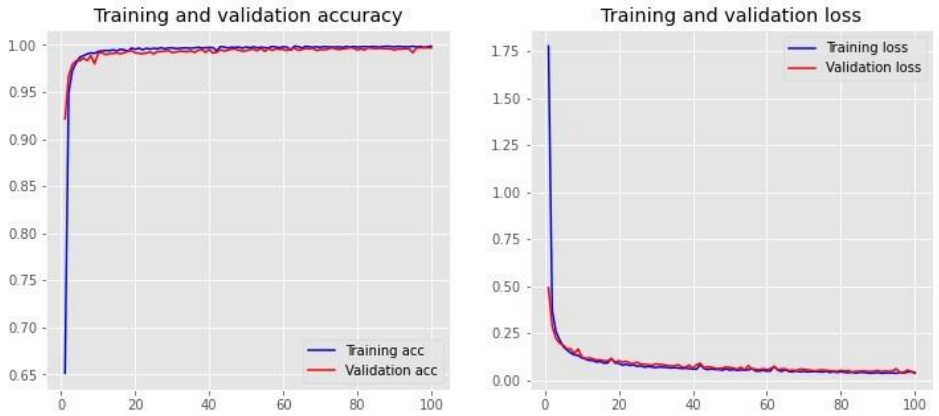


Figure 9. Learning performance of the proposed model

Table 5. Evaluation of protein sequences classification

No	Protein Sequence	True Class	Predict Class	Evaluation
1	AVRRHIRSSPLKMRRVINLVRDRSVPEAVAILDYMPQKVTGVVEKTIRSAVYNLMDQHDERFDEGALKLKEIRADEGPTFQRHQARARGRAAPIRKRTTHLKVVVA	PF00237.19	PF00237.19	True
2	INVSGSGEIMAKPDIAYLSIGVQSEGNTAAAAQKANAAKINKVTQLLKEKWSISADDIQTSQFSVQPNYTYNEKEGQKLKGYMANHTLSVKYRNLDKIGQLLDEATNSGANNVDNIQFSVENPSAYEEAAIAKALDNAQSKASAVAKSAKRGLGALVNVTVDGGEAQVYTQRENAMSKALMDMSGGTEIQSGQVTVKVQVSAQY	PF04402.14	PF04402.14	True
3	DIAGCEISAALKNAIAIGGGILKAYNAGDNAHATLLTLGLNEMYEFGKHFGAKLETFLNFAGLGDLILTASSKKSRNFRLGERIVELNDAKKALESFNLTVEGVETARIAHEIGVKYQISMNFFEIIYNILYNNVKPISLL	PF07479.14	PF07479.14	True
4	ATITITVSSVNEFPSAVNDTATTNEDSPISGNLATNDSPSKDGGNVWSLVGDNGGALHGTIAMTPDGSYTY	PF17803.1	PF17803.1	True
5	KRLYRSYTDKMLGGVCGGLGEYFDIDPVIIRVLFVVAVLFGGGGILAYIILWIVIPQK	PF04024.12	PF04024.12	True
6	SVPRLKKAEGMDLKMFPHLVDDETFKQIVAILRIAVPFTGIILSTRESAEMRKEVIEYGVSQVSAGSCAGVGGYKEREEGKNTNQFIIADHRSHLDVLKELIEEGHI	PF06968.13	PF06968.13	True
7	DEAKQIAEKLKTICPKLKIKTGENGKVFGGVTAKDIADVLNKEYKINIDKKKIDLKESIKTLGVTKVNVKLYEGVMGEIKIDVIP	PF03948.14	PF03948.14	True
8	FSVGDHIVYPLQGVGIIKCIEERNFQGEPQPYYVIHIAISDMIVKIPIAKAAEMGIRAIVPPSEAQEAIDSISSKYDPLPVDWKTRYQMNVDLLQQGSIASIAQVVQAL	PF02559.16	PF02559.16	True
9	DITDGVNALISKAKMQQGLCHLFVPHTTAGITVNENCDPDVARDILATLDQLVPIHGNYRHAEGNSHAHLKASLMGHALSVPIEGGKLLLGTWQGLYFTEFDGPRARKVILTL	PF01894.17	PF01894.17	True
10	IAAGICVGIAACGAGIGMGLTTAKSSEGISRQPEAADKIRTNMMLGLVFIETAIIYALVVAILI	PF00137.21	PF00137.21	True
11	DYNSDGNTDIAAVYMDEKSLNHIAFYTNDGSGNLQSYPSLLSTGFGEKTVIASGDINMDGQPDLIV	PF13517.6	PF13517.6	True
12	KTCKPVPNTRFGFYTKDNIYDIDGNVILEADSKITTVTTGADGTAKIPFSVPVMSEGYGEVEAPLNSGDYYFLEESVSDSYYISEEPTFVHLEYEN	PF17802.1	PF17802.1	True

13	AAVALVIDQASKYWILHDVLEDKAMIIFTPFFSLVRAWNTGVSFSMFNDWGLSGVIVLSLVAFVIIAFLVNWLRKEPSKLIQVSLGLIIGGALGNVIDRIRLGAVFDFLDFSIGTYHWPAFNAADSFICVGALIVIFHGLW	PF01252.18	PF01252.18	True
14	RGVTFFKATGAFSREDKQVVYCVISRTQLSQIKEIIHTRDDEAFLAISEVPEVVG	PF10035.9	PF10035.9	True
15	PWRTAVKIATGAATPLPPSVVPSSEATGRVLAAPVLARGDLPGFDASAMDGYAVAGRGPWRVLGQVYAGGPVWPAPLGAGEAVEIMTGAVVPAGTVAVLPYETADLNAAGPGPKAHIRRAGEDARAGDELLPAGRLVTPAVAGLLAQAG	PF03453.17	PF03453.17	True
16	ASQIARPFRHSLKKKLDKCTSKPTLIGLLANQDPASKKYAEWTAKTCQETGVQFELAQVNKHELELEIIKANHNKDIHGIMVYYPVFGYPLDMSLQNRVNYFKDVE	PF00763.23	PF02410.15	False
17	KWRKNDQDTGSPRIVVAILTEKIVYLTKHMQQHPHDYHSRRGLITMVNKRRRQLNYYFKKEPKECLEMCATLGIR	PF00312.22	PF13976.6	False
18	LDYVVHRQYGIGKYMGITTKEIEGIHKDFMRILYRDGDELFVPLEQFNLVRKFMSREAASVRLSKLGTSTWQKNKERIKQDVADVADKLVTLY	PF02559.16	PF02021.17	False
19	ETIASARQRMQHRNPGGVCNRDLPATALGDRSGFGAAALQLWERLVAHRGLSTRSGIRLLRVARTVADLNGENEASADAVAQASHYR	PF13335.6	PF10589.9	False
20	FIATDLTNFDRWIFIPPKFRMKCTEGCYITAKVTQHPFKDGRAQAKITQNVGDDNTPYIEKLYSVCKHRLDNEFS	PF17876.1	PF14698.6	False
21	HSVVMHITTGVCRIQDIQEKQFTEDQHQKYYVLQPIFEKGTTLFVPIENDPVRIRPLLTKEAITELLHELSAQEDEPWIHNQHQRTAHFKTILKNGNEQEILSMLH	PF02559.16	PF02033.18	False
22	QYNRLRSGMDDVQRRLAELRASADSDDGLIRATVGPRGQLLDLRLDRRIYRDMDAAELSRKIVTTVEQATAKATQQVEQLMADY	PF02575.16	PF02583.17	False
23	QVEIVTVSRLVEKKGVEYGIRAVAKLLKNQKKNINYTIVGDGPLKESLQELVQQLDVANHVQLLGSKQQQEVIEILKNSHIMLAPSVTSSNGDQEGIPVALMETMAMGLPVVSTFHSGIPELVEDGVAGFLVPERNVDALAEKIGYLVEH	PF13692.6	PF05697.13	False

Table 6. Benchmarking of several models trained on the same dataset

Model	Validation Accuracy	Validation Loss	Testing Accuracy	Elapsed Time
Improved Bi-LSTM	**99.68%**	**0.042**	**99.13%**	135’
LSTM	98.47%	0.061	98.38%	69’
Bi-GRU	98.55%	0.054	98.66%	91.2’
GRU	96.84%	0.112	98.08%	**55’**
Transformer	98.89%	0.051	98.37%	270’

We trained and tested the models using other RNN algorithms of LSTM, BiGRU, GRU, and Transformer. Each algorithm was run with 100 epochs and a 0.01 learning rate to build the model. Similar hyperparameter tuning was performed, resulting in the performance as shown in Table 6. The result shows that Bi-LSTM outperforms other models regarding loss and accuracy rate. Compared with the existing studies in Table 2, our improved Bi-LSTM achieves higher performance. Nevertheless, more work needs to be conducted to improve the accuracy rate of the protein sequence classifier.

For future work, we recommend the following strategies to improve the classifier performance:

1. More representative dataset. Dataset preparation is an important key to success in building a prediction model. The network will learn better if fed with the representative dataset.
2. Merging the well trained-model. With deep analysis of the trained model performance, the well-trained model can be merged to get a better model by considering the learning fit.
3. Fusion the result. Applying the fusion strategy, especially at the stage of getting the result, may enhance the definite decision of the learning.

4. Model deployment. The high-performance model could be deployed into bioinformatics devices. Integrating the model into bioinformatics devices or application platforms will allow for high usage of protein sequence classifiers and ease protein sequence work.

5. CONCLUSION

This study presented an improved Bi-LSTM model for classifying protein sequences. The model utilizes well-known properties with an improved formula. With a test accuracy rate of 99.13%, improved Bi-LSTM outperforms previous protein sequence classifier models in classifying the protein sequence. Thus, our proposed improved Bi-LSTM could be used to solve various computational biology problems, such as DNA classification, virus classification, and other biological recognition cases. For future work, more representative datasets should be added to improve the classifer accuracy. Implementing a lightweight neural network would also improve the stability of the model and allow the deployment of the predicted model in any bioinformatics devices or application platforms.

REFERENCES

[1] B. Alberts, D. Bray, K. Hopkin, A. Johnson, J. Lewis, M. Raff, K. Roberts, P. Walter, *Protein structure and function, Essential cellbiology,* pp.120–170, 4th Ed, New York, Garland Science, Taylor and Francis Group, 2010.

[2] A. K. Wong, E.-S. A. Lee, “Aligning and clustering patterns to reveal theprotein functionality of sequences”, *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol.11, no.3, pp 548–560, 2014.

[3] Z. Zulkarnain, H. Sujuti, D. W. Soeatmadji, D. H. Utomo, et al., “Tshr169 antigen specifically binds to the thyroid-stimulating autoan- tibody, representing an effective biomarker for graves’ disease”, *International Journal Bioautomation*, vol. 23, no. 1, pp 51-60, 2019.

[4] D. Zhang, M. R. Kabuka, “Protein family classification with multi-layergraph convolutional networks”, *2018 IEEE International Conference on Bioinformatics and Biomedicine (BIBM),* IEEE, pp. 2390–2393, 2018.

[5] L. Deng, D. Yu, et al., “Deep learning: methods and applications”, *Foundations and trends® in signal processing* vol.7, no.3–4, pp. 197–387, 2014.

[6] S. J. Giri, P. Dutta, P. Halani, S. Saha, “Multipredgo: deep multimodal protein function prediction by amalgamating protein structure, sequence, and interaction information”, *IEEE Journal of Biomedical and Health Informatics*, vol. 25, no.5, pp. 1832–1838, 2020.

[7] R. S. Singh, D. J. Gelmecha, S. Mishra, G. Dengia, D. K. Sinha, “A novel machine learning approach for detection of coronary artery dis- ease using reduced non-linear and chaos features”, *International Journal Bioautomation*, vol. 26, no. 3, 2022.

[8] Bihter, D.A.Ş. and Toraman, S., “Classifying protein sequences using convolutional neural network”, *Bitlis Eren Üniversitesi Fen Bilimleri Dergisi*, vol. 9, no. 4, pp.1663-1671, 2020.

[9] Pandey, A. and Roy, S.S., “Protein sequence classification using convolutional neural network and natural language processing”. In *Handbook of Machine Learning Applications for Genomics* (pp. 133-144). Singapore: Springer Nature Singapore. 2022.

[10] Nadav Brandes, Dan Ofer, Yam Peleg, Nadav Rappoport, Michal Linial, “ProteinBERT: a universal deep-learning model of protein sequence and function”, *Bioinformatics*, vol. 38, no. 8, pp 2102–2110, 2022.

[11] Abu-Qasmieh, I., Al Fahoum, A., Alquran, H. and Zyout, A., “An Innovative Bispectral Deep Learning Method for Protein Family Classification”. *Computers, Materials & Continua*, vol. 75, no.2, 2023.

[12] UniProt, *Aligning multiple protein sequences*, https://www.ebi.ac.uk/training/online/courses/uniprot-exploring-protein-sequen, 2022.

[13] NCBI, Protein, https://www.ncbi.nlm.nih.gov/protein, 2022.

[14] Graves, Alex, Abdel-rahman Mohamed, and Geoffrey Hinton. “Speech recognition with deep recurrent neural networks”. *2013 IEEE international conference on acoustics, speech and signal processing*. IEEE, 2013.

[15] Hu, H., Li, Z., Elofsson, A. and Xie, S., “A Bi-LSTM based ensemble algorithm for prediction of protein secondary structure”. *Applied Sciences*, vol. 9, no. 17, p.3538, 2019.

[16] Jin, H., Du, W., Gu, J., Zhang, T. and Shi, X., “Combining GCN and Bi-LSTM for protein secondary structure prediction”. *2021 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)* (pp. 44-49). IEEE, 2021.

[17] Sharma, A.K. and Srivastava, R., “Protein secondary structure prediction using character bi-gram embedding and Bi-LSTM”. *Current Bioinformatics*, vol. 16, no. 2, pp.333-338, 2021.

[18] Ema, R.R., Khatun, A., Hossain, M.A., Akhond, M.R., Hossain, N. and Arafat, M.Y., “Protein Secondary Structure Prediction using Hybrid Recurrent Neural Networks”. *Journal of Computer Science*, vol. 18, no. 7, pp.599-611, 2022.

[19] Hochreiter S, Schmidhuber J. “Long short-term memory”. *Neural Computation*, vol.9, pp. 1735-1780, 1997.

[20] Noumi, T., Inoue, S., Fujita, H., Sadamitsu, K., Sakaguchi, M., Tenma, A. and Nakagami, H., “Epitope prediction of antigen protein using attention-based LSTM network”. *Journal of Information Processing*, vol. 29, pp.321-327. 2021.

[21] Chung J, Gulcehre C, Cho K, Bengio Y., “Empirical Evaluation of Gated Recurrent Neural Networks on Sequence Modeling”, arXiv; arXiv:1412.3555. 2014.

[22] Li, J., Wang, L., Zhang, X., Liu, B. and Wang, Y., “Gonet: a deep network to annotate proteins via recurrent

convolution networks". *2020 IEEE international conference on bioinformatics and biomedicine (BIBM), IEEE*, pp. 29-34, 2020.

[23] Sharma, L., Deepak, A., Ranjan, A. and Krishnasamy, G., "A novel hybrid CNN and BiGRU-Attention based deep learning model for protein function prediction", *Statistical Applications in Genetics and Molecular Biology*, vol. 22, no. 1, p.20220057, 2023.

[24] Wang, Z., Lin, T., Yang, X., Liang, Y. and Shi, X., "Protein Subcellular Localization Prediction by Combining ProtBert and BiGRU", *2022 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, IEEE, pp. 86-89, 2022.

[25] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł. and Polosukhin, I., "Attention is all you need". *Advances in neural information processing systems*, vol. 30, 2017.

[26] Cao, Y. and Shen, Y., "TALE: Transformer-based protein function Annotation with joint sequence–Label Embedding", *Bioinformatics*, vol. 37, no. 18, pp.2825-2833, 2021.

[27] Clauwaert, J. and Waegeman, W., "Novel transformer networks for improved sequence labeling in genomics", *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol. 19, no. 1, pp.97-106, 2020.

[28] A. Villegas-Morcillo, A. M. Gomez, J. A. Morales-Cordovilla and V. Sanchez, "Protein Fold Recognition From Sequences Using Convolutional and Recurrent Neural Networks," *in IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol. 18, no. 6, pp. 2848-2854, 1 Nov.-Dec. 2021.

[29] Rani TS, Babu AY, Haritha D. Wrapper, "Fuzzy Approach with 3d Fast Convolution Neural Network (FCNN) Based Feature Selection in Protein Sequence Classification", *International Journal of Intelligent Systems and Applications in Engineering*, vol. 10, no. 2s, pp. 28–34, 2022.

[30] Qabel A, Ennadir S, Nikolentzos G, Lutzeyer JF, Chatzianastasis M, Boström H, Vazirgiannis M. "Structure-Aware Antibiotic Resistance Classification Using Graph Neural Networks". *InNeurIPS 2022 AI for Science: Progress and Promises*, 2022.

[31] J. Brownlee, Ordinal and one-hot encodings for categorical data, https://machinelearningmastery.com/one-hot-encoding-for-categorical-data/, 2020.

[32] C. Spearman, *The proof and measurement of association between twothings*. 1961.

[33] H. Wang, Y. Zhang, J. Liang, L. Liu, "Dafa-bilstm: Deep autoregression feature augmented bidirectional lstm network for time series prediction", *Neural Networks*, vol. 157, pp. 240–256, 2023.

[34] R. Wang, X. Liang, X. Zhu, Y. Xie, "A feasibility of respiration predic- tion based on deep bi-lstm for real-time tumor tracking", *IEEE Access*, vol. 6, pp. 51262–51268, 2018.

[35] R. Roslidar, M. Syaryadhi, K. Saddami, B. Pradhan, F. Arnia, M. Syukri, K. Munadi, R. Roslidar, M. Syaryadhi, K. Saddami., "Breacnet: A high-accuracy breast thermogram classifier based on mobile convolutional neural network", *Math. Bioscience Enginering*, vol. 19, pp.1304–1331, 2022.

[36] M. Desai, M. Shah, "An anatomization on breast cancer detection and diagnosis employing multi-layer perceptron neural network (MLP) and convolutional neural network (CNN), *Clinical eHealth*, vol. 4, pp. 1–11, 2021.

[37] A. C. Hughes, M. Mort, L. Elliston, R. M. Thomas, S. P. Brooks, S. B. Dunnett, L. Jones, "Identification of novel alternative splicing events in the huntingtin gene and assessment of the functional consequences us- ing structural protein homology modelling", *Journal of molecular biology*, vol. 426, no. 7, pp. 1428–1438, 2014.

BIOGRAPHY OF AUTHORS

Roslidar received her Bachelor Degree in Electrical Engineering in 2001 from Universitas Syiah Kuala. In 2009 she graduated from the Master programme in Telecommunication Engineering, University of Arkansas, USA, under Fulbright scholarship. In January 2022, she received her PhD in Doctoral School of Engineering Science, Universitas Syiah Kuala.

Since 2001 she has been the lecturer and researcher at the Department of Electrical and Computer Engineering in Universitas Syiah Kuala. Her research interest is developing the e-health monitoring system based on thermal imaging. She is also active in any research related with electrical engineering and deep learning.

Novia Brilianty is a Computer Engineering undergraduate, at the Faculty of Engineering, Universitas Syiah Kuala, currently in 3rd year. She was a lecture assistant and laboratory assistant from 2022-2023. Her research interest focused on computer vision and deep learning in the bioinformatical issue.

Muhammad Jurej Alhamdi received his bachelor degree from the Department of Electrical Engineering, Faculty of Engineering, Universitas Syiah Kuala, Banda Aceh in 2022. He was a Research Assistant at the control system Laboratory in 2020-2021. His research interests include creating and designing autonomous car using Deep Learning, and application of deep learning in the world of health, a. Currently, he is pursuing her study in Master of Electrical Engineering, Universitas Syiah Kuala

Cut Nanda Nurbadriani received her bachelor degree from the Department of Electrical Engineering, Faculty of Engineering, Universitas Syiah Kuala, Banda Aceh in 2022. She was a Research Assistant at the electronics and digital logic Laboratory in 2019-2022. Her research interests include creating and designing microcontroller-based prototypes for medicine purposes. Currently, she is pursuing her study in Master of Electrical Engineering, Universitas Syiah Kuala

Essy Harnelly, an associate professor in Biology Department, Mathematics and Natural Science, Universitas Syiah Kuala. She finished the Ph.D. program at Georg-August University, Germany. Her interest is in plant molecular biology especially medical plants. Plant adaptation and timber tracking

Zulkarnain worked as a senior lecturer and a researcher at the Physiology Department of Medical Faculty, Universitas Syiah Kuala. He passed his Ph.D. in 2020 at the Medical Science Program, Faculty of Medicine, Universitas Brawijaya, Malang, East Java. His research activities focused on preventive medicine, autoimmune biomarker, recombinant protein, and drug discovery