RESEARCH ARTICLE

Editorial Process: Submission:11/24/2024 Acceptance:11/15/2025 Published:11/22/2025

Prediction of Proto-Oncogene Using Bidirectional GRU and Attention

Seeja R.D[1]*, Bernice Rufus A[2]

Abstract

Objective: One of the key responsibilities of bioinformatics is now protein sequence prediction, thanks to the advancements in genome sequencing technology. The primary means of uncontrolled cancer growth is the absence of tumour suppression gene (TSG) regulatory ability and proto-oncogene (OG) mutations. Even though a cancer is a complicated mixture of several disorders, computational research may be able to identify genes linked to OG or TSG activity, which may help with the creation of drugs that target the condition directly. **Methods:** Recently, the attention mechanism in deep learning has emerged as a cutting-edge method for protein sequence classification. The attention-based strategy can provide a reliable and comprehensible way to help overcome current challenges in characterising deep neural networks for protein sequence classification. This study proposes two approaches like Attention with Convolutional Neural Network (ACNN) and Attention with Bi directional Gated Recurrent Units (ABiGRU) to predict Proto-oncogene protein sequence. The proposed deep learning with Attention model is validated using Independent test and K-fold cross-validation test. Moreover this study has performed Ablation study and Statistical significant Testing to access the superiority of the proposed model. **Results:** The results are analyzed by the benchmark Uniprot dataset. Independent testing of ACNN model gives 96.85% of accurate results and ABiGRU model gives 97.53%of accurate results. **Conclusions:** According to these findings, the suggested model may be crucial in determining a cancer patient's early prognosis and in helping researchers identify cancer-fighting systems.

Keywords: Bi-directional Gated Recurrent Units (BiGRU)- Proto-oncogenes- Deep learning- Attention Mechanism

Asian Pac J Cancer Prev, **26 (11)**, 3959-3966

Introduction

The gene, the smallest unit of DNA, is a two-fold helix particle composed of direct sets of nucleotides [1]. The building blocks of each nucleotide are the gene bases. Each gene is made up of a sequence of nucleotide bases that convey information about how cells develop and function. In essence, this occurs when the genetic information is translated into proteins by the cells. Every protein in the human body serves a specific function. Common cellular genes called proto-oncogenes regulate human cell division and development [2]. The lack of control over the cell cycle has long been linked to cancer. A sequence of genetic alterations leading to the inactivation of tumour-suppressing genes and the activation of proto-oncogenes into oncogenes causes the loss of control.

The process of activation, which includes insertion mutations, point mutations, protein-protein interactions, retroviral transduction, gene amplification, chromosomal translocation, and transposon integration, can turn proto-oncogenes into oncogenes. Proto-oncogenes are frequently classed according to how closely their sequences resemble those of known proteins or according to how they typically behave inside cells [3]. The study of the genes associated with the onset of cancer is known as oncogenomics. In transformed cells, proto oncogenes are often activated by point mutations or gene amplification [4]. The discovery of these genes may provide new insights into the aetiology and management of cancer [5]. These genes may also play a part in the genesis of cancer.

Given the relationship between the impacts of mutations on gene activity, oncogenes are believed to be identifiable from other genes based on their specific mutation profile [6]. Finding new oncogenes other than those that are often mutated are difficult because of the considerable variability of mutations across persons and different kinds of cancer [7]. For this reason, developing computational methods for the discovery is imperative.

Protein sequence analysis has been the subject of a great deal of research in recent years due to its many applications in protein bioinformatics and medical proteomics, as work by [8] shows. Studying protein

*[1]SRMIST, Department of Computer Application,Faculty of Science and Humanities, Kattankulathur - 603 203, Chengalpattu District, Tamil NaduIndia. [2]A. Bernice Rufus, MPhil Scholar, Scott Christian College, Nagercoil, India. *For Correspondence: sheejarufus.r.d@gmail.com*

sequences with the primary objective of predicting the structures and functions of proteins is known as in silico protein sequence characterization. It has been shown in recent research that protein sequence comparisons are more accurate than direct DNA comparisons. In order to identify members of the same protein superfamily that are related to one another physically, functionally, and historically, protein sequence classification is essential to protein sequence analysis. The benefit of accurately classifying a member protein sequence as belonging to a superfamily is that, instead of studying the sequences of each individual member protein, it just necessitates doing some molecular studies within that superfamily.

To further detect oncogene from an amino acid sequence, many techniques have been developed [9]. Generally, the study of proto-oncogene protein sequence identification covers a wide range of activities, not just those connected to cancer. In order to overcome the shortcomings of the most recent innovative work, this study attempts to make major advances in the field of proto-oncogene prediction. In order to treat and cure the oncogene, this may help identify it early on.

Machine learning is making great strides in the rapid identification of cancer every day. Numerous scholarly articles and papers have been disseminated across several platforms using diverse approaches. Numerous computational techniques have been developed to find tumour suppressor genes in silico.

Khan et al. [10] used a provided original protein sequence to extract location-related characteristics for the detection of S-nitrosocysteine sites, the most prevalent posttranslational modification of proteins. Statistical moments were employed for position-relative feature extraction, and a multilayer neural network trained using techniques for adaptive learning and gradient descent was employed. Modular radial basis function and conventional radial basis function neural networks were introduced by Zainuddin et al. [11] in order to classify protein sequences into many categories. The n-gram method is used to translate the properties of proteins into numerical numbers. One kind of offered learning strategy is the self-organized selection of centers. In this case, a subtractive clustering-based training methodology is used to train the network.

The work by Malebary et al. [12] computes position-based characteristics and statistical moments that are integrated into pseudo amino-acid composition (PseAAC) using Chou's five-step criteria. Random forest classifier is then employed to accurately predict proto-oncogenes. To extract features from the protein sequence, Yang et al. [13] used the word segmentation technique. The SVM was then used to classify the attributes. Mahmood et al. [14] published a technique for finding hydroxylysine sites that is based on a strong statistical and mathematical approach that considers the shape of each element inside protein sequences as well as the influence of sequence order.

Wang et al. [15] created a novel approach that considers both domain sequence similarity and total sequence similarity in order to determine the evolutionary divergence between a given protein and a protein family. A 60-dimensional space was constructed using the natural vector technique, in which a vector uniquely represents each protein. They also combine all the natural vectors pertaining to a family of proteins to form a convex hull. The "Sorting Tolerant from Intolerant" (SIFT) approach was used by [16, 17] to determine if an amino acid substitution (AAS) impacts protein function. Lyu et al. [18] developed the method Discovery of Oncogenes and Tumour SupressoR genes using Genetic and Epigenetic features (DORGE) to identify TSGs and OGs by combining large-scale genetic and epigenetic data.

Moreover, deep learning techniques have been used to increase accuracy. A unique approach was presented by Tavanaei et al. [19] to predict proto-oncogenes (OGs) and tumour suppression genes (TSGs) based on the three-dimensional structures of the Protein Data Bank (PDB). Convolutional neural networks (CNNs) are created by them to categorize feature map sets that are taken from the protein structures. In order to classify the cancer genes, proto-oncogenes, tumour suppressor genes, and fusion genes, Anandanadarajah et al. [20] provided an effective preprocessing for the 3D convolutional deep learning stage and several fundamental structure classification approaches.

Alotaibi et al. [21] proposed deep learning methods to help identify stomach cancer growth at the best possible time, such as bi-LSTM, gated recurrent units, and long- and short-term memory. This study identified 61 carcinogenic driver genes, wherein mutations may cause stomach cancer. A deep learning model with minimal supervision was developed by Tomita et al. [22] to identify somatic mutations in LUAD patients. Extracted CNN-based features are merged and analyzed to predict the genetic mutation for a patient. They used CNN-based ResNet18 and ImageNet pre-trained CNN to study two categories of picture characteristics: LUAD sub type specific features and general image features.

Some of the constraints that have been faced by the most recent novel efforts in the field of proto-oncogene cancer mutations require consideration. The attention mechanism should be integrated with deep learning models in order to solve this. For future research projects to assess and improve accuracy, it is essential to create more thorough and reliable evaluation procedures based on this study.

Materials and Methods

Overview of BiGRU

An expansion of the GRU (Gated Recurrent Unit) neural network is the Bidirectional Gated Recurrent Unit (BiGRU). The forward and backward GRU units make up the BiGRU network used in this study. Here, $\overrightarrow{hi_t}$ represents the hidden layer of the forward GRU unit, while $\overleftarrow{hi_t}$ represents the hidden layer of the backward GRU unit. Formulas 1 and 2 display the unidirectional GRU's hidden layer outputs at time t. As indicated by formula 3, the hidden layer output of the forward GRU unit and the backward GRU unit are spliced through the hidden layer output of the BiGRU at time t. Capturing the sentence sequence's contextual properties is the aim of BiGRU. Figure 1 shows the architecture of BiGRU.

$$\overrightarrow{hi_t} = \mathrm{GRU}(x_t, \overrightarrow{hi_{t-1}}) \quad (1)$$

$$\overleftarrow{hi_t} = \mathrm{GRU}(x_t, \overleftarrow{hi_{t-1}}) \quad (2)$$

$$hi_t = \left[\overrightarrow{hi_t}, \overleftarrow{hi_t}\right] \quad (3)$$

Attention

In deep learning, attention is a method that allows neural networks to concentrate on particular portions of the input data while processing information or forming predictions. It draws inspiration from how humans digest information and assign varying degrees of attention to different parts of our environment. Rather than considering every input element identically, attention gives each element a variable relevance score, which enables the model to compute the weights of the components differently. By ignoring unimportant features and focusing on pertinent information, the model is better able to produce insightful outputs or make accurate forecasts.

Proposed Methodology

This study proposed two approaches for proto-oncogene prediction from the given sequence. This two approaches are used to identify whether the given protein sequence is normal or it will be changed into oncogene. In the first method, an attention with Convolutional Neural Network (ACNN) is employed to identify proto-oncogene. In the second method, an Attention with Bi directional Gated Recurrent Units (ABiGRU) is used for predicting the given proto-oncogene protein sequence.

ACNN based Proto-oncogene prediction

The architecture of the attention with CNN approach for predicting proto-oncogene is illustrated in Figure 2.

This ACNN architecture consists of two convolutional layers, two maxpool layers, one normalization layer, flattern layer, attention layer, dropout layer and finally a softmax layer for identification. The input protein sequences I_{seq} are fed into the convolution layer, which then applies a 64-bit filter with a 1-bit kernel size. The feature maps received from Convolutional layer is fed into maxpooling layer1 (MPL1) with size 1x2. The normalized output from the normalized layer is applied to the second convolutional layer with 128 filter size. The concatenation layer with 1/3 kernel size receives the features from the convolutional layer2 and the maxpool layer 1. Again the output of the concatenated layer is fed into the maxpool layer2 (MPL2) of size 1x2.

After that, the output will go to a layer for flattening and then to an attention layer. Attention layers results will be received by dropout layer. Finally the softmax layer classifies the protein sequence by oncogene or normal.

The feature extraction procedure using ACNN is described in the following equations (4) – (11).

$$ConL_1 = 2Dcon_{64\times1\times5}(I_{seq}) \quad (4)$$

$$MPL_1 = maxpool_{1\times2}(ConL_1) \quad (5)$$

$$CC = Concat(MPL_1 + ConL_2) \quad (6)$$

$$MPL_2 = maxpool_{1\times2}(CC) \quad (7)$$

$$FL = Flatt(MPL2) \quad (8)$$

$$AL = Attent(FL) \quad (9)$$

$$DL = Drop(AL) \quad (10)$$

$$Result = Softmax(DL) \quad (11)$$

ABiGRU based Proto-oncogene prediction

The architecture of the attention with BiGRU approach for predicting proto-oncogene is illustrated in Figure 3.

This ABiGRU architecture consists of two BiGRU with size 128 and 64, two maxpool layers, two ReLU layer, flattern layer, dense layer. attention layer, dropout layer and finally a softmax layer for classification. The input protein sequences I_{seq} is sent to the BiGRU with 128 size filter. The result of BiGRU is directed to the size 2 maxpooling layer MPL_1. The features obtained from maxpool layer1 is given to the ReLu layer. The attention layer received the results from the ReLU layer and the most relevant data is fed into the dropout layer. The output from the attention layer is again sent to the BiGRU of 64 size filter followed by maxpooling layer2 (MPL_2) with size 2. The result will then be applied to the ReLu layer and then dense layer of size 128. Finally the softmax layer classifies the protein sequence by oncogene or normal. The following Equation (12) – (21) illustrates the BiGRU with attention model.

$$BiGRU_1 = BiGRU_{128}(I_{seq}) \quad (12)$$
$$MPL_1 = MaxPool_{1\times2}(BiGRU_1) \quad (13)$$
$$ReLU_1 = relu(MPL_1) \quad (14)$$
$$AL = Attent(ReLU_1) \quad (15)$$
$$DL_1 = dropout(AL) \quad (16)$$
$$BiGRU_2 = BiGRU_{64}(DL_1) \quad (17)$$
$$MPL_2 = MaxPool_{1\times2}(BiGRU_1) \quad (18)$$
$$ReLU_2 = relu(MPL_2) \quad (19)$$
$$DEL = Densel_{128}(ReLU_2) \quad (20)$$
$$Result = softmax(DEL) \quad (21)$$

Results

Dataset Description

An extensive collection of protein data is available

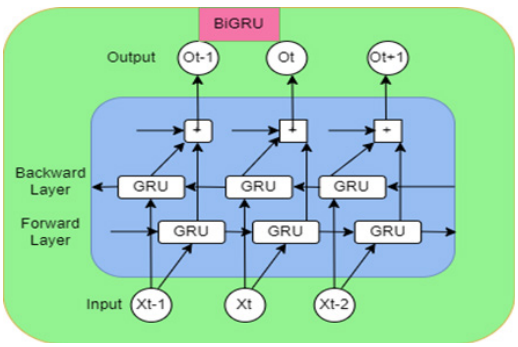

Figure 1. Architecture of BiGRU

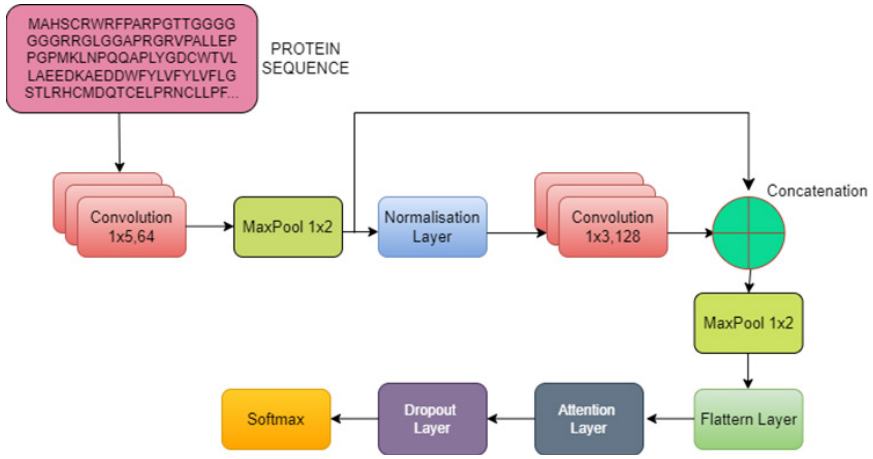

Figure 2. Architecture of Proto-Oncogene Prediction based on Attention CNN

in the well-known protein database Uniprot, the availability of which has been empirically verified. Moreover, a significant amount of its protein sequences are unclassified. Every protein has a distinct promotion number, and watchwords and other features are used to represent it based on its known capabilities. The protein sequences in the Uniprot database that contained the term "proto-oncogene" were selected for data collection. We used the pre-handled dataset acquired from [23] for the evaluation of our model. We used both independent and k-fold techniques to testing, even though the benchmark records are divided into two groups for various independent test combinations. On the other hand, the evidence is used in independent tests and tenfold subsampling test with prediction models.

The benchmark dataset is expressed as Y=Y+UY-(3) because Y+ and Y- are made up of 630 negative samples and 252 positive samples, respectively. 252+630=882 samples are included in the Supplementary Information S1 file for the convenience of the readers. The training and test datasets for statistical prediction make up the benchmark dataset. Once the preparation record has been used to prepare the proposed model, it is tested using that record. We used both independent and k-fold techniques to testing, even though the benchmark records are divided into two groups for various independent test combinations. On the other hand, the evidence is used in independent tests and tenfold subsampling tests with prediction models.

Performance Metrics

The effectiveness of the suggested proto-oncogene protein sequence detection method is assessed using the following statistical metrics: recall, accuracy, precision, F1 score and MCC. The terms True-Positive, True-Negative, False-Positive and False-Negative are represents TruPos,TruNeg,FalPos and FalNeg respectively. The Matthews Correlation Coefficient (MCC) is a performance metric that evaluates the overall efficacy of a binary classification model, accounting for negatives as well as true and false positives, and is utilised in proto-oncogene prediction. A model with a higher MCC value has a better capacity for prediction. The mathematical representation of the metrics is displayed in the following equations (22) – (26).

$$Accuracy = \frac{(TruPos+TruNeg)}{(TruPos+FalPos+TruNeg+FalNeg)} \quad (22)$$

$$Precision = \frac{TruPos}{(TruPos+FalPos)} \quad (23)$$

$$Recall = \frac{TruPos}{(TruPos+FalNeg)} \quad (24)$$

$$F1Score = 2 \times \frac{(Recall \times Precision)}{(Recall+Precision} \quad (25)$$

$$MCC = \frac{TruPos \times FalPos - TruPos \times FalNeg}{\sqrt{(TruPos+FalPos)(TruPos+FalNeg)(TruNeg+FalPos)(TruNeg+FalNeg)}} \quad (26)$$

Experimental Results

This work has utilised 70% of data for training purpose

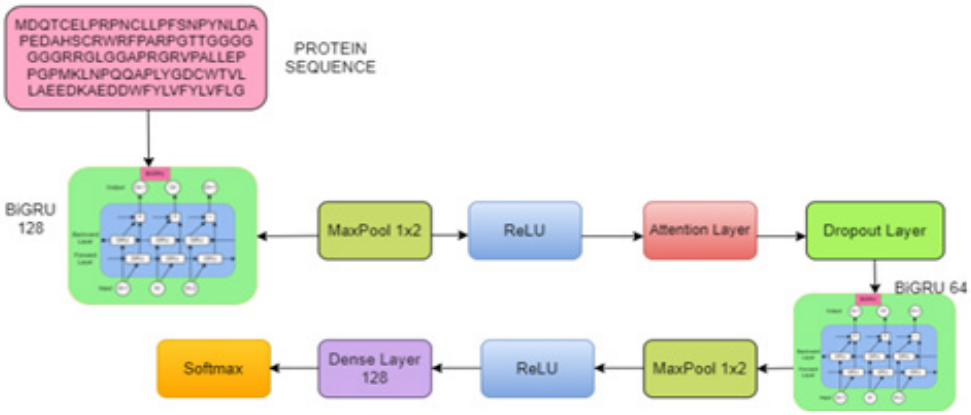

Figure 3. Architecture of Proto-Oncogene Prediction based on Attention BiGRU

Table 1. Results of the Independent test of ACNN and ABiGRU Method

Methods	Accuracy	Prcision	Recall	FScore	MCC
PSSM [24]	80.75	76.62	79.03	77.55	55.61
PseAAC [25]	84.52	80.84	83.25	81.83	64.04
ProtoPred_RF [12]	96.96	93.94	98.05	95.78	91.91
ACNN	96.85	95.06	96.38	95.72	91.34
ABiGRU	97.53	96.54	97.9	97.32	93.57

and the remaining 30% of data for testing purpose. The proposed deep learning with Attention model is validated using Independent test and K-fold cross-validation test.

Independent Test

The independent test results in terms of accuracy, precision, recall, Fscore and MCC are displayed in Table 1.

The proposed ACNN model achieves 96.85 % of accuracy 95.06% of precision 96.38% of recall 95.72% of Fscore and 91.34% of MCC. The ABiGRU model achieves 97.53% of accuracy 96.54% of precision 97.9% of recall 97.32% of Fscore and 93.57% of MCC. Moreover the MCC value of our ABiGRU model achieves 93.57%, which is better performance compared to the other approaches.

The following figures demonstrate the performance of the proposed technique and the existing approaches. Figure 4 shows the accuracy of five different methods. Compared to the other methods, our proposed ABiGRU method outperforms other approaches. ABiGRU enhances the accuracy of +0.65%, +0.57%, +13.01%, and +16.78% than the ACNN, Protopred_RF, PseAAC and PSSM approaches respectively. The proposed ACNN model gives almost same accuracy of Protopred-RF model.

Figure 5 shows the precision of five different methods. Compared to the other methods, our proposed ABiGRU method outperforms other approaches. ABiGRU enhances the precision of +1.48%, +2.6%, +15.7%, and +19.92% than the ACNN, Protopred_RF, PseAAC and PSSM approaches respectively.

Supplementary Figure 1 shows the recall of five different methods. Compared to the other methods, protopred_RF method outperforms other approaches. The ABiGRU model achieves 97.9% of recall, which is just 0.15% less value than the protopred_RF model. But ABiGRU enhances the recall of other methods except protopred_RF.

Supplementary Figure 2 shows the Fscore of five different methods. Compared to the other methods, our proposed ABiGRU method outperforms other approaches. ABiGRU enhances the precision of +1.6%, +1.54%, +15.49%, and +19.77% than the ACNN, Protopred_RF, PseAAC and PSSM approaches respectively.

Supplementary Figure 3 shows the MCC of five different methods. Compared to the other methods, our proposed ABiGRU method outperforms other approaches. ABiGRU enhances the precision of +2.23%, +1.66%, +29.53%, and +37.96% than the ACNN, Protopred_RF, PseAAC and PSSM approaches respectively.

Table 2. Results of K-Fold test of ACNN and ABiGRU Model

Fold #	Accuracy	Precision	Recall	F1-Score
1	96	93	94	95
2	98	96	97	97
3	96	95	96	95
4	96	95	96	95
5	96	94	95	97
6	97	97	98	97
7	97	95	97	96
8	98	97	98	98
9	99	98	99	98
10	100	100	100	100
Average	97	96	97	97

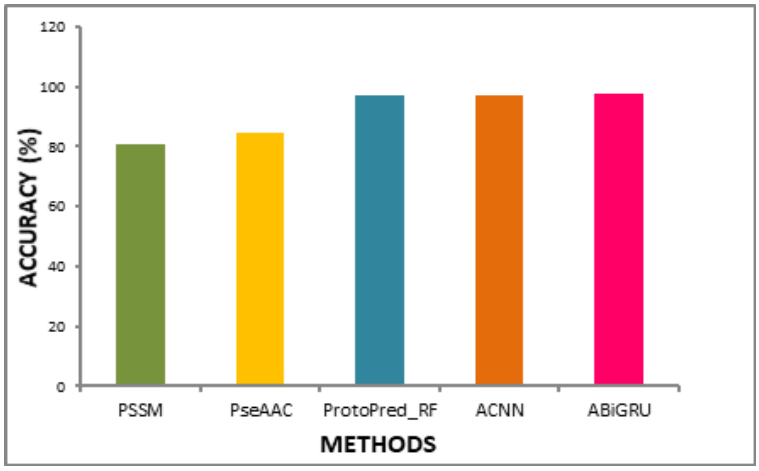

Figure 4. Comparison of Accuracy of ABiGRU and ACNN with the Existing Methods

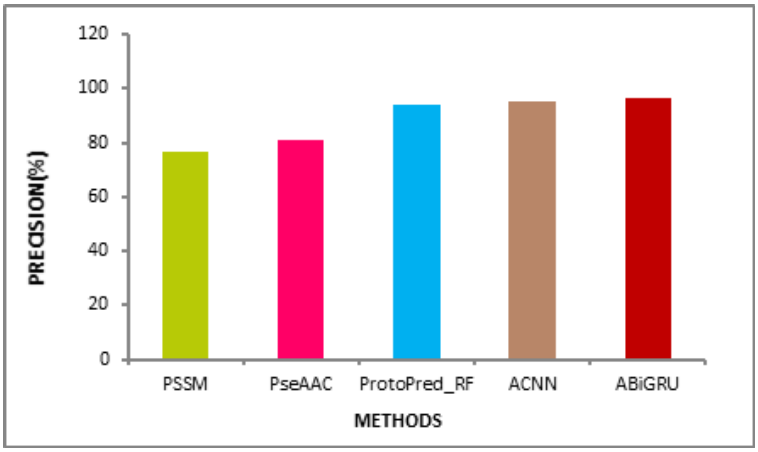

Figure 5. Comparison of Precision of ABiGRU and ACNN with the Existing Methods

Table 3. Study of Individual Components

Model Architecture	Accuracy	Precision	Recall	F1-Score	MCC
CNN only	92.34	90.15	91.87	91	82.76
BiGRU only	95.21	93.45	94.76	94.1	88.5
ACNN (CNN + Attention)	96.85	95.06	96.38	95.72	91.34
ABiGRU (Full model)	97.53	96.54	97.9	97.32	93.57

Table 4. Analysis of Statistical Significance Testing

Model Pair	Mean Accuracy Difference	p-value
ABiGRU vs ACNN	+0.68%	0.004
ABiGRU vs ProtoPred_RF	+0.57%	0.007

K-fold Cross Validation Test

The benchmark data set is split into k(10) disjoint fold partitions for cross-validation. For validation, each fold serves as a mutually exclusive data partition. The remaining data were used to train the model. Therefore, during testing and training, a significant portion of the whole data set is used. The result is the average of the results obtained from every fold. The method handles data samples, both positive and negative, in the same way. Using k=10, arbitrary parcels were generated. Due to the comprehensive evaluation of all the facts, cross-approval outperforms alternative confirmation techniques. Table 2 displays the findings of the K- fold cross validations of the proposed ABiGRU model.

Ablation Study

We performed ablation studies to assess the contribution of each component in the proposed ABiGRU model. The results of the ablation studies are presented in Table 3. The combined application of ABiGRU shows more effective results than using each component separately. This indicates that these elements work together and increase the predictive performance of the ABiGRU model in a synergistic manner. The ablation study confirms the importance of including these components in the overall framework and further bolsters the claim that the ABiGRU model is superior in Proto-Oncogene prediction tasks.

These results confirm that both the BiGRU units and the attention layer contribute significantly to the performance of the model. The attention layer enables the model to focus on relevant amino acid patterns, while BiGRU captures long-range dependencies in the sequence. Supplementary Figure 4 shows the accuracy of ablation study in the proposed model.

The objective of this ablation study is to conduct comparative analysis of each component in the proposed ABiGRU model. The results, as depicted in Supplementary Figure 4, demonstrate that the BiGRU method integrated within Attention (ABiGRU) exhibits superior performance compared all the other methods.

Statistical Significance Testing

To statistically validate the superiority of the ABiGRU model, we performed a two-tailed paired t-test on the accuracy scores obtained over 10 different runs of k-fold cross-validation. We compared ABiGRU with ACNN and Protopred_RF. Table 4 illustrates the performance of the proposed work by statistical significance testing with p-value

The improvements observed are statistically significant ($p < 0.01$), confirming that the ABiGRU model outperforms the existing techniques with high confidence. All p-values are less than 0.01, indicating statistically significant improvements by ABiGRU over existing methods.

Model Interpretability via Attention Weights

To enhance interpretability analyzed the attention scores assigned to individual amino acids in sample sequences from the UniProt database. Attention weights allow us to identify which regions of a sequence the model considers most relevant for classifying a protein

as a proto-oncogene.

We selected representative sequences that were correctly classified with high confidence. For these sequences, attention heatmaps were overlaid with domain annotations from UniProt. This correspondence between high-attention regions and biologically annotated domains confirms that the model is not only accurate but also biologically grounded in its decision-making.

Supplementary Figure 5 illustrates a heatmap where the high-attention regions clearly align with a kinase domain in a known proto-oncogene sequence.

Discussion

One of the main ways that exposure to a mutagen promotes cancer is through mutations in proto-oncogenes. Translated proto-oncogenes become proto-oncogene proteins. These proteins function as a biomarker for this kind of cancer susceptibility. The suggested study offers a reliable method for locating these proteins. The attention mechanism in deep learning has emerged as a innovative method for protein sequence classification. This study proposed two approaches like Attention with Convolutional Neural Network (ACNN) and Attention with Bi directional Gated Recurrent Units (ABiGRU) to predict Proto-oncogene protein sequence. The performance was evaluated on the benchmark Uniprot dataset. The proposed deep learning with Attention model is validated using Independent test and K-fold cross-validation test. Also we performed ablation studies to assess the contribution of each component in the proposed ABiGRU model. The results demonstrate that the BiGRU method integrated within Attention (ABiGRU) exhibits superior performance compared all the other methods. Moreover statistically validate the superiority of the ABiGRU model, we performed a two-tailed paired t-test on the accuracy scores obtained over 10 different runs of k-fold cross-validation. The ABiGRU achieves 97.53% of accuracy and the ACNN model achieves 96.85% of accuracy. The results were analysed with three existing techniques, our ABiGRU model outperforms all the other methods.

Author Contribution Statement

R.D. Seeja contributed Methodology, conceptualization and investigation. A Bernice Rufus has focused on methodology and drafting the original manuscript..

Acknowledgements

The authors sincerely thank all contributors whose invaluable insights and efforts were essential to the successful completion of this manuscript.

Conflict of interest

The authors declare that they have no conflict of interest.

References

1. Arshad A, Khan Y, DNA Computing A Survey, In Proceedings of the 2019 International Conference on Innovative Computing (ICIC), 2019; pp.1–5.
2. Williams DE, Eisenman J, Baird A, Rauch C, Van Ness K, March CJ, et al. Identification of a ligand for the c-kit proto-oncogene. Cell. 1990;63(1):167-74. https://doi.org/10.1016/0092-8674(90)90297-r.
3. Cooper G. M, Oncogenes, II-nd edition, Jones and Bartlett Publishers Inc. Boston, 1995; 384.
4. Mulligan LM, Kwok JB, Healey CS, Elsdon MJ, Eng C, Gardner E, et al. Germ-line mutations of the ret proto-oncogene in multiple endocrine neoplasia type 2a. Nature. 1993;363(6428):458-60. https://doi.org/10.1038/363458a0.
5. Croce CM. Oncogenes and cancer. N Engl J Med. 2008;358(5):502-11. https://doi.org/10.1056/NEJMra072367.
6. Vogelstein B, Papadopoulos N, Velculescu VE, Zhou S, Diaz LA, Jr., Kinzler KW. Cancer genome landscapes. Science. 2013;339(6127):1546-58. https://doi.org/10.1126/science.1235122.
7. Pon JR, Marra MA. Driver and passenger mutations in cancer. Annu Rev Pathol. 2015;10:25-50. https://doi.org/10.1146/annurev-pathol-012414-040312.
8. Cong H, Zhang M, Zhang Q, Gong J, Cong H, Xin Q, et al. Analysis of structures and epitopes of surface antigen glycoproteins expressed in bradyzoites of toxoplasma gondii. Biomed Res Int. 2013;2013:165342. https://doi.org/10.1155/2013/165342.
9. Radivojac P, Clark WT, Oron TR, Schnoes AM, Wittkop T, Sokolov A, et al. A large-scale evaluation of computational protein function prediction. Nat Methods. 2013;10(3):221-7. https://doi.org/10.1038/nmeth.2340.
10. Khan YD, Batool A, Rasool N, Khan SA, Chou K-C. Prediction of nitrosocysteine sites using position and composition variant features. Lett Org Chem. 2019;16(4):283-93. https://doi.org/10.2174/1570178615666180802122953.
11. Zainuddin Z, Kumar M, Radial Basis Function Neural Networks in Protein Sequence Classification. Malaysian Journal of Mathematical Sciences. 2008; 2(2): 195-204..
12. Malebary S, Khan R, Khan Y. Protopred: Advancing oncological research through identification of proto-oncogene proteins. IEEE Access. 2021;PP:1-. https://doi.org/10.1109/ACCESS.2021.3076448.
13. Yang Y, Lu BL, Yang WY. Classification of protein sequences based on word segmentation methods. InProceedings of the 6th Asia-Pacific Bioinformatics Conference 2008, pp. 177-186.
14. Mahmood MK, Ehsan A, Khan YD, Chou KC. Ihyd-lyssite (epsv): Identifying hydroxylysine sites in protein using statistical formulation by extracting enhanced position and sequence variant feature technique. Curr Genomics. 2020;21(7):536-45. https://doi.org/10.2174/1389202921999200831142629.
15. Wang Y, Tian K, Yau SS. Protein sequence classification using natural vector and convex hull method. J Comput Biol. 2019;26(4):315-21. https://doi.org/10.1089/cmb.2018.0216.
16. Vaser R, Adusumalli S, Leng SN, Sikic M, Ng PC. Sift missense predictions for genomes. Nat Protoc. 2016;11(1):1-9. https://doi.org/10.1038/nprot.2015.123.
17. Kumar P, Henikoff S, Ng P. Predicting the effects of coding non—synonymous variants on protein function using the sift algorithm. Nat Protoc. 2009;4:1073-81. https://doi.org/10.1038/nprot.2009.86.
18. Lyu J, Li JJ, Su J, Peng F, Chen YE, Ge X, et al. Dorge: Discovery of oncogenes and tumor suppressor genes using

genetic and epigenetic features. Sci Adv. 2020;6(46). https://doi.org/10.1126/sciadv.aba6784.
19. Tavanaei A, Anandanadarajah N, Maida A, Loganantharaj R. A deep learning model for predicting tumor suppressor genes and oncogenes from pdb structure. bioRxiv. 2017:177378. https://doi.org/10.1101/177378.
20. Anandanadarajah N, Chu CH, Loganantharaj R. An integrated deep learning and dynamic programming method for predicting tumor suppressor genes, oncogenes, and fusion from pdb structures. Comput Biol Med. 2021;133:104323. https://doi.org/10.1016/j.compbiomed.2021.104323.
21. Alotaibi FM, Khan YD. A framework for prediction of oncogenomic progression aiding personalized treatment of gastric cancer. Diagnostics (Basel). 2023;13(13). https://doi.org/10.3390/diagnostics13132291.
22. Tomita N, Tafe LJ, Suriawinata AA, Tsongalis GJ, Nasir-Moin M, Dragnev K, et al. Predicting oncogene mutations of lung cancer using deep learning and histopathologic features on whole-slide images. Transl Oncol. 2022;24:101494. https://doi.org/10.1016/j.tranon.2022.101494.
23. Wang D, Lee NK, Dillon TS. Extraction and optimization of fuzzy protein sequences classification rules using GRBF neural networks. Neural Information Processing-Letters and Reviews. 2003;1(1):53-7.
24. Delorenzi M, Speed T. An hmm model for coiled-coil domains and a comparison with pssm-based predictions. Bioinformatics. 2002;18(4):617-25. https://doi.org/10.1093/bioinformatics/18.4.617.
25. Jia J, Liu Z, Xiao X, Liu B, Chou KC. Isuc-pseopt: Identifying lysine succinylation sites in proteins by incorporating sequence-coupling effects into pseudo components and optimizing imbalanced training dataset. Anal Biochem. 2016;497:48-56. https://doi.org/10.1016/j.ab.2015.12.009.