IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 9 April 2023, accepted 1 June 2023, date of publication 9 June 2023, date of current version 6 July 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3285197

RESEARCH ARTICLE

iPro-TCN: Prediction of DNA Promoters Recognition and Their Strength Using Temporal Convolutional Network

ALI RAZA[1], WALEED ALAM[2], SHAHZAD KHAN[1,3], MUHAMMAD TAHIR[3,4], AND KIL TO CHONG[2,5]

[1]Physical and Numerical Sciences, Qurtuba University of Science and Information Technology, Peshawar, Khyber Pakhtunkhwa 25000, Pakistan
[2]Department of Electronics and Information Engineering, Jeonbuk National University, Jeonju 54896, South Korea
[3]Department of Computer Science, Abdul Wali Khan University, Mardan, Khyber Pakhtunkhwa 23200, Pakistan
[4]Department of Electrical and Computer Engineering, University of Manitoba, Winnipeg, MB R3T5V6, Canada
[5]Advanced Electronics and Information Research Center, Jeonbuk National University, Jeonju 54896, South Korea

Corresponding authors: Muhammad Tahir (M.tahir@umanitoba.ca) and Kil To Chong (kitchong@jbnu.ac.kr)

This work was supported in part by the National Research Foundation of Korea (NRF) funded by the Korean Government [Ministry of Industry Science and Technology (MIST)] under Grant 2020R1A2C2005612.

ABSTRACT Promoters are an important regulatory element in the genome that control gene expression, and their abnormalities have been linked to various diseases. Therefore, accurately promoter identification is essential for biological research as we as drug development. But the identification of the promoter using laboratory approaches is highly costly. In order to address this issue, we proposed a computational model called iPro-TCN to predict promoter and their strength using temporal convolutional network (TCN) with a word2vec feature representation. This model includes a feature descriptor known as Word2Vec and achieved high performance to predict promoters, including strong and weak promoters. The iPro-TCN model obtained accuracy of 91.86% to predict promoter in the first layer for, and an accuracy of 84.63% to predict strong and week promoter in the second layer using cross validation test. On benchmark datasets, the proposed iPro-TCN model produced better performance than previous computational models in term of all performance metrics.

INDEX TERMS Deep learning, promoters, temporal convolution network, natural language processing, DNA, word2vec.

I. INTRODUCTION

Promoter is a region of DNA where RNA polymerase begins to transcribe a gene. Gene expression regulation in prokaryotes is generally simpler than in eukaryotes because prokaryotes have a smaller and more compact genome, and their transcription and translation machinery is less complex [1]. The promoter sequence contains specific DNA motifs that bind transcription factors and determine the level of gene expression. Promoters are usually located immediately upstream of the gene, and they typically contain a ribosome binding site, a start codon, and a promoter sequence. The activity of promoters can be regulated by a variety of factors, including DNA methylation, histone modification, and the binding of transcription factors [2], [3]. Dysregulation of gene expression due to changes in promoter activity can have significant consequences, and understanding the mechanisms that regulate promoter activity is an active area of research in genetics and molecular biology.

Promoters are important in many biological processes, including development, aging, and disease. Understanding the role of promoters in gene expression can be important for the diagnosis and treatment of diseases [4], and there are many techniques that are used to study promoters and their role in gene regulation, including DNA sequencing, DNA microarrays, and RNA sequencing. Promoters are DNA sequences that serve as binding sites for RNA polymerase, the enzyme responsible for transcribing DNA into RNA.

The associate editor coordinating the review of this manuscript and approving it for publication was Xinyu Du.

In bacteria, the promoter region typically consists of two hexameric sequences known as the "-35" and "-10" elements, which are centered around 35 and 10 base pairs upstream of the transcription initiation site, respectively [5], [6]. The -35 element consists of the sequence TTGACA, while the -10 element consists of the sequence TATAAT [7]. These elements serve as binding sites for specific transcription factors that help to regulate transcription. Sigma factors ($\sigma - factors$) are proteins that are part of the RNA polymerase enzyme complex in bacteria [8]. They play a critical role in the initiation of transcription, which is the process of synthesizing RNA from a DNA template. When a bacterium needs to transcribe a particular gene, the sigma factor helps to recognize and bind to the promoter region of the DNA, which is a specific sequence located upstream of the gene. This binding helps to position the RNA polymerase enzyme complex at the transcription initiation site, so that it can begin synthesizing RNA. There are several different sigma factors in bacteria, each of which is specialized to recognize and bind to specific promoter sequences. Different sigma factors are responsible for transcribing different sets of genes, depending on the promoter sequences they recognize. For example, the $\sigma 70$ sigma factor is responsible for transcribing the majority of genes in Escherichia coli, while other sigma factors may be specialized for transcribing genes under specific environmental conditions or in response to particular signaling pathways [9].

The identification of promoters is an important area in genome research, as understanding the regulatory sequences that control gene expression can provide insight into the function of different genes and the regulation of biological processes. DNA sequencing and bioinformatics analysis are often used to identify and characterize promoters in bacterial genomes. The PCSM model (position-correlation scoring matrix) is a computational approach that was developed by Li et al., for predicting $\sigma 70$ promoters in the bacterial species Escherichia coli K-12 [9]. The PCSM model uses a scoring matrix approach to analyze DNA sequences and identify promoter regions that are likely to be recognized by the $\sigma 70$ sigma factor. Similarly, vwZ-curve model was developed by Song et al., for analyzing prokaryotic promoters and predicting their transcriptional activity [10] used a variable-window Z-curve method to extract general features of prokaryotic promoters. The vwZ-curve model analyzes DNA sequences by breaking them down into smaller window sizes and applying a Z-curve analysis to each window. This allows the model to identify patterns and features that are characteristic of promoters, such as specific DNA sequence motifs or structural features. likely, Silva et al., proposed a computational approach called Stabilty for predicting promoter regions in bacterial genomes [11]. The Stability model based on NN algorithm that integrates DNA duplex stability into its predictions. The iPro54-PseKNC model is a computational approach established by Lin et al. to identify sigma-54 promoters in prokaryotes [12]. The iPro54-PseKNC model is based on pseudo k-tuple nucleotide composition (PseKNC). In the iPro54-PseKNC model, PseKNC extract features from DNA sequences that are relevant for identifying sigma-54 promoters. Then, incremental feature selection procedure was employed to select the relevant features. Finally, used SVM for classification. Similarly, the iPromoter-2L method [8] was established by Liu et al., to predict promoters and their types using random forest for classification. The iPromoter-2L model based on multi-window-based PseKNC method. In iPromoter-2L model, the PseKNC was employed to extract hidden feature from promoters sequences and multi-window approach was used to divided the promoters sequences into smaller windows and the PseKNC features are extracted from each window. This allows the model to identify patterns and features that are characteristic of promoters, such as specific DNA sequence motifs or structural elements. Likewise, Patiyal et al., established a computational model called: Sigma70Pred for the identification of sigma70 promoters. This model used various feature extraction such as timer count, dimer count, nucleotide composition and so on to extract hidden feature from promoter sequences and various classifier namely KNN, SVM and so for classification [13]. Similarly, Shujaat et al., developed a model namely: iPromphage for prediction of phage promtoers. This used one-hot encoding schemes, and convolution layers to extraction hidden feature from promoter sequences [14]. In the same way, iPSW(2L)-PseKNC approach was established by Xiao et al. for promoters and their strength prediction [15]. The model based on hybrid features, which are a combination of structural and sequence-based features, and PseKNC to extract hidden featrue from promoters sequences and for classification used SVM classifier.

The existing computational methods such as PCSM model [9], vwZ-curve model [10], Stability model [11], iPro54-PseKNC model [12], and iPromoter-2L model [8] were mostly based on machine learning (ML) approaches. These methods rely on extracting various features from DNA sequences, and using these features as input to train a model. These models based on machine learning approaches obtained good performance and solutions for the promoter sequences data, but they were strongly dependent on handcrafted engineering feature map and required field knowledge to extract hidden features and pattern in promoter sequences data. In this regard, we propose a model called iPro-TCN based on deep learning approaches for prediction of promoters and their strength. The iPro-TCN model include two phases, in first phase we employed the most popular natural language processing (NLP) methods, it split the promoter sequence into words i.e., 3-mer, 4-mer, 5-mer, and 6-mer and each word is then mapped to its corresponding feature. In second phase, we employed the deep learning algorithm temporal convolutional network (TCN) to predict promoter and their strengths. Overall, our model is a powerful tool for promoter prediction with high accuracy and performance.

II. MATERIALS AND METHODS

A rigorous experimental process in order to obtain reliable and valid results. This process typically involves carefully planning and designing the study, selecting appropriate materials and methods, conducting the experiments, analyzing the data, and interpreting and reporting the results. A flow chart can be used to illustrate the overall process of the proposed model, including the selection of materials and methods is presented in Figure 2.

A. BENCHMARK DATASETS

The benchmark datasets that have been widely used to evaluate the performance of deep learning models for promoter identification. These datasets typically consist of DNA sequences labeled as promoters or non-promoters, and they are used to train and evaluate deep learning models such as neural networks. In this paper, we selected benchmark datasets from Xiao et al. [15], the Xiao et al. collected these datasets of experimentally confirmed promoter sequences from a database called RegulonDB [16]. These datasets were used to build a classifier for predicting promoters in DNA sequences. The benchmark dataset contained 3382 promoter sequences and 3382 non-promoter sequences, and further split the promoter sequences into two types, 1591 strong promoters and 1791 weak promoters based on different kinds of transcriptional activation and expression [15], [17]. The promoter strength often relies on both the condition of the cell and the DNA sample. Particularly, the data was processed using CD-HIT to remove any sequences with a similarity above 80% [18], [19]. This is often done to ensure that the dataset contains a diverse set of sequences and to avoid biasing the results by including highly similar sequences.

B. DNA REPRESENTATION WITH MODEL

DNA is a molecule that carries the genetic information of an organism. It is made up of a chain of nucleotides, which are the basic units of genetic information. The nucleotides are arranged in a specific sequence that determines the genetic information carried by the DNA molecule [15], [19]. A language model is a type of ML method that can predict the likelihood of a sequence of words or characters in a language [16]. Language models are often used in NLP tasks including text generation, speech recognition, and,language translation,. It is possible to represent DNA sequences as a language model by treating each nucleotide as a ''word'' in the language. For example, a DNA sequence could be represented as a sequence of characters ''ACGTGCA…'', with the language model predicting the likelihood of each nucleotide based on the context of the previous nucleotides in the sequence. To train a language model on DNA sequences, you would need a dataset of DNA sequences and their corresponding labels (if the task involves predicting labels).

C. TEMPORAL CONVOLUTIONAL NETWORK

TCN is a prominent deep learning architectures with casual convolution layers and dilations used for one dimensional data and specifically designed to process sequential data. They are called ''causal'' because the output of the model at any time step is only dependent on the past inputs, and not on future inputs [20], [21]. This makes them suitable for tasks such as predicting the next value in a time series or generating a translation of a sentence, where the output at each time step depends only on the input up to that point. TCNs use dilated causal convolutions, which allow the model to have a large receptive field while using less memory or more parameters required to store the model. They also use residual blocks, which enable the model to learn changes to the identity mapping rather than the entire transformation, which helps the model learn more efficiently. A TCN model consists of a series of convolutional layers, each of which operates on a different portion of the input sequence. The output of each layer is fed into the next layer, allowing the model to build up a representation of the entire input sequence over time.

D. THE PROPOSED PREDICTOR MODEL

1) OVERVIEW OF iPro-TCN MODEL

Here, we proposed a deep learning-based approch namely: iPro-TCN to predict promoters and their types. The iPro-TCN takes a DNA sequence as input and converts into a feature matrix using the word2vec method. This involves dividing the sequence into overlapping k-mers (short subsequences of fixed length) and converting each k-mer into a vector representation using the word2vec algorithm. The resulting feature matrix is then input into a temporal convolutional network (TCN) consisting of convolutional, causal convolutions, Padding and activation function, which are designed to extract meaningful features from the data. The output of the TCN is then transfer to fully connected layer and sigmoid layer, which is used to classify the input as either promoters or non-promoters. Figure 2 display the architecture of the proposed iPro-TCN method.

2) FEATURE EXTRACTION TERMINOLOGY

The genetic data are represented as sequences, it is recognized as a language by the neurons and cells that transmits information. Furthermore, the Natural Language Processing (NLP) methods namely: word2vec are used. The word2vec is a method for converting words (or in this case, k-mers) into numerical vector representations that capture the meaning of the words. It can be used to generate these representations using skip-gram model or continuous bag-of-words (CBOW) methods [15]. Based on the context of the adjacent words, the CBOW method predicts the target word. This word is represented by vector). On the other hand, the skip-gram model makes predictions regarding the surrounding words based on the target word. The skip-gram model is generally better for infrequent words, as it is able to generate high-quality vector representations for words that appear less frequently in the training data. The genome is divide according to their 23 chromosomes (Chr1 to Chr23). Additionally it is divided

TABLE 1. List of training parameters for Word2Vec.

List of Parameters	Word2Vec Model
Context words	k-mer(k=3,4,5,6)
Vector Size	100
Training Method	CBOW
Epochs	20
Window-Size and Minimum-Count	5 and 5

into 100nt length of sentences. Further, employed 3-mer, 4-mer and etc on each sentence [22]. In our model iPro-TCN, the skip-gram model is being used to generate vector representations for k-mers, which are then used as a preliminary feature matrix for the model. This allows the model to capture the meaning of the k-mers in the DNA sequence and use this information to classify the sequence as a Promoters or non-promoters.

The probability of observing the context words given a target word. This is done by learning a set of word vectors such that the dot product between the vectors for the target word and the context words is large when the words co-occur frequently in the training data, and small otherwise.The Figure 1 display the architecture of a CBOW model. Mathematically, the objective of the skip-gram model can be expressed as follows:

$$\text{Maximize} \sum(\mathrm{i}=1)^{\mathrm{n}} \sum(\mathrm{j} \in \text{context(i)})^{\mathrm{m}} \log \mathrm{p}\left(\mathrm{x_i} \mid \mathrm{x_i}\right)$$

where n is the number of words in the training data, m is the size of the context for each word (i.e., the number of words considered as context for each target word), context(i) is the set of context words for the target word xi, and p(xj | xi) is the probability of observing context word xj given target word xi. By maximizing this objective, the skip-gram model learns a set of word vectors that capture the relationships between words in the training data and can be used for various NLP tasks.

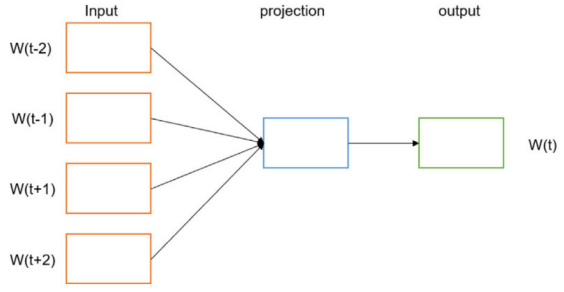

FIGURE 1. Shows the distribution of Promoters and their Strength.

3) TCN BUILDING BLOCK

In this paper, the proposed model: iPro-TCN consist of word2vec algorithm and a temporal convolutional network to predict promoters and their types. The word2vec converting words (k-mers) into numerical vector representations that capture the meaning of the words. It is commonly used in natural language processing and has been shown to be effective for a variety of tasks. The TCN is similar to a standard CNN, but is designed to handle input data with a temporal dimension (e.g., a sequence of words or time points). The TCN applies convolutional filters to the input data in a way that preserves the temporal relationships between the input elements. The iPro-TCN model is able to extract meaningful features from the DNA sequence and used them to predict the sequences as promoters and non-promoters along with strong and weak promoters. The TCN is particularly well-suited for this task, as it is able to capture the temporal relationships between the k-mers in the DNA sequence and use this information to make more accurate predictions.

III. PERFORMANCE EVALUATION

The following four metrics have been widely used in literature to determine the rate of success of this type of prediction model. They are accuracy (Acc), sensitivity (Sn), Mathew's correlation coefficient (MCC) and specificity (Sp) as given below [23], [24], [25], [26], [27], [28], [29], [30], [31], [32], [33], [34], [35], [36], [37], [38], [39]:

$$ACC = \frac{TN + TP}{TP + FN + TN + FP} \tag{1}$$

$$SN = \frac{TP}{TP + FN} \tag{2}$$

$$SP = \frac{TN}{FP + TN} \tag{3}$$

$$MCC = \frac{TP \times TN\text{-}FP \times FN}{\sqrt{(TN + FP)(TN + FN)(TP + FP)(TP + FN)}} \tag{4}$$

where TP, TN, FP, FN represents true positive, true negative, false positive, false negative, respectively). The receiver operating characteristic (ROC) curve was created by plotting the TF rate (1-SN) against the FP rate (1-SP) [40], [41], [42], [43], [44], [45]. On the ROC Figure 5, the AUC (area under the ROC curve) was also calculated and provided as an effective performance parameter. In this study, we employed the cross-validation approach and split the data into training, validation and testing. The Precision-Recall (PR) curves as well as the ROC curve, were used to provide an accessible method of measuring the model's prediction performance. The ROC curve compares the TP rate (TPR; 1-specificity) with the FP rate (FPR; 1-specificity) at different thresholds, whereas the precision-recall curve calculates the precision (the proportion of real positives out of all predicted positives) versus recall (sensitivity) at different thresholds, respectively. Moreover, the AUC serves as an objective measure of the quality of the prediction model. Which is a function of the number of observations. The AUC is in the range of 0.5-1, which is acceptable. The AUC indicates how accurate a predictor is [46] and [47]. The higher the AUC, the better the predictor. Finally, the confusion matrix, which serves as a visual representation of performance, is displayed.

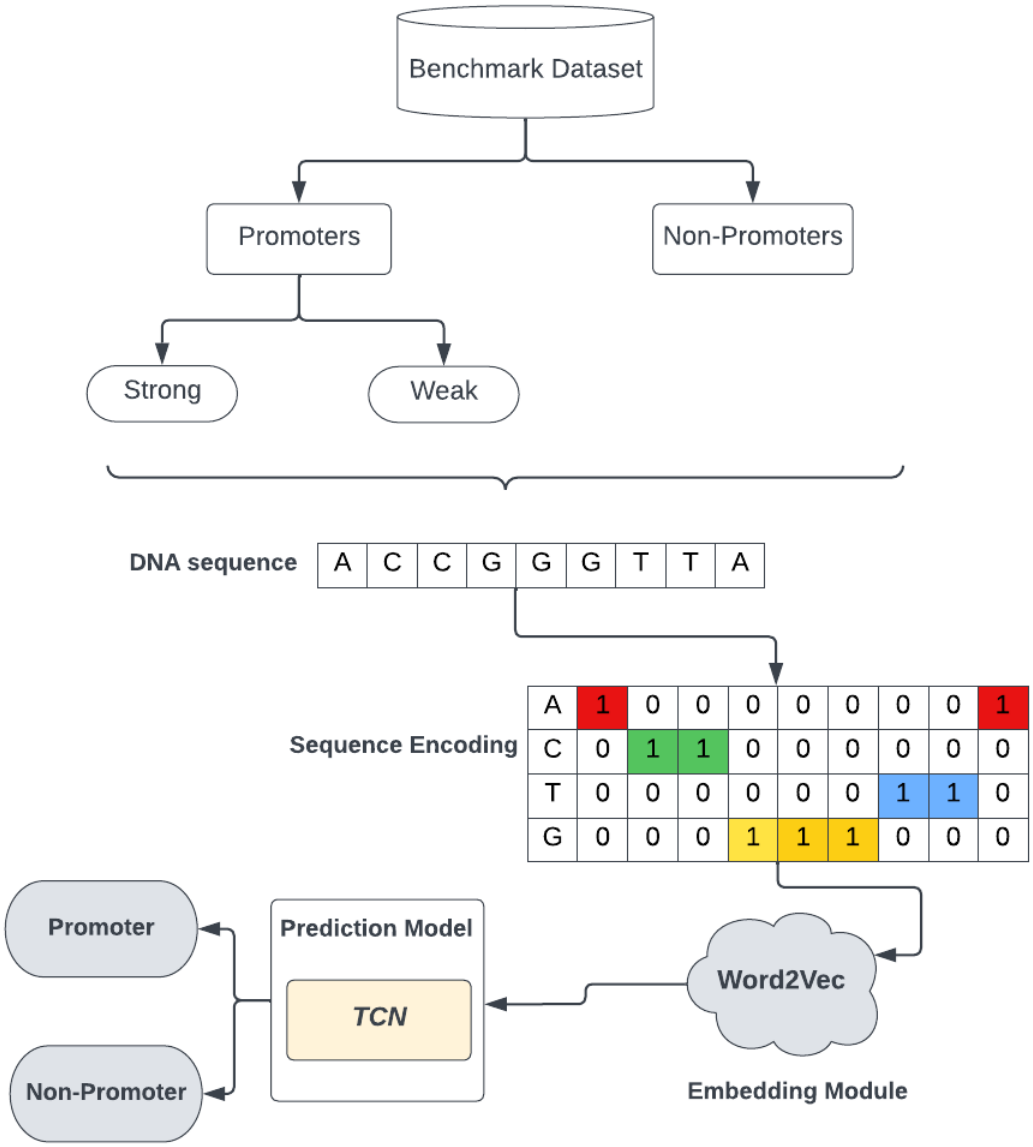

FIGURE 2. Illustration of the proposed model architecture.

IV. RESULTS AND DISCUSSION

A. COMPARISON OF DIFFERENT ENCODING SCHEMES

Here, we present the performance of our iPro-TCN model on 3-mer, 4-mer and so on. In first layer our model obtained accuracy of 89.65%, sensitivity of 86.81%, specificity of 83.47%, and MMC 0.79 on 3-mer encoding scheme. Similarly, on 4-mer the model produced 90.76% of accuracy, 90.26% of sensitivity, 91.21% of specificity and 0.81 of MCC. Likewise, on 5-mer the model produced 90.09% of accuracy, 90.81% of sensitivity, 89.38% of specificity and 0.80 of MCC. Finally, on 6-mer the model produced 91.86% of accuracy, 92.74% of sensitivity, 91.00% of specificity and 0.83 of MCC. The 6-mer produced better outcome the other 3, 4 and 5-mers shown in Table 2. Similarly, the prediction performance for the second layer of our proposed model is presented in the Table 3.

TABLE 2. Different encoding schemes comparison for promoter identification in layer 1 using benchmark dataset.

Encoding Scheme	ACC	SN	SP	MCC
3-mer	89.65%	86.81%	92.47%	0.79
4-mer	90.76%	90.26%	91.21%	0.81
5-mer	90.09%	90.81%	89.38%	0.80
6-mer	91.86%	92.74%	91%	0.83

B. PERFORMANCE COMPARISON OF iPro-TCN MODEL WITH EXISTING MODELS

The performance comparison of iProTCN model and existing model were presented in Table 4. The results of our proposed deep learning method iProTCN obtianed better performance then existing methods, namely iPSW(2L)-PseKNC [15], dPromoter-XGBoost [48], BERT-Promoter [49]. The best results of our proposed model on 1st layer are accuracy

TABLE 3. Different encoding schemes comparison for promoter strength in layer 2 using benchmark dataset.

Encoding Scheme	ACC	SN	SP	MCC
3-mer	84.34%	84.78%	83.94%	0.68
4-mer	84.63%	81.98%	87.04%	0.69
5-mer	83.75%	83.85%	83.66%	0.67
6-mer	84.49%	82.29%	86.47%	0.68

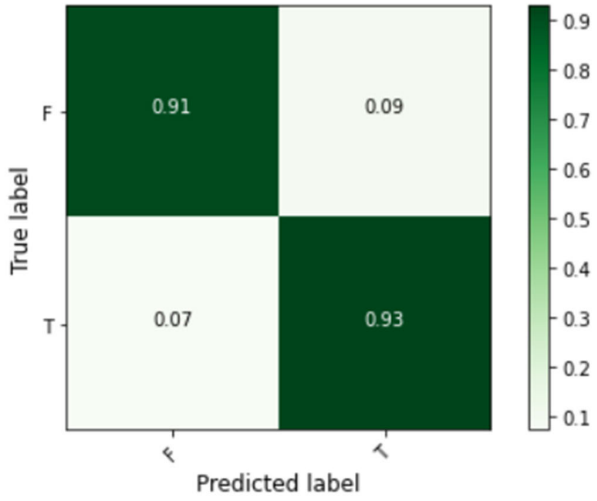

FIGURE 3. The proposed model's confusion matrix for 1st Layer.

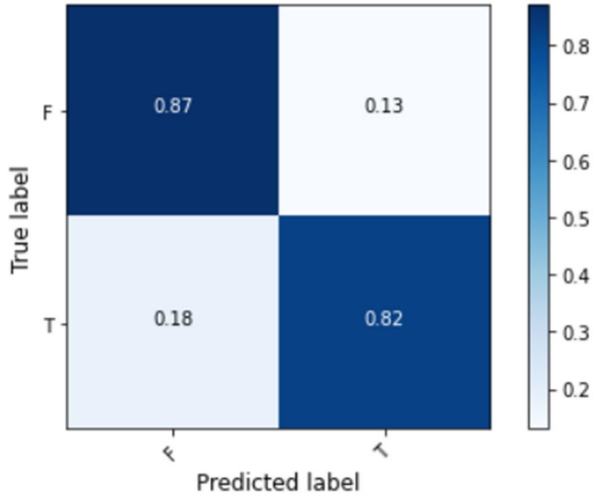

FIGURE 4. The proposed model's confusion matrix for 2nd Layer.

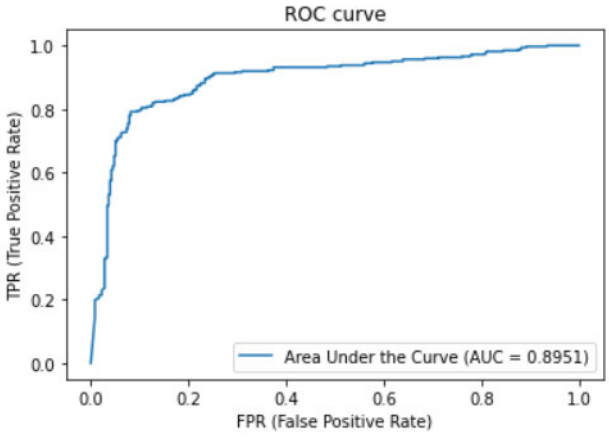

FIGURE 5. The ROC curves of the iPro-TCN model at 2nd Layer.

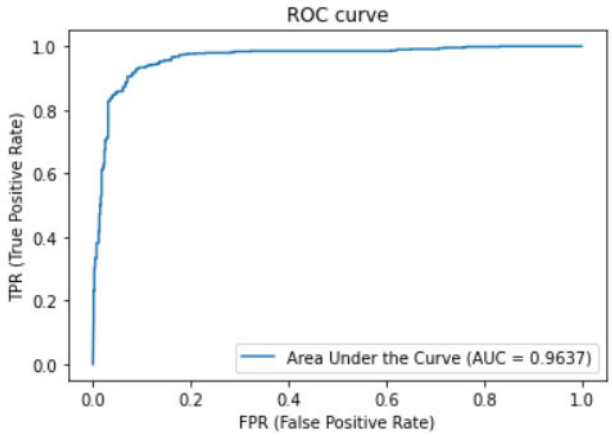

FIGURE 6. The ROC curves of the iPro-TCN model at 1st Layer.

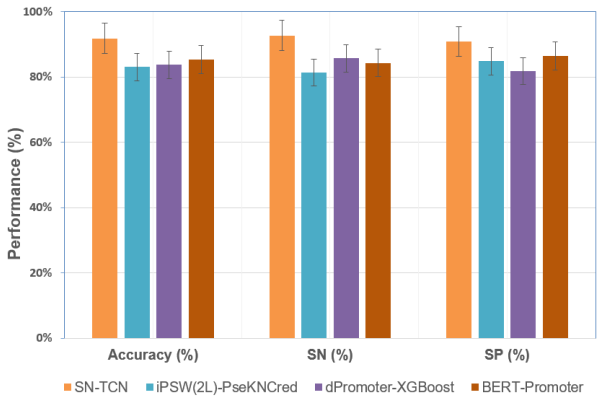

FIGURE 7. The iPro-TCN model comparison with existing method on first layer.

91.86%, sensitivity 92.74%, specificity 91.0%, and Matthews correlation coefficient (MCC) 0.83 values. Similarly, the 2nd layer of the model also obtained high accuracy 84.63%, specificity 87.04%, and sensitivity 81.98% values, as well as a high MCC 0.69 value. The results in Figure 7, Figure 8 and Table 4 show that the proposed model performs better than the iPSW(2L)-PseKNC [15], dPromoter-XGBoost [48], and BERT-Promoter [49] models in terms of all performance evaluation metrices, with better improvement.

The results indicate that the iPro-TCN model performs better than the iPSW(2L)-PseKNC [15], dPromoter-XGBoost [48], BERT-Promoter [49] model in terms of all measures for first layer and second layer, with improvements in accuracy 5.9% and 7.71%, sensitivity, 8.4% and 11.13%, specificity, 2.82% and 5.41% of the model. Receiver operating characteristic (ROC) also presented in Figure 5, and

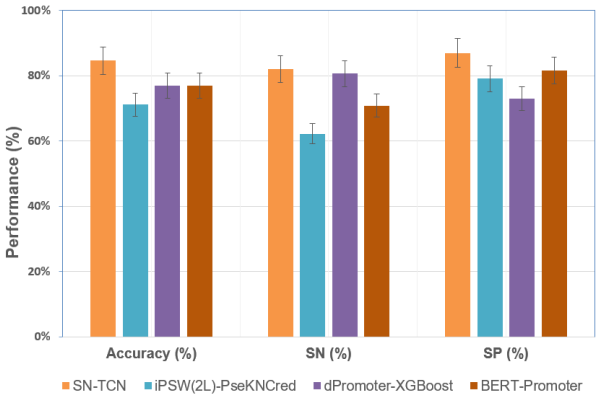

FIGURE 8. The iPro-TCN model comparison with existing method on second layer.

Figure 6 respectively, which provide additional visualizations of the model's performance. Figure 3 and 4 shows the Confusion Matrix for both layers of the model.

TABLE 4. Performance evaluation of proposed method by benchmark dataset with existing method.

Promoter Identification			
Methods	ACC	SN	SP
iPro-TCN	91.86%	92.74%	91%
iPSW(2L)-PseKNC	83.13%	81.37%	84.89%
dPromoter-XGBoost	83.81%	85.72%	81.92%
BERT-Promoter	85.45%	84.34%	86.56%
Promoter Strength			
iPro-TCN	84.63%	81.98%	87.04%
iPSW(2L)-PseKNC	71.2%	62.23%	79.17%
dPromoter-XGBoost	77%	80.65%	72.91%
BERT-Promoter	76.92%	70.85%	81.63%

According to [50], the development of web servers for computational biology models is a promising direction that can help push medical science into an ongoing revolution. In future work, it is planned to establish a web server for the iPro-TCN model in order to make it more widely accessible to researchers and practitioners in the field. This will likely involve hosting the model on a server and providing a user interface that allows users to interact with the model and obtain its predictions for specific input data. Web servers for computational biology models can be useful for a variety of purposes, such as testing the performance of the model on different datasets, comparing the results of the model to other models, and integrating the model into larger workflow systems.

V. CONCLUSION

Identifying promoters in DNA sequences is significant step towards understanding gene transcription regulation, as promoters are responsible for initiating transcription of a gene. In this research, a computational model namley: iPro-TCN was proposed based on word embedding method and a Temporal Convolutional Network to accurately predict promoters and their strength in DNA sequences. In the first layer the model predict the promoter and non-promoter and second layer of the model prdict strong and week promoter. The improved performance of the model is due to the use of dilated causal convolution in the temporal convolutional layer, which allows the model to consider the state of each promoter identification modification feature for each state. The Temporal convolutional network is effectively capture the features generated by the word embedding process. Ultimately our model has outperformed the current state-of-the-art model in both layers, it clearly indicate that the model is a significant improvement over previous methods and a useful tool for predicting promoter identification and their strength.

AVAILABILITY OF DATA AND MATERIALS

https://github.com/malikmtahir/iPro-TCN.git

ACKNOWLEDGMENT

(Ali Raza and Waleed Alam equally contributed to this work.)

REFERENCES

[1] I. A. Shahmuradov, R. Mohamad Razali, S. Bougouffa, A. Radovanovic, and V. B. Bajic, "BTSSfinder: A novel tool for the prediction of promoters in cyanobacteria and Escherichia coli," *Bioinformatics*, vol. 33, no. 3, pp. 334–340, Feb. 2017.

[2] M. Kozak, "Initiation of translation in prokaryotes and eukaryotes," *Gene*, vol. 234, no. 2, pp. 187–208, Jul. 1999.

[3] D. Sweetser, M. Nonet, and R. A. Young, "Prokaryotic and eukaryotic RNA polymerases have homologous core subunits.," *Proc. Nat. Acad. Sci. USA*, vol. 84, no. 5, pp. 1192–1196, Mar. 1987.

[4] S. Döhr, A. Klingenhoff, H. Maier, M. H. de Angelis, T. Werner, and R. Schneider, "Linking disease-associated genes to regulatory networks via promoter organization," *Nucleic Acids Res.*, vol. 33, no. 3, pp. 864–872, Feb. 2005.

[5] D. K. Hawley and W. R. McClure, "Compilation and analysis of Escherichia colipromoter DNA sequences," *Nucleic Acids Res.*, vol. 11, no. 8, pp. 2237–2255, 1983.

[6] W. Ross, K. K. Gosink, J. Salomon, K. Igarashi, C. Zou, A. Ishihama, K. Severinov, and R. L. Gourse, "A third recognition element in bacterial promoters: DNA binding by the α subunit of RNA polymerase," *Science*, vol. 262, no. 5138, pp. 1407–1413, Nov. 1993.

[7] E. E. Blatter, W. Ross, H. Tang, R. L. Gourse, and R. H. Ebright, "Domain organization of RNA polymerase α subunit: C-terminal 85 amino acids constitute a domain capable of dimerization and DNA binding," *Cell*, vol. 78, no. 5, pp. 889–896, Sep. 1994.

[8] B. Liu, F. Yang, D.-S. Huang, and K.-C. Chou, "IPromoter-2L: A two-layer predictor for identifying promoters and their types by multi-window-based PseKNC," *Bioinformatics*, vol. 34, no. 1, pp. 33–40, Jan. 2018.

[9] Q.-Z. Li and H. Lin, "The recognition and prediction of σ70 promoters in Escherichia coli K-12," *J. Theor. Biol.*, vol. 242, no. 1, pp. 135–141, Sep. 2006.

[10] K. Song, "Recognition of prokaryotic promoters based on a novel variable-window Z-curve method," *Nucleic Acids Res.*, vol. 40, no. 3, pp. 963–971, Feb. 2012.

[11] S. de Avila e Silva, F. Forte, I. T. S. Sartor, T. Andrighetti, G. J. L. Gerhardt, A. P. Longaray Delamare, and S. Echeverrigaray, "DNA duplex stability as discriminative characteristic for Escherichia coli σ54-and σ28-dependent promoter sequences," *Biologicals*, vol. 42, no. 1, pp. 22–28, Jan. 2014.

[12] H. Lin, E.-Z. Deng, H. Ding, W. Chen, and K.-C. Chou, "IPro54-PseKNC: A sequence-based predictor for identifying sigma-54 promoters in prokaryote with pseudo K-tuple nucleotide composition," *Nucleic Acids Res.*, vol. 42, no. 21, pp. 12961–12972, Dec. 2014.

[13] S. Patiyal, N. Singh, M. Z. Ali, D. S. Pundir, and G. P. S. Raghava, "Sigma70Pred: A highly accurate method for predicting sigma70 promoter in Escherichia coli K-12 strains," *Frontiers Microbiology*, vol. 13, Nov. 2022, Art. no. 1042127.
[14] M. Shujaat, J. S. Jin, H. Tayara, and K. T. Chong, "IProm-phage: A two-layer model to identify phage promoters and their types using a convolutional neural network," *Frontiers Microbiology*, vol. 13, Nov. 2022, Art. no. 1061122.
[15] X. Xiao, Z.-C. Xu, W.-R. Qiu, P. Wang, H.-T. Ge, and K.-C. Chou, "IPSW(2L)-PseKNC: A two-layer predictor for identifying promoters and their strength by hybrid features via pseudo K-tuple nucleotide composition," *Genomics*, vol. 111, no. 6, pp. 1785–1793, Dec. 2019.
[16] N. Q. K. Le, E. K. Y. Yapp, N. Nagasundaram, and H.-Y. Yeh, "Classifying promoters by interpreting the hidden information of dna sequences via deep learning and combination of continuous fasttext N-grams," *Frontiers Bioeng. Biotechnol.*, vol. 7, p. 305, Nov. 2019.
[17] S. Gama-Castro, "RegulonDB version 9.0: High-level integration of gene regulation, coexpression, motif clustering and beyond," *Nucleic Acids Res.*, vol. 44, no. D1, pp. D133–D143, Jan. 2016.
[18] W. Li and A. Godzik, "Cd-hit: A fast program for clustering and comparing large sets of protein or nucleotide sequences," *Bioinformatics*, vol. 22, no. 13, pp. 1658–1659, Jul. 2006.
[19] Y. Huang, B. Niu, Y. Gao, L. Fu, and W. Li, "CD-HIT suite: A web server for clustering and comparing biological sequences," *Bioinformatics*, vol. 26, no. 5, pp. 680–682, Mar. 2010.
[20] S. Bai, J. Zico Kolter, and V. Koltun, "An empirical evaluation of generic convolutional and recurrent networks for sequence modeling," 2018, *arXiv:1803.01271*.
[21] H. V. Dudukcu, M. Taskiran, Z. G. Cam Taskiran, and T. Yildirim, "Temporal convolutional networks with RNN approach for chaotic time series prediction," *Appl. Soft Comput.*, vol. 133, Jan. 2023, Art. no. 109945.
[22] N. Q. K. Le, "IN6-methylat (5-step): Identifying DNA N6-methyladenine sites in rice genome using continuous bag of nucleobases via chou's 5-step rule," *Mol. Genet. Genomics*, vol. 294, no. 5, pp. 1173–1182, Oct. 2019.
[23] M. Tahir, M. Hayat, and S. A. Khan, "INuc-ext-PseTNC: An efficient ensemble model for identification of nucleosome positioning by extending the concept of chou's PseAAC to pseudo-tri-nucleotide composition," *Mol. Genet. Genomics*, vol. 294, no. 1, pp. 199–210, Feb. 2019.
[24] M. Tahir and M. Hayat, "INuc-STNC: A sequence-based predictor for identification of nucleosome positioning in genomes by extending the concept of SAAC and chou's PseAAC," *Mol. BioSystems*, vol. 12, no. 8, pp. 2587–2593, 2016.
[25] F. Li, C. Li, T. T. Marquez-Lago, A. Leier, T. Akutsu, A. W. Purcell, A. I. Smith, T. Lithgow, R. J. Daly, J. Song, and K.-C. Chou, "Quokka: A comprehensive tool for rapid and accurate prediction of kinase family-specific phosphorylation sites in the human proteome," *Bioinformatics*, vol. 34, no. 24, pp. 4223–4231, Dec. 2018.
[26] F. Li, Y. Wang, C. Li, T. T. Marquez-Lago, A. Leier, N. D. Rawlings, G. Haffari, J. Revote, T. Akutsu, K.-C. Chou, A. W. Purcell, R. N. Pike, G. I. Webb, A. I. Smith, T. Lithgow, R. J. Daly, J. C. Whisstock, and J. Song, "Twenty years of bioinformatics research for protease-specific substrate and cleavage site prediction: A comprehensive revisit and benchmarking of existing methods," *Briefings Bioinf.*, vol. 20, no. 6, pp. 2150–2166, Nov. 2019.
[27] J. Li, Z. Yang, B. Yu, J. Liu, and X. Chen, "Methylation protects miRNAs and siRNAs from a 3'-end uridylation activity in arabidopsis," *Current Biol.*, vol. 15, no. 16, pp. 1501–1507, Aug. 2005.
[28] B. Liu, F. Yang, and K.-C. Chou, "2L-piRNA: A two-layer ensemble classifier for identifying piwi-interacting RNAs and their function," *Mol. Therapy Nucleic Acids*, vol. 7, pp. 267–277, Jun. 2017.
[29] B. Liu, K. Li, D.-S. Huang, and K.-C. Chou, "IEnhancer-EL: Identifying enhancers and their strength with ensemble learning approach," *Bioinformatics*, vol. 34, no. 22, pp. 3835–3842, Nov. 2018.
[30] N. Q. K. Le and V.-N. Nguyen, "SNARE-CNN: A 2D convolutional neural network architecture to identify SNARE proteins from high-throughput sequencing data," *PeerJ Comput. Sci.*, vol. 5, p. e177, Feb. 2019.
[31] T.-T.-D. Nguyen, N.-Q.-K. Le, R. M. I. Kusuma, and Y.-Y. Ou, "Prediction of ATP-binding sites in membrane proteins using a two-dimensional convolutional neural network," *J. Mol. Graph. Model.*, vol. 92, pp. 86–93, Nov. 2019.
[32] L. Malambo, S. C. Popescu, S. C. Murray, E. Putman, N. A. Pugh, D. W. Horne, G. Richardson, R. Sheridan, W. L. Rooney, R. Avant, M. Vidrine, B. McCutchen, D. Baltensperger, and M. Bishop, "Multitemporal field-based plant height estimation using 3D point clouds generated from small unmanned aerial systems high-resolution imagery," *Int. J. Appl. Earth Observ. Geoinf.*, vol. 64, pp. 31–42, Feb. 2018.
[33] 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," *Comput. Methods Programs Biomed.*, vol. 177, pp. 81–88, Aug. 2019.
[34] L. Wei, S. Wan, J. Guo, and K. K. Wong, "A novel hierarchical selective ensemble classifier with bioinformatics application," *Artif. Intell. Med.*, vol. 83, pp. 82–90, Nov. 2017.
[35] L. Wei, P. Xing, J. Zeng, J. Chen, R. Su, and F. Guo, "Improved prediction of protein–protein interactions using novel negative samples, features, and an ensemble classifier," *Artif. Intell. Med.*, vol. 83, pp. 67–74, Nov. 2017.
[36] Z. Hong, X. Zeng, L. Wei, and X. Liu, "Identifying enhancer–promoter interactions with neural network based on pre-trained DNA vectors and attention mechanism," *Bioinformatics*, vol. 36, no. 4, pp. 1037–1043, Feb. 2020.
[37] H. Wang, J. Tang, Y. Ding, and F. Guo, "Exploring associations of non-coding RNAs in human diseases via three-matrix factorization with hypergraph-regular terms on center kernel alignment," *Briefings Bioinf.*, vol. 22, no. 5, Sep. 2021, Art. no. bbaa409.
[38] X. Ma, B. Xi, Y. Zhang, L. Zhu, X. Sui, G. Tian, and J. Yang, "A machine learning-based diagnosis of thyroid cancer using thyroid nodules ultrasound images," *Current Bioinf.*, vol. 15, no. 4, pp. 349–358, Jun. 2020.
[39] Y. Ding, J. Tang, and F. Guo, "Identification of drug–target interactions via dual Laplacian regularized least squares with multiple kernel fusion," *Knowl.-Based Syst.*, vol. 204, Sep. 2020, Art. no. 106254.
[40] P. Feng, H. Ding, H. Yang, W. Chen, H. Lin, and K.-C. Chou, "iRNA-PseColl: Identifying the occurrence sites of different RNA modifications by incorporating collective effects of nucleotides into PseKNC," *Mol. Therapy Nucleic Acids*, vol. 7, pp. 155–163, Jun. 2017.
[41] W. Chen, P. Feng, H. Yang, H. Ding, H. Lin, and K.-C. Chou, "iRNA-AI: Identifying the adenosine to inosine editing sites in RNA sequences," *Oncotarget*, vol. 8, no. 3, pp. 4208–4217, Jan. 2017.
[42] H. Yang, W.-R. Qiu, G. Liu, F.-B. Guo, W. Chen, K.-C. Chou, and H. Lin, "iRSpot-Pse6NC: Identifying recombination spots in Saccharomyces cerevisiae by incorporating hexamer composition into general PseKNC," *Int. J. Biol. Sci.*, vol. 14, no. 8, pp. 883–891, 2018.
[43] W. Chen, H. Ding, X. Zhou, H. Lin, and K.-C. Chou, "iRNA(m6A)-PseDNC: Identifying N6-methyladenosine sites using pseudo dinucleotide composition," *Anal. Biochemistry*, vols. 561–562, pp. 59–65, Nov. 2018.
[44] Z.-D. Su, Y. Huang, Z.-Y. Zhang, Y.-W. Zhao, D. Wang, W. Chen, K.-C. Chou, and H. Lin, "iLoc-lncRNA: Predict the subcellular location of lncRNAs by incorporating octamer composition into general PseKNC," *Bioinformatics*, vol. 34, no. 24, pp. 4196–4204, Dec. 2018.
[45] H. Yang, H. Lv, H. Ding, W. Chen, and H. Lin, "iRNA-2OM: A sequence-based predictor for identifying 2'-O-methylation sites in homo sapiens," *J. Comput. Biol.*, vol. 25, no. 11, pp. 1266–1277, Nov. 2018.
[46] M. Greiner, D. Sohr, and P. Göbel, "A modified ROC analysis for the selection of cut-off values and the definition of intermediate results of serodiagnostic tests," *J. Immunological Methods*, vol. 185, no. 1, pp. 123–132, Sep. 1995.
[47] J. Chen, R. Long, X.-L. Wang, B. Liu, and K.-C. Chou, "DRHP-PseRA: Detecting remote homology proteins using profile-based pseudo protein sequence and rank aggregation," *Sci. Rep.*, vol. 6, no. 1, p. 32333, Sep. 2016.
[48] H. Li, L. Shi, W. Gao, Z. Zhang, L. Zhang, Y. Zhao, and G. Wang, "DPromoter-XGBoost: Detecting promoters and strength by combining multiple descriptors and feature selection using XGBoost," *Methods*, vol. 204, pp. 215–222, Aug. 2022.
[49] N. Q. K. Le, Q.-T. Ho, V.-N. Nguyen, and J.-S. Chang, "BERT-promoter: An improved sequence-based predictor of DNA promoter using BERT pre-trained model and SHAP feature selection," *Comput. Biol. Chem.*, vol. 99, Aug. 2022, Art. no. 107732.
[50] K.-C. Chou and H.-B. Shen, "REVIEW: Recent advances in developing web-servers for predicting protein attributes," *Natural Sci.*, vol. 1, no. 2, pp. 63–92, 2009.

ALI RAZA received the Master of Science (M.S.) degree in computer science (CS) from the City University of Science and Information Technology, Peshawar, in 2018. He is currently pursing the Ph.D. degree with the Department of Computer Science, Qurtuba University of Science and Information Technology, Peshawar, Pakistan. He is also a Lecturer with the Department of Computer Science, Muslim Youth University, Islamabad. His research interests include bioinformatics, machine learning, and deep learning.

WALEED ALAM received the M.Sc. degree in information technology from the Institute of Information and Technology, Quaid-i-Azam University, Islamabad, Pakistan, in 2018, and the M.S. degree in electronic and information engineering from Jeonbuk National University, Jeonju, South Korea, in 2021, where he is currently pursuing the Ph.D. degree in electronics and information engineering. His research interests include artificial intelligent, machine learning, deep learning, and image processing. His current research interest includes bioinformatics application using deep learning.

SHAHZAD KHAN received the Master of Computer Science (M.Sc.) degree from Gomal University D. I. Khan, in 2004, and the M.S. degree in computer science from Abdul Wali Khan University Mardan, Pakistan, in 2017. He is currently pursing the Ph.D. degree with the Department of Computer Science, Qurtuba University of Science and Information Technology, Peshawar, Pakistan. His research interests include bioinformatics, machine learning, and deep learning.

MUHAMMAD TAHIR received the Master of Information (M.I.T.) degree from Gomal University D. I. Khan, in 2005, the M.S. (CS) degree in multimedia and communication from Mohammad Ali Jinnah University (MAJU), Islamabad, in 2011, and the Ph.D. degree from the Department of Computer Science, Abdul Wali Khan University Mardan (AWKUM), Pakistan. He has completed postdoctoral research with the Advanced Electronics and Information Research Center, Jeonbuk National University, Jeonju, South Korea. He has been a Lecturer with the Department of Computer Science, AWKUM, since November 2010. His research interests include bioinformatics, machine learning, and deep learning.

KIL TO CHONG received the Ph.D. degree in mechanical engineering from Texas A&M University, in 1995. Currently, he is a Professor with the School of Electronics and Information Engineering, Jeonbuk National University, Jeonju, South Korea, and the Head of the Advanced Research Center of Electronics. His research interests include machine learning, signal processing, motor fault detection, network system control, and time-delay systems.