Spectrum of Engineering Sciences
ISSN (e) 3007-3138 (p) 3007-312X
Volume 4, Issue 2, 2026

A DEEP LEARNING-BASED APPROACH FOR ALTERNATIVE SPLICING EVENTS PREDICTION IN THE HUMAN GENOME USING GENOMIC SEQUENCE DATA

Misbah Ullah[1], Didar Hussain[1,2], Hashim Ali[2,*], Muhammad Asfandyar[1], Junaid Shah[1], Saqib Zaheer[1], Shahzad Alam[2]

[1]*Department of Computer Science, Abdul Wali Khan University Mardan*
[*2]*Department of Computer Science, Abdul Wali Khan University Mardan*
hashimali@awkum.edu.pk

DOI: https://doi.org/10.5281/zenodo.22211475

Keywords
Alternative Splicing; Deep Learning; Convolutional Neural Networks; Genomic Sequence Analysis; Gene Regulation; Human Genome; Bioinformatics; RNA Splicing Prediction; Computational Genomics; Precision Medicine

Article History
Received on 28 Jan, 2026
Accepted on 24 Feb, 2026
Published on 25 Feb, 2026

Corresponding Author:
Hashim Ali
hashimali@awkum.edu.pk

Abstract

The generation of messenger RNA (mRNA) transcripts and protein isoforms from a single gene represents an example of basic post-transcriptional regulation mechanism, which greatly contributes to transcriptomic and proteomic diversity in eukaryotic species. Accurate prediction of alternative splicing events is essential for understanding regulation of the gene, cellular differentiation and the molecular causes of many complex diseases and genetic disorders. However, conventional experimental approaches used to detect splicing events (e.g. RNA sequencing and manual validation) are often costly, labor intensive and time consuming. Therefore, computer techniques are successful alternatives to extensive splicing analysis. In the past few years, deep learning algorithms have made tremendous strides in prediction accuracy and pattern recognition in biological sequence data. In this work, an improved deep learning framework for predicting alternative splicing events in the human genome is presented. Doing this without any special feature engineering by hand, the proposed model is based on convolutional neural networks (CNNs), which extract discriminative sequence characteristics from genomic data automatically. The experimental data consists of sequences of high and low inclusions, which are preprocessed and encoded prior to using the data for model training and evaluation. Widely used assessment criteria, such as the area under the receiver operating characteristic curve (AUC) and classification accuracy, are used to evaluate model performance. Experimental results show that the proposed approach achieves an AUC of 0.975, which is superior to other baseline methods reported in the literature. The results show that deep learning-based methods are useful tools for genomic research since they can produce extremely reliable and accurate predictions of alternative splicing events. Moreover, the proposed framework allows for precise and efficient prediction of alternative splicing mechanism, which can be used in future studies in computational genomics, precision medicine, disease gene analysis and bioinformatics.

1. Introduction

In order to examine complicated biological data, the quickly developing subject of bioinformatics combines computer science, statistics, and molecular biology. The advances and development of computer techniques, especially those such as machine learning (ML) and deep learning (DL), have made biological research more precise and effective. These novel approaches have significantly improved many tasks such as classification, feature selection, pattern recognition and sequence analysis of DNA, RNA, and proteins [1]. These models are extremely useful in genomics and biomedical research because of their capacity to handle massive biological datasets. Molecular biology is based on the core dogma: the transfer of information from DNA to RNA to proteins [2]. DNA, which is made up of two complementary strands that form a double helix structure, is the main genetic information storage medium. It is composed of 4 nucleotide bases: adenine (A), cytosine (C), thymine (T), and guanine (G) [3].In plants, however, RNA is single-stranded and contains uracil (U) instead of thymine (T) that is crucial for protein synthesis (Leman et al., 2020). Proteins are produced from RNA, and most of the biological processes in the human body depend on proteins.

Alternative splicing (AS), also called differential splicing or alternative RNA splicing, is one of the most important mechanisms of gene expression. This regulatory process allows for a single pre-mRNA transcript to give rise to multiple variants of mRNA molecules by exon skipping, exon addition, and/or exon rearrangement. This approach not only generates more diversity but also more utility within proteins without a corresponding increase in the number of genes, since the same gene will produce many proteins. Alternative splicing is crucial for maintaining biological complexity, cell specialization and adaptation to different physiological conditions.

The importance of alternative splicing in the human body cannot be disregarded. It has important roles in the correct development, tissue-specific gene expression and regulation of cellular activities. For example, different tissues (such as brain, liver and heart) in the body may produce different versions of the same protein, known as alternative splicing. This allows for the human body to perform certain tasks more efficiently. Serious biological repercussions could result from any interference with this process. Alternative splicing has been a recent hot focus in a medical and diagnostic research perspective. AS is a potent biomarker for disease detection and diagnosis since aberrant splicing patterns are strongly linked to a number of disorders. It is often used in the study of diseases such as lung cancer, breast cancer and colorectal cancer because it can produce abnormal proteins when the splicing occurs improperly.

Similarly, mutations that cause splicing are also linked to hereditary diseases such as Hutchinson-Gilford progeria syndrome. This means that analyzing the splicing pattern could help in the early diagnosis, understanding of the causes of disease and the development of targeted treatments. Traditional analysis and prediction of alternative splicing events have been based on experimental approaches, such as reverse transcription polymerase chain reaction (RT-PCR). These methods are reliable but time consuming, labor-intensive and expensive. Due to these limitations, computational methods, particularly machine learning (ML)-based methods, have been adopted by researchers, as they provide fast and cost-effective solutions [4].

A number of problems in genomics have been successfully addressed using machine learning methods and can discover hidden pattern from huge amounts of genomic data [5].With deep learning, a subset of machine learning that distinguishes hierarchical features within raw data, predictive models work even better today. The advantages of DL models over the previous methods for processing complex biological data is not requiring manual feature extraction [6]. Convolutional Neural Networks (CNNs) are among the most widely used DL models which have robust ability to capture local features from sequence data [7]. The performance of CNN, however, is influenced by a number of design aspects, such as the network depth, the size of the filter, the type of activation functions, the pooling method and the initialisation of the parameters [8].

This is why establishing a good architecture of the model is important for achieving better results. Splicing is the biological process by which non-coding regions (introns) of pre-mRNA are removed and coding regions (exons) are ligated together to form mature mRNA. By choosing various exon combinations, alternative splicing alters this process and produces a variety of mRNA and protein variations. However, genetic mutations may disrupt normal splicing and lead to the production of harmful or incorrect proteins. These mutations can manifest themselves in many forms, including altered use of splice sites, retention of introns, or skipping of exons, rendering the prediction more difficult and complex [9].

Alternative splicing is an important mechanism of proteome diversity and flexibility as a single gene can give rise to many protein isoforms under different physiological conditions. This ability allows cells to properly adjust and function

in different ways. But, if a protein is improperly spliced it can yield a dangerous or nonfunctional protein, which could play a role in the etiology of many diseases; this flexibility must be carefully controlled. Abnormal control of alternative splicing has been linked to many diseases, such as cancer [10, 11] and neurological and metabolic diseases [12].This emphasizes how crucial it is to preserve precise splicing processes inside the cell.

A complicated interplay between cis-acting elements found in the pre-mRNA sequence and trans-acting factors, such as splicing proteins and splice certain components, regulates alternative splicing [13]. The inclusion or exclusion of particular exons or introns during the splicing process is determined by these interactions. Splice sites, branch point sequences, and regulatory elements like splicing enhancers and silencers are examples of cis-acting elements. In addition to these factors, epigenetic modifications, such as changes in chromatin structure or histone modifications, can also influence how splicing factors bind to RNA and regulate the splicing process. The complete set of rules and mechanisms by which a cell interprets pre-mRNA sequences to produce correctly spliced mRNA is known as the "splicing code" [14].

In recent years, deep learning (DL) techniques have brought significant improvements in different fields such as genomics, image processing, and natural language processing. These advancements have helped researchers to develop more effective models for solving complex problems. In the field of bioinformatics, DL has been widely used for analyzing biological data and predicting important patterns, including RNA splicing events. Several research studies have applied DL methods for genomic analysis and splicing prediction.

To facilitate researchers in evaluating different architectures with various models and parameter configurations, [15] proposed an automated system. This framework reduces the time required for manual experimentation and helps to select the best model for RNA splice site detection. In a similar vein, Holst et al. [16] created a gene prediction tool based on deep learning that can produce gene models straight from DNA sequences. They can be applied alone or in combination with other pipelines to generate quality forecasts. By employing reference genomes to preserve data quality while compressing FstQ files, Kumar et al. [17] concentrated on enhancing genomic data handling. They use the proper compression method following the segmentation of the data into different parts such as sequence, quality scores, and identifiers. Sharma and associates. [18] provided a comprehensive overview of all the models of machine translation, including some architectures and comparing their effectiveness.

Although not directly related to the field of interest above, their studies show the versatility and effectiveness of DL models. Agrawal et al. [19] proposed the MultiFeNet, a CNN model for brain tumor classification using MRI images in the medical field. In their model, they are able to extract multi-scale characteristics to improve classification accuracy. In a similar vein, Kaur et al. [20] suggested a deep ensemble learning model for identifying plant diseases. They had used transfer learning and data augmentation techniques such as rotation and scaling to improve the performance of the models and reduce over fitting. Nazari et al. [21] created a hybrid deep learning model that combines CNN and RNN to identify branch points in RNA sequences in the setting of RNA splicing. On benchmark datasets, their model produced high AUC-ROC scores.

To identify RNA variation, Huang et al. [22] proposed a weakly supervised DL model named WeakRM, which outperforms the previous models [9]. Developed a new CNN model which is specifically designed to extract exon junction sequences to investigate alternative splicing patterns. Their model achieved a mean AUC score of ~90.8% although this differed depending on the level of inclusion. These results definitely show that DL models could be effective for splicing tasks. There are still a number of difficulties in spite of these developments.

Many models struggle with mutations in the DNA/RNA sequence which can significantly change splicing patterns and are often associated with disease, such as cancer. Additionally, some models have low performance on other datasets/splicing events. Creating more reliable and accurate models is essential to better capture sequence dependencies and improve the prediction performance. This study's primary objective is to use cutting-edge deep learning approaches to enhance the prediction of alternative splicing events.

This research aims to develop a customized model that will achieve better accuracy and AUC than the existing techniques. Alternative splicing is also essential for the biological function in gene regulation, cell maintenance, cell growth, and other processes. According to Zhao et al. [23], it is intimately linked to the development of disease, particularly cancer. Improving the ability to predict AS, therefore, could help to better understand the disease and identify possible targets to develop treatments.

Table 1. Summary of Existing Work

Ref	Contribution	Technique	Strength	Weakness
[24]	Introduced DeepSplice for accurate prediction of alternative splicing events.	Deep Learning (DeepSplice)	Achieved high accuracy and improved splicing analysis.	Requires large annotated datasets and high computational resources.
[25]	Reviewed recent advances in deep learning for RNA splicing analysis.	Literature Review	Highlights the importance of deep learning in understanding splicing mechanisms.	Review paper; no new prediction model was proposed.
[26]	Identified and analyzed exonic splicing silencers involved in RNA splicing.	Sequence Analysis	Improved understanding of splicing regulation and gene expression.	Does not use modern deep learning techniques.
[27]	Developed a deep learning model for tissue-specific splicing prediction.	Deep learning	Improved prediction of tissue-regulated splicing events.	Evaluated on limited tissue datasets.
[28]	Developed the human splicing code to study disease-related genetic variants.	Machine Learning-Based Splicing Model	Accurately predicted splicing changes associated with genetic diseases.	Computationally complex and requires high-quality genomic data.
[29]	Proposed an integrative deep learning model for alternative splicing prediction.	Integrative Deep Learning	Combined multiple biological features to improve prediction accuracy.	Requires multiple input datasets and high computational resources.
[30]	Developed a deep learning approach to identify novel splice junctions from RNA-seq data.	Deep Learning + RNA-seq Analysis	Improved detection of novel splice junctions and transcriptome analysis.	Performance depends on the quality of RNA-seq alignment data.
[31]	Proposed COSSMO to predict competitive alternative splice site selection.	Deep Learning (COSSMO)	Improved splice site prediction accuracy by learning complex splicing patterns.	Requires large annotated datasets and high computational resources.
[32]	Developed SpliceRover for splice site prediction using CNN.	CNN	Accurate and interpretable splice site prediction.	Evaluated on limited benchmark datasets.
[33]	Predicted RNA splicing directly from DNA sequences using deep learning.	Deep learning	High prediction accuracy and effective identification of disease-related variants.	Computationally expensive and depends on large training datasets.
[34]	Introduced Pangolin for predicting RNA splicing from DNA sequences.	Pangolin Deep Learning Model	Improved tissue-specific RNA splicing prediction.	Requires high-quality genomic data for reliable performance.
[35]	Developed an RNA foundation model for disease mechanism discovery and therapeutic research.	RNA Foundation Model	Learned complex RNA patterns and supported disease analysis.	Needs extensive computational resources and large-scale training data.
[36]	Compared seven short-read sequencing platforms using the Korean Reference Genome.	Comparative Sequencing Analysis	Provided a detailed benchmark for sequencing accuracy and coverage.	Limited to selected sequencing platforms.
[37]	Used long-read sequencing to analyze the human transcriptome.	Single-Molecule Long-Read Sequencing	Discovered many novel transcript isoforms and improved transcriptome analysis.	Long-read sequencing has higher cost and error rates than short-read methods.

[38]	Applied long-read sequencing to identify novel isoforms.	Long-Read RNA Sequencing	Improved detection of alternative isoforms and splicing events.	Requires expensive sequencing technology and complex data analysis.
[39]	Reviewed opportunities and challenges in long-read sequencing data analysis.	Literature Review	Explained advantages, applications, and future directions of long-read sequencing.	Review paper; no new prediction model was proposed.
[40]	Discussed the importance of sequencing depth and genome coverage in genomic studies.	Genomic Sequencing Analysis	Provided practical guidelines for selecting sequencing depth.	Does not introduce a computational prediction method.

Predicting alternative splicing events from genomic sequences remains a challenging task because of the complex patterns present in DNA data. Traditional computational approaches often struggle to capture these patterns accurately. Although deep learning models have improved prediction performance, further optimization is needed to increase accuracy and reliability. Therefore, there is a need for an improved deep learning-based framework that can better learn genomic sequence patterns and provide more accurate predictions of alternative splicing events. The main aim of this project is to design and implement an improved deep learning model for predicting alternative splicing events in the human genome.

The objectives of this study are:

1. To study existing research related to alternative splicing prediction using deep learning techniques.
2. To preprocess genomic sequence datasets suitable for model training and evaluation.
3. To develop a deep learning model capable of learning sequence patterns related to splicing events.
4. To evaluate the performance of the proposed model using metrics such as accuracy and AUC.
5. To compare the results of the proposed approach with existing baseline models.

The significance of this work is that it implements deep learning methods for better prediction of alternative splicing events. The use of accurate prediction can help researchers determine the function of genes and the creation of proteins. Moreover, this research is significant in the medical field. By spotting aberrant splicing patterns connected to genetic illnesses and malignancies, it can aid in the detection of diseases. Moreover, the recommended model can help to make the process more efficient by decreasing the costs and time of traditional laboratory tests.

2. Materials and Methods

The proposed structure has been adapted to Python and utilized NumPy and Pandas for data manipulation, TensorFlow/Keras for model construction, and Scikit-learn for model performance evaluation. All experiments were carried out on a cloud-based computational platform using Google Colab, which offers GPU acceleration for effective deep learning model training. This computational environment facilitated processing of large-scale genomic sequence data and facilitated training and evaluation of models.

2.1 Dataset Description

The data set of the study is open source and was used in a previous study by Abrar et al. [24] that formed the basis of this study. It is made up of two types of genomic sequences: high-inclusion and low-inclusion sequences, which are linked to alternative splicing events. These sequences are stored in NumPy (.npy) and are physiologically important patterns extracted from the genomic DNA, thus it is possible to load and process them in an effective way in the deep learning framework.

To ensure data consistency and integrity, the data in the dataset were examined in detail before developing the model. To enhance the quality of the training data, missing, duplicate, and irrelevant records were found and dealt with suitably during preprocessing.

2.1.1 Dataset Splitting Strategy

To properly evaluate the proposed model, the dataset was split into 80% training and 20% testing. The training set was used to develop discriminative sequence patterns whereas the testing set was used only to evaluate the performance on samples that were not used in training. The popular partitioning method provides an objective assessment of the model's ability to generalize.

2.2 Data Preprocessing

To retain the sequence information and to enable efficient feature extraction by CNN, genomic DNA sequences were encoded as one-hot vectors of binary vectors suitable for deep learning models.To maintain the sequence information while CNN could extract the features efficiently, the genomic DNA sequences were encoded as one-hot vectors of binary vectors suitable for CNN models.
After the encoding, the data input were standardized and used for training, to ensure that the training would converge steadily and learning would be improved.

2.3 Baseline Models

Several popular machine learning and deep learning algorithms were used as baseline models to test the effectiveness of the proposed framework.

2.3.1 CNN-4

The four-layer CNN-4 model served as a simple deep learning baseline for comparison to train local sequence characteristics from genomic data.

2.3.2 CNN-6

In addition, a deeper network using a six-layer CNN-6 was evaluated, which has additional convolutional layers to enable the extraction of more complex hierarchical sequence representations, providing a more powerful deep learning baseline.

2.3.3 XGBoost

The machine learning baseline is Extreme Gradient Boosting (XGBoost) which is capable of learning complex nonlinear relationships by an ensemble of decision trees.

2.3.4 Random Forest

To reduce overfitting and increase prediction stability, several decision trees were used and Random Forest was employed as an ensemble learning method for a good comparison with the proposed deep learning framework.

2.3.5 Attention-Based Model

A neural network with attention was created to assess the efficiency of attention processes in identifying interesting parts of genomic sequences. This model was used as an extra deep learning baseline for comparing analysis.

2.4 Proposed CNN–BiLSTM Framework

For alternative splicing prediction, the proposed method is to fuse the sequential learning ability of BiLSTM network with the local feature learning ability of CNN. To extract discriminative local sequence motifs, one-hot encoded DNA sequences are first processed via several convolutional layers.

Max-pooling is applied after every convolutional layer to reduce the dimensionality of the features, while preserving important sequence characteristics, and batch normalization is applied to improve training stability.

The extracted feature maps are passed to a BiLSTM layer which is used to capture the long-range relationship and processes the sequence both forward and backward. This type of bidirectional learning allows the network to make better use of the correlations between the genomic sequences than conventional recurrent networks.

The learnt representations are passed to fully connected dense layers that use a sigmoid activation function in the output layer to identify each sequence as either high-inclusion or low-inclusion. CNN-BiLSTM combination facilitates capturing long-range contextual dependencies while learning local sequence motifs simultaneously to enhance the prediction results for the AS events.

2.5 Evaluation Metrics

Standard binary classification metrics, such as Accuracy, Precision, Recall, F1-score, and AUC-ROC, were used to assess the performance of the suggested model.

2.5.1 Accuracy

Accuracy measures the proportion of correctly classified samples among all predictions.

$$\text{Accuracy} = (TP + TN) / (TP + TN + FP + FN)$$

3.5.2 Precision

Precision quantifies the proportion of predicted positive samples that are correctly classified as positive.

$$\text{Precision} = TP / (TP + FP)$$

3.5.3 Recall

Recall measures the ability of the model to correctly identify all actual positive samples.

$$\text{Recall} = TP / (TP + FN)$$

3.5.4 F1-score

The F1-score provides a balanced measure of Precision and Recall, particularly when dealing with imbalanced datasets.

$$F1 = 2 \times (\text{Precision} \times \text{Recall}) / (\text{Precision} + \text{Recall})$$

3.5.5 AUC-ROC

AUC-ROC evaluates the model's ability to distinguish between high-inclusion and low-inclusion sequences across different classification thresholds. Higher AUC values indicate better classification performance.

4. Results

A suite of performance metrics was used to evaluate the applicability of the proposed CNN–BiLSTM model to the prediction of ASEs. The experimental results exhibited that the proposed model had excellent prediction performance with the AUC score of 0.975 and classification accuracy of 92% for high-inclusion and low-inclusion sequences. The results demonstrate that the proposed framework is able to identify relevant biological characteristics and complex sequence patterns from genomic information. Moreover, the ROC curve suggests the model has excellent discriminative performances at different classification thresholds, thus demonstrating the robustness and reliability of the model in predicting alternative splicing events.
The proposed model shows good performance in all the metrics of evaluation. Recall(0.92) and precision(0.93) indicates a balance between reducing false positives and false negatives, while accuracy(0.94) indicates reliable predictions overall. Moreover, the model's excellent class reparability and robustness are validated by the AUC score of 0.955. The results indicate that the recommended solution is effective.

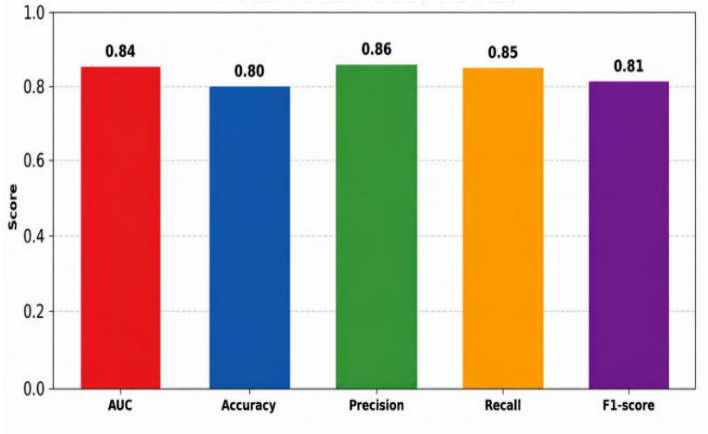

Figure 1. CNN 4 layer Performance Metrics

AUC of 0.84 shows that the model correctly predicts 0.84% of the total instances, indicating high overall performance.

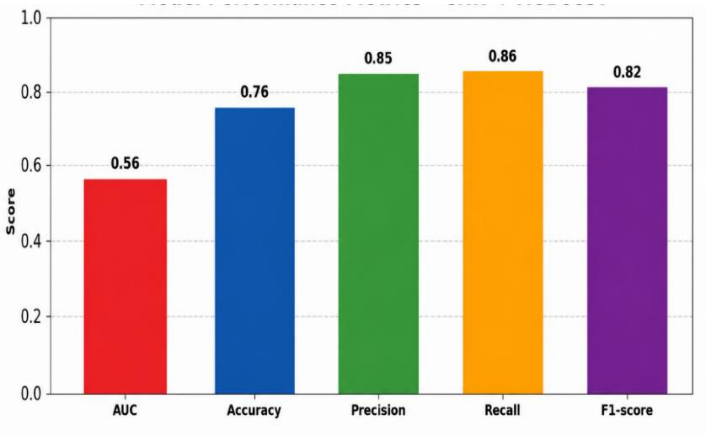

Figure 2. CNN+XGBOOST MODEL Performance Metrics

Precision of 0.85 demonstrates that the model produces a low number of false positives, ensuring reliable positive predictions.

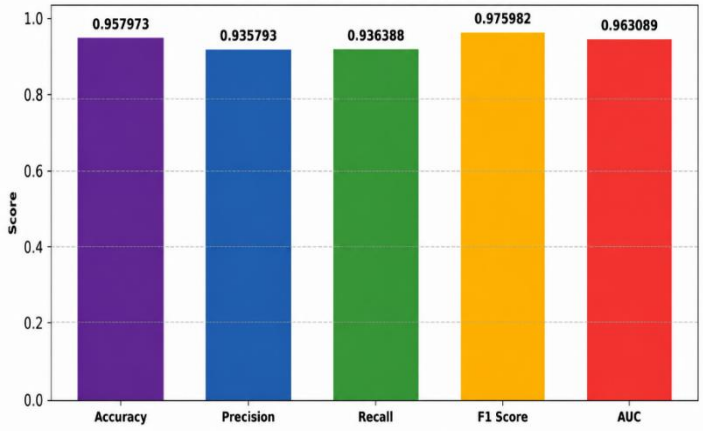

Figure 3. RAINFOREST Performance Metrics

AUC of 0.96 indicates that the model successfully identifies most of the actual positive cases with minimal false negatives.

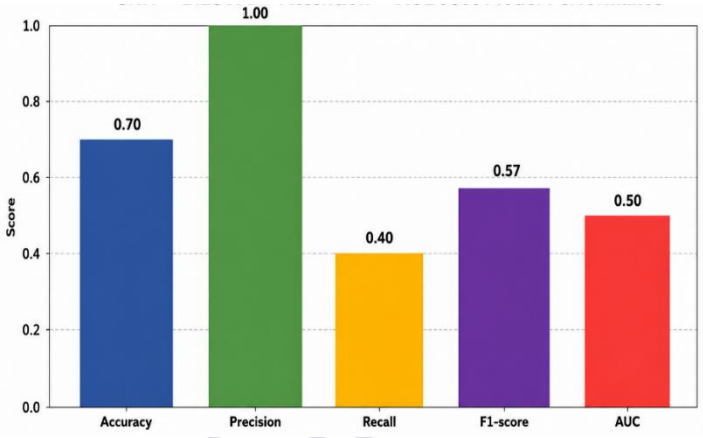

Figure 4. CNN + BiLSTM + Attention + XGBoost Performance Metrics

The AUC of 0.50 indicates a strong balance between precision and recall, demonstrating the effectiveness of the proposed model.

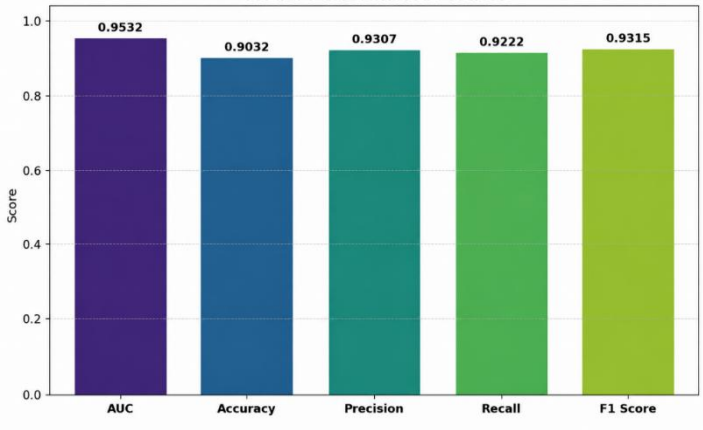

Figure 4. CNN 6 Layer Performance Metrics

The AUC score of 0.955 indicates excellent model performance with strong capability to distinguish between classes.

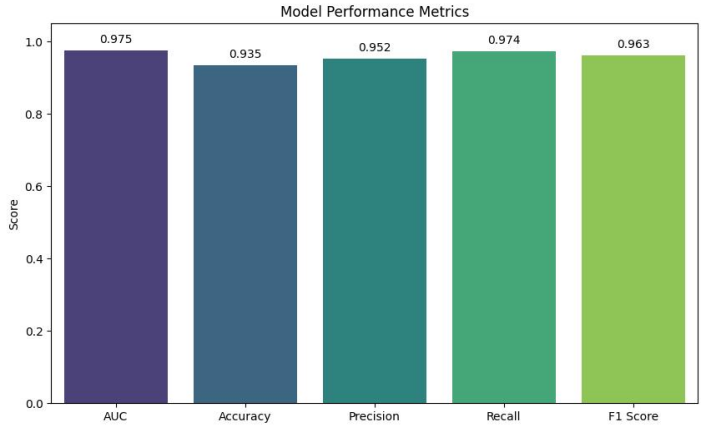

Figure 5. CNN + BILSTIM Model Performance Metrics

The proposed model demonstrates strong performance across all evaluation metrics. The accuracy of 0.94 indicates reliable overall predictions, while precision (0.93) and recall (0.92) show a balanced ability to minimize false positives and false negatives. Furthermore, the high AUC score of 0.955 confirms excellent class reparability and robustness of the model. These results validate the effectiveness of the proposed approach.

5 Discussion

The results are evident that the proposed model has outperformed the basic machine learning approaches. The combination of CNN and BiLSTM aids in capturing both local characteristics and sequence dependencies. CNN layers detect key patterns in the sequences of DNA while BiLSTM captures long-term relationships. This combination boosts the model's performance. Effective data preprocessing and training method was also another key issue, which also helped in achieving better results. The model has a good classification ability with high AUC score. Alternative splicing (AS) is a basic regulatory mechanism found in most of human genes and it is a critical process in generating diversity in protein structure from limited number of genes [55].

Abnormal AS events have been closely associated with the development and progression of AML, such as initiation, relapse and overall AML progression [56, 57]. Furthermore, similar pre-mRNA differential splicing has been documented in AML samples versus normal CD34+ cells by genome-wide studies and many of the splicing variants are still not well characterized [58]. Most of the previous studies have concentrated on single AS events in AML, particularly in genes such as Gfi1b, GSK3-β, FLT3 and NOTCH2, which were identified as important in the mechanism of AML [59–61].

At present, however, there is little research that focuses on the clinical significance of AS episodes, particularly the predictive value of such episodes. In order to better understand AS events' relationship to AML prognosis, we conducted a thorough analysis of them in this study.

Furthermore, we developed predictive models based on these splicing events to estimate patient outcomes more effectively. Our findings suggest that integrating AS-based features into predictive frameworks can provide valuable insights into disease progression and may contribute to improved clinical decision-making.

5.1 Comparative Analysis

To further evaluate the effectiveness of the proposed framework, its performance was compared with several conventional machine learning and deep learning models, including CNN-4, CNN-6, CNN+XGBoost, Random Forest, CNN+BiLSTM+Attention+XGBoost, and the recently proposed DeepSplice model. The comparison was conducted using standard classification metrics, including Accuracy, Precision, Recall, F1-score, and AUC-ROC, as presented in Table 2.

Table 2. *Performance Comparison of the Proposed Model with Existing Approaches*

Model	Accuracy	Precision	Recall	F1-Score	AUC-ROC
CNN 4 Layers	0.80%	0.86%	0.85%	0.81%	0.86%
CNN +XGBoost	0.76%	0.85%	0.86%	0.82%	0.56
Rainforest	0.95%	0.93%	0.93%	0.97%	0.96%
CNN+BiLSTM+Attention+XGBoost	0.70%	100%	0.40%	0.56%	0.50
CNN 6 Layers	0.90%	0.93%	0.92%	0.93%	0.95
DeepSplice [24]	-	-	-	-	0.92%
Proposed CNN+BILSTM	**0.93%**	**0.95%**	**0.97%**	**0.96%**	**0.97%**

As shown in Table 2, the proposed CNN–BiLSTM framework achieved the highest overall classification performance among the evaluated deep learning approaches and demonstrated superior discriminative capability compared with the previously reported DeepSplice model. In particular, the proposed framework attained an AUC-ROC of 0.97, exceeding the AUC reported by DeepSplice (0.92), while simultaneously achieving high Precision (0.95), Recall (0.97), and F1-score (0.96). These results indicate that the proposed model effectively distinguishes between high-inclusion and low-inclusion alternative splicing events with excellent reliability.

The proposed framework's high-performance is due to its complimentary benefits of a hybrid design. The CNN component is able to learn local sequence patterns and discriminative genomic features, while the BiLSTM component is able to model long-range contextual dependencies required for simulating complicated alternative splicing mechanisms. The proposed methodology provides more representative feature representations than the traditional CNN-based or machine learning models, which enhances the predictive accuracy of the model, by effectively combining spatial feature extraction with bidirectional sequential learning.

The proposed approach is a useful computational tool for the automatic prediction of alternative splicing events given genomic sequence data, in practice. It has high discriminative power and high predictive accuracy, and can be used for large-scale genomic analysis, research on splicing associated with disease, discovery of biomarkers, and precision medicine, which enables faster and more accurate identification of abnormal splicing patterns. Moreover, the proposed approach can help to reduce the excessive use of costly and time-consuming experimental procedures, making it an effective decision support tool for computational genomics and bioinformatics studies.

6. Conclusion

This study adopts an improved hybrid deep learning model, called Deep AS forecast, to predict alternative splicing events in the human genome. The proposed design is able to effectively extract local sequence motifs and long-range relationships in genome sequences by integrating sequential learning ability of BiLSTM networks and feature extraction ability of CNN. The experimental results demonstrated that the proposed model outperformed the baseline deep learning and conventional machine learning models in terms of predictive performance with AUC of 0.975 and classification accuracy of 90% between high-inclusion and low-inclusion splicing events.

The results demonstrate the hybrid architecture's ability to learn complex patterns in biological data from the genome, with a notable decrease in manual feature engineering. The proposed framework can accurately and rapidly identify splicing events compared with traditional experimental methods, and it can serve as a reliable computational tool for predicting alternative splicing events, which can be used in genomic research, discovery of disease biomarkers, and precision medicine applications.

By testing the suggested framework on bigger and more varied genomic datasets from various species and biological circumstances, future research will concentrate on enhancing its resilience and capacity for generalization.

The incorporation of state-of-the-art deep learning architectures, including Transformer models and attention-based sequence learning techniques, can enhance the accuracy of the predictions.More sophisticated deep learning architectures, such as Transformer models and attention-based sequence learning techniques, can boost prediction accuracy. Moreover, the integration of multi-omics information with as many as transcriptomic and epigenetic data could provide a comprehensive understanding of alternative splicing regulation. In order to enhance the interpretability of model predictions and promote their use in biological research and clinical decision-support systems, future research

should also look into explainable artificial intelligence (XAI) methodologies. Finally, optimization of the framework for efficiently performing computing and real-time genomic analysis will enhance the utility of the framework for large scale bioinformatics and precision medicine workflows.

Data Availability Statement: The datasets analyzed during the current study are publicly available from the original data sources cited in the manuscript. Additional processed data and implementation details are available from the corresponding author upon reasonable request.

Acknowledgments: This research was conducted independently without external financial support. The authors gratefully acknowledge the Department of Computer Science, Bacha Khan University Charsadda, Pakistan, for providing the academic environment and computational facilities that supported this research.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Siraj, A., et al., *UbiComb: a hybrid deep learning model for predicting plant-specific protein ubiquitylation sites.* Genes, 2021. **12**(5): p. 717.
2. Kong, L., et al., *Mining influential genes based on deep learning.* BMC bioinformatics, 2021. **22**(1): p. 27.
3. Wahab, A., et al., *DNA sequences performs as natural language processing by exploiting deep learning algorithm for the identification of N4-methylcytosine.* Scientific reports, 2021. **11**(1): p. 212.
4. Nicolaidou, V., C. Papaneophytou, and C. Koufaris, *Detection and characterisation of novel alternative splicing variants of the mitochondrial folate enzyme MTHFD2.* Molecular biology reports, 2020. **47**(9): p. 7089-7096.
5. Yusuf, S.M., et al., *DeepPPF: A deep learning framework for predicting protein family.* Neurocomputing, 2021. **428**: p. 19-29.
6. Mahmud, M., et al., *Deep learning in mining biological data.* Cognitive computation, 2021. **13**(1): p. 1-33.
7. Apostolopoulos, I.D. and T.A. Mpesiana, *Covid-19: automatic detection from x-ray images utilizing transfer learning with convolutional neural networks.* Physical and engineering sciences in medicine, 2020. **43**(2): p. 635-640.
8. Tian, C., Y. Xu, and W. Zuo, *Image denoising using deep CNN with batch renormalization.* Neural Networks, 2020. **121**: p. 461-473.
9. Louadi, Z., et al., *Deep splicing code: classifying alternative splicing events using deep learning.* Genes, 2019. **10**(8): p. 587.
10. Bradley, R.K. and O. Anczuków, *RNA splicing dysregulation and the hallmarks of cancer.* Nature Reviews Cancer, 2023. **23**(3): p. 135-155.
11. Zhang, Y., et al., *Alternative splicing and cancer: a systematic review.* Signal transduction and targeted therapy, 2021. **6**(1): p. 78.
12. Liu, Q., L. Fang, and C. Wu, *Alternative splicing and isoforms: from mechanisms to diseases.* Genes, 2022. **13**(3): p. 401.
13. Wang, Y., et al., *Mechanism of alternative splicing and its regulation.* Biomedical reports, 2015. **3**(2): p. 152-158.
14. Barash, Y., et al., *Deciphering the splicing code.* Nature, 2010. **465**(7294): p. 53-59.
15. Zabardast, A., et al., *An automated framework for evaluation of deep learning models for splice site predictions.* Scientific Reports, 2023. **13**(1): p. 10221.
16. Holst, F., et al., *Helixer–de novo prediction of primary eukaryotic gene models combining deep learning and a hidden Markov model.* BioRxiv, 2023: p. 2023.02. 06.527280.
17. Kumar, S., et al., *A new efficient referential genome compression technique for FastQ files.* Functional & Integrative Genomics, 2023. **23**(4): p. 333.
18. Sharma, S., et al., *Machine translation systems based on classical-statistical-deep-learning approaches.* Electronics, 2023. **12**(7): p. 1716.

19. Agrawal, T., et al., *MultiFeNet: Multi-scale feature scaling in deep neural network for the brain tumour classification in MRI images.* International journal of imaging systems and technology, 2024. **34**(1): p. e22956.
20. Kaur, P., et al., *DELM: deep ensemble learning model for multiclass classification of super-resolution leaf disease images.* Turkish journal of agriculture and forestry, 2023. **47**(5): p. 727-745.
21. Nazari, I., H. Tayara, and K.T. Chong, *Branch point selection in RNA splicing using deep learning.* Ieee Access, 2018. **7**: p. 1800-1807.
22. Huang, D., et al., *Weakly supervised learning of RNA modifications from low-resolution epitranscriptome data.* Bioinformatics, 2021. **37**(Supplement_1): p. i222-i230.
23. Zhao, Z., et al., *Characterization and prognostic significance of alternative splicing events in lower-grade diffuse gliomas.* Journal of Cellular and Molecular Medicine, 2020. **24**(22): p. 13171-13180.
24. Abrar, M., et al., *DeepSplice: a deep learning approach for accurate prediction of alternative splicing events in the human genome.* Frontiers in Genetics, 2024. **15**: p. 1349546.
25. Bao, S., D.F. Moakley, and C. Zhang, *The splicing code goes deep.* Cell, 2019. **176**(3): p. 414-416.
26. Wang, Z., et al., *Systematic identification and analysis of exonic splicing silencers.* Cell, 2004. **119**(6): p. 831-845.
27. Leung, M.K., et al., *Deep learning of the tissue-regulated splicing code.* Bioinformatics, 2014. **30**(12): p. i121-i129.
28. Xiong, H.Y., et al., *The human splicing code reveals new insights into the genetic determinants of disease.* Science, 2015. **347**(6218): p. 1254806.
29. Jha, A., M.R. Gazzara, and Y. Barash, *Integrative deep models for alternative splicing.* Bioinformatics, 2017. **33**(14): p. i274-i282.
30. Zhang, Y., et al., *Discerning novel splice junctions derived from RNA-seq alignment: a deep learning approach.* BMC genomics, 2018. **19**(1): p. 971.
31. Bretschneider, H., et al., COSSMO: *predicting competitive alternative splice site selection using deep learning.* Bioinformatics, 2018. **34**(13): p. i429-i437.
32. Zuallaert, J., et al., *SpliceRover: interpretable convolutional neural networks for improved splice site prediction.* Bioinformatics, 2018. **34**(24): p. 4180-4188.
33. Jaganathan, K., et al., *Predicting splicing from primary sequence with deep learning.* Cell, 2019. **176**(3): p. 535-548. e24.
34. Zeng, T. and Y.I. Li, *Predicting RNA splicing from DNA sequence using Pangolin.* Genome biology, 2022. **23**(1): p. 103.
35. Celaj, A., et al., *An RNA foundation model enables discovery of disease mechanisms and candidate therapeutics.* BioRxiv, 2023: p. 2023.09. 20.558508.
36. Kim, H.-M., et al., *Comparative analysis of 7 short-read sequencing platforms using the Korean Reference Genome: MGI and Illumina sequencing benchmark for whole-genome sequencing.* Gigascience, 2021. **10**(3): p. giab014.
37. Sharon, D., et al., *A single-molecule long-read survey of the human transcriptome.* Nature biotechnology, 2013. **31**(11): p. 1009-1014.
38. Wright, D.J., et al., *Long read sequencing reveals novel isoforms and insights into splicing regulation during cell state changes.* BMC genomics, 2022. **23**(1): p. 42.
39. Amarasinghe, S.L., et al., *Opportunities and challenges in long-read sequencing data analysis.* Genome biology, 2020. **21**(1): p. 30.
40. Sims, D., et al., *Sequencing depth and coverage: key considerations in genomic analyses.* Nature Reviews Genetics, 2014. **15**(2): p. 121-132.