IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 16 November 2023, accepted 9 December 2023, date of publication 11 January 2024, date of current version 22 January 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3352823

RESEARCH ARTICLE

Unveiling the Potential Pattern Representation of RNA 5-Methyluridine Modification Sites Through a Novel Feature Fusion Model Leveraging Convolutional Neural Network and Tetranucleotide Composition

WALEED ALAM[1], MUHAMMAD TAHIR[2,3], SHAHID HUSSAIN[4], SARAH GUL[5], MAQSOOD HAYAT[2], REYAZUR RASHID IRSHAD[6,7], AND FABIANO PALLONETTO[4]

[1]Department of Electronics and Information Engineering, Jeonbuk National University, Jeonju 54896, South Korea
[2]Department of Computer Science, Abdul Wali Khan University, Mardan, Khyber Pakhtunkhwa 23200, Pakistan
[3]Department of Electrical and Computer Engineering, University of Manitoba, Winnipeg, MB R3T 5V6, Canada
[4]Innovation Value Institue (IVI), School of Business, National University of Ireland Maynooth (NUIM), Maynooth, W23 F2H6 Ireland
[5]Department of Biological Sciences, Faculty of Basic and Applied Sciences, International Islamic University Islamabad, Islamabad 44000, Pakistan
[6]Department of Computer Science, College of Science and Arts, Najran University, Sharurah, Najran 68341, Saudi Arabia
[7]Science and Engineering Research Center, Najran University, Najran 68341, Saudi Arabia

Corresponding authors: Muhammad Tahir (M.tahir@umanitoba.ca) and Shahid Hussain (shahid.hussain@mu.ie)

This work was supported in The authors are thankful to Deanship of Scientific Research and under the supervision of the Science and Engineering Research Centre at Najran University for funding this work under the Research Centers Funding program grant code (NU/RCP/SERC/12/5). This work was also supported by the Science Foundation Ireland under Grant 21/SPP/3756.

ABSTRACT The 5-Methyluridine (m5U), predominantly present in RNA and especially enriched in transfer RNA (tRNA), significantly enhances translational accuracy and protein synthesis by ensuring precise genetic information decoding and optimal tRNA functionality within cellular mechanisms. The identification of m5U modification sites is crucial, as this modification has gained significant attention in diseases such as breast cancer, stress response, and viral infections, offering insights into its molecular mechanisms and regulatory functions in disease contexts. Nevertheless, due to the arduous nature, intricate procedures, reliance on sophisticated and expensive instrumentation, and the need for specialized expertise, conventional biochemical approaches for identifying m5U modification sites result in substantial resource expenditures and notable temporal investments. Consequently, the pressing need for a precise and efficient computational method highlights the urgency for alternative approaches in identifying m5U modification sites. In this study, we introduce a novel computational approach called "Deep-m5U," which combines the strengths of Convolutional Neural Networks (CNNs) and tetranucleotide composition to accurately identify methyluridine modification sites and improve overall performance. The developed Deep-m5U method leverages CNNs to accurately detect protein-coding regions aand capture relevant motifs, while incorporating tetra-nucleotide composition to capture global compositional characteristics, resulting in a more robust model that significantly enhances performance. We evaluated the Deep-m5U model on two publicly available benchmark datasets: the full transcript and mature mRNA datasets. Our results showcase superior performance, achieving accuracies of 91.26% and 95.63% respectively, surpassing the current cutting-edge methods. Moreover, the open-source code for Deep-m5U is freely accessible at: https://github.com/waleed551/Deep-m5U.

INDEX TERMS 5-methyluridine, convolutional neural network, Deep-m5U, feature fusion, protein-coding regions, RNA modification, tetranucleotide composition.

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

I. INTRODUCTION

The 5-Methyluridine (m5U) is a prominent non-canonical nucleoside that pervades RiboNucleic Acid (RNA) molecules, exerting a pivotal influence on their functionality. This vital molecule assumes a central role in the complex orchestration of RNA splicing, stability, and translation processes, thereby intricately regulating gene expression. Beyond its foundational involvement in RNA processing, m5U is a versatile participant in diverse cellular phenomena, encompassing essential functions in cell differentiation, proliferation, and programmed cell death [1]. In recent years, the investigation of RNA transcriptome modifications has rapidly advanced as a frontier in biological research, facilitated by the development of state-of-the-art high-resolution transcriptome quantification and mapping techniques [2], [3]. In recent research endeavors, the exploration of RNA modifications has revealed a substantial number, surpassing 170, with the majority of these modifications found within ribosomal ribonucleic acids (rRNAs) and transfer ribonucleic acids (tRNAs) [4], [5], [6]. These RNA modifications play crucial roles in various biological processes, encompassing embryonic stem cell (ESC) growth, metabolism, migration, cancer cell survival, DNA damage response, and environmental exposure response [6], [7], [8]. However, despite being the most common RNA modification, m5U still lacks comprehensive exploration in the scientific literature, with limited research efforts focused on its identification and in-depth functional characterization [1].

The nucleoside m5U is consistently found at position 54 within the T-loop of tRNAs in both eukaryotic and bacterial organisms. Notably, its occurrence in human mitochondrial tRNAs is observed but to a lesser extent compared to other organisms [9]. Interestingly, emerging evidence suggests that m5U may play a crucial role in the pathogenesis of breast cancer within the human genome. The implications of m5U as a potential biomarker or contributing factor to breast cancer development in humans warrant further investigation into its exact mechanisms and interactions, shedding light on its therapeutic implications and significance as a diagnostic marker [10]. On the other hand, in the context of plants, m5U has been reported to regulate both developmental processes and stress responses, while also being implicated in Systemic Lupus Erythematosus (SLE) [11]. Consequently, the precise and reliable identification of m5U sites is imperative across all species to comprehensively comprehend fundamental biological functions and processes. In this context, several advanced biochemical laboratory-based experimental techniques have been developed, including FICC-seq, iCLIP, and miCLIP-seq, to facilitate the identification of m5U sites [9], [12]. However, the demanding nature of the procedures, intricate protocols, dependency on sophisticated and expensive instrumentation, and the requirement for specialized expertise associated with conventional biochemical approaches for identifying m5U modification sites lead to substantial resource expenditures and considerable time investments [13]. Therefore, the pressing demand for a more accurate, robust, and novel computational model arises, as it can provide a cost-effective and time-efficient alternative to conventional biochemical approaches, while ensuring higher accuracy and improved robustness in the identification of m5U sites [14].

Subsequently, a myriad of computational models and predictors have emerged to address RNA modifications, encompassing diverse epigenetic marks, such as N1-methyladenosine (m1A) [15], N6-methyladenosine (m6A) [16], [17], pm6A-CNN [18], iRNA-Methyl [19], SRAMP [20], RNAm5Cfinder [21], 5-methylcytosine (m5C) [22], M6AMRFS [23], and XG-ac4C [24]. These computational tools, driven by cutting-edge algorithms and leveraging machine learning techniques [25], have significantly advanced the identification and understanding of RNA modifications, showcasing their potential to revolutionize epitranscriptomics research. In parallel, Jiang et al. [12] presented an innovative machine learning-driven computational model called m5UPred, specifically designed for the accurate identification of m5U modification sites. The m5UPred model employed two distinct feature extraction methodologies, namely nucleotide density (ND) and nucleotide chemical property (NCP), to discern crucial features from the RNA sequences. This novel approach holds promising potential to advance the field of m5U site prediction, representing a noteworthy contribution to the landscape of epitranscriptomics research [12]. Moreover, Diverse cutting-edge classification algorithms, spanning the generalized linear model (GLM) [26], Naive Bayes (NB) [27], random forest (RF) [28], Support Vector Machine (SVM) [29], Particle Swarm Optimization [30], Buffalo-Based Secure Edge-Enabled Computing [31], [32], and neural network-based models [33], were extensively employed to predict m5U sites. In a recent breakthrough, Li et al. [21] developed a machine learning-based computational model, namely iRNA-m5U, to predict m5U sites in RNA. The model leverages ND and NCP feature encoding schemes to transform RNA samples into a discrete feature space. Subsequently, a SVM was utilized as the classifier to accurately predict m5U sites [21]. Despite being extensively trained on a vast corpus of human data, the aforementioned state-of-the-art computational models, namely m5UPred [12] and iRNA-m5U [21], surprisingly demonstrated an unsatisfactory level of prediction performance when it came to accurately identifying elusive m5U sites within the genome of the widely studied Saccharomyces cerevisiae, commonly known as baker's yeast [34]. This unexpected limitation in their predictive capabilities has highlighted the need for further advancements and tailored adaptations of these models to tackle the intricacies and idiosyncrasies present in the unique RNA characteristics of this yeast species.

To bridge the existing knowledge gap, a diligent endeavor was undertaken to devise an innovative, precise, and resilient computational model, designated as Deep-m5U, with the

primary objective of predicting m5U sites. This model was meticulously designed to address the shortcomings observed in the earlier methodologies, m5UPred and to provide a more comprehensive and reliable solution to the challenging task of m5U site prediction. Through the integration of cutting-edge deep learning techniques [35] and sophisticated feature engineering, Deep-m5U strives to achieve superior performance, ensuring accurate and robust identification of m5U sites across diverse genomic contexts. The developed novel Deep-m5U computational model adhered to the well-established bioinformatics 5-steps rule, encompassing the following key stages: 1) Rigorous dataset construction or selection, 2) meticulous feature encoding, 3) judicious selection of a classification algorithm, 4) comprehensive cross-validation testing, and 5) provision of a webserver or GitHub repository for accessibility and reproducibility. In this work, our contribution can be summarized in three main aspects:

- We proposed a novel computational approach named "Deep-m5U," which synergistically combines the strengths of CNNs features and tetranucleotide composition features to accurately identify methyluridine modification sites and significantly improve overall performance. Our method employs a CNN to extract essential and generalized features from the input RNA sequences, effectively identifying accurate m5U sites by merging these features with the tetranucleotide feature space. Utilizing multiple convolutional layers as feature extractors, we capture intricate patterns and discriminative aspects for identifying methyluridine modification sites. The integration of CNN-based features with tetranucleotide information enriches the representation of RNA sequences, enhancing the model accuracy in discriminating between m5U and non-m5U sites. Finally, employing dense layers for classification based on the enriched feature map enables precise predictions, distinguishing methyluridine modification sites from background noise.
- In the developed mode, we have employed the highly effective one-hot encoding technique to transform RNA sequences into discrete feature vectors. By employing this approach, we were able to accurately capture the essential features of RNA sequences, while simultaneously mitigating data redundancy and preserving crucial structural characteristics. Consequently, our model demonstrated enhanced performance and interpretability.
- We applied the proposed Deep-m5U model to publicly available single-nucleotide resolution m5U sequencing datasets from HEK293 and HAP1 cell lines, obtained through FICC-seq and miCLIPseq technologies. Its performance was compared against the state-of-the-art m5uPred method, utilizing accuracy, sensitivity, specificity, Matthews's correlation coefficient, and the area under the curve as evaluation metrics. Impressively, the Deep-m5U model achieved remarkable accuracy rates of 91.26% and 95.63% on the full transcript and mature mRNA training datasets, respectively, and 89.43% and 93.08% accuracy on the full transcript and mature mRNA test datasets, showing its superior predictive capabilities for m5U sites on the benchmark datasets.

This article is structured as follows: Section II covers dataset collection and construction, outlining the methodology for dataset splitting into training, validation, and testing subsets. Section III provides a concise explanation of the proposed deep learning layers and the process of constructing feature maps for m5U site classification. Section V discusses the evaluation metrics, while Section VI presents a comparative analysis against related state-of-the-art computational models. Lastly, in Section VII, we conclude by discussing the findings, including limitations and possible future directions.

II. MATERIALS AND METHODS

This section discusses the underlying mechanisms of the proposed Deep-m5U model, specifically focusing on illustrating the fusion of features from CNN and the integration of tetranucleotide information. Further, the implementation of the core 5-step bioinformatics regulations with on emphasizing dataset management to enhance result accuracy are also presented. The section also outlines the mechanism for generating tetranucleotide composition feature vectors which is a vital component of the model's architecture and performance. The integration of these functional components strengthening the Deep-m5U model capability to achieve imporved results in m5U site prediction compared to existing models and thereby marking a significant advancement in epitranscriptomics research. A schematic representation of the Deep-m5U model's architecture illustrating the fusion of these functional components and feature integration is presented in Figure 1. The figure demonstrates the intrinsic coupling of CNN features with nucleotide composition, enabling the formation of a fully connected layer for effective training and future anticipation. The subsequent sections elaborate on the comprehensive integration of features, along with a detailed explanation of the CNNs and nucleotide composition mechanisms. To ensure the comprehensive integration of features and delving into the detailed mechanisms of CNNs and nucleotide composition, we provide a deeper understanding of the model's architecture, which ultimately enhances its capacity to make highly accurate predictions and advances the reliability of m5U site prediction.

A. THE MECHANISM OF TRAINING AND TESTING DATASETS

In this research, we meticulously compiled experimentally identified m5U sites from publicly available single-nucleotide resolution m5U sequencing data that encompassed two distinct cell lines: HEK293 and HAP1. These datasets were generated through advanced FICC-seq

and miCLIPseq technologies, known for their high precision in capturing m5U modifications [1], [25], [28], [36], [37], [38], [39], [40]. The data used in this study were publicly available through the Gene Expression Omnibus (GEO) database, with the accession number GEO: GSE109183. Following established research precedents, we carefully selected a sequence length of 41nt, based on its demonstrated capability to produce the most promising prediction results. This deliberate choice ensured the dataset's suitability for conducting our analysis with a focus on m5U site prediction. The positive samples in the dataset were characterized by the presence of the m5U site positioned at the center of the sequence. Conversely, the negative samples were also centered around uridine (U), but these positions did not contain m5U modifications. To construct the negative datasets, we randomly selected 10 samples from the same positive transcripts, ensuring that these negative samples only comprised unmodified uridine sites. This approach maintained the dataset's integrity and balanced representation for training and testing the model effectively. To preserve a balanced dataset, each negative set was combined with the positive set, generating 10 unique datasets with an equal 1:1 ratio of positive and negative samples. To alleviate batch variance during performance evaluation, the average voting technique was applied. This approach ensured the maintenance of a balanced representation and unbiased assessment of the model's prediction performance across the datasets. Moreover, our dataset encompasses two distinct data modalities: full transcript and mature mRNA. As a result, the predictive performance of the model was rigorously evaluated on both modalities. The full transcript data includes both the exon and intron regions of the transcript, while the mature mRNA dataset exclusively focuses on the exon region, known to harbor m5U sites. This differentiation in data modalities enables a comprehensive assessment of the model's predictive capabilities, accounting for variations in gene expression and processing between these transcript regions. Each full transcript dataset comprised 2447 positive and negative samples, while the mature mRNA dataset contained 1673 positive and negative samples. For the evaluating the Deep-m5U model we petitioned the benchmark dataset into a ratio of 80% and 20% for training and independent testing parts and characterized each of these datasets as presented in Table 1.

TABLE 1. A representation of the different techniques and the dataset used for Deep-m5U.

Techniques	Cell Line	Dataset	Number of Sites	
			Site 1	Site 2
miCLIP-Seq	HEK293	GSE78040	1,282	1,218
miCLIP-Seq	HEK293	GSE63753	1,165	455
miCLIP-Seq	HEK293	Combined	2,447	1,673

B. THE CROSS VALIDATION MECHANISM

In this section, we delve into the mechanism of cross-validation, a well-known method for analyzing the performance of ML and deep learning models. It is considered highly effective, especially when dealing with limited data sample sizes [25], [41], [42], [43]. This method involves dividing the dataset into distinct subsets, namely training, validation, and test sets [41]. Subsequently, for each sub-test set, a portion of the data is selected, while the remaining samples are utilized for the training subset. This process iterates until each data point has been considered at least once in both the training and test phases [42], [44]. This process leads to obtain robust performance metrics that help avoid the potential bias impact in the dataset, enabling an unbiased assessment of the model's generalization capabilities across various data distributions [43]. Thus, this mechanism allows each test set to represent unknown data against the trained models, resulting in an unbiased evaluation of the model's performance [45].

C. THE TETRANUCLEOTIDE COMPOSITION FEATURE VECTOR

Given the efficiency of k-mers in extracting suitable features from genomics and proteomics datasets [46], we utilize this technique to divide the long sequences into overlapping subsequences with each of length k. For example considering a DNA sequence, which can be broken down into possible subsequences of length k, with each subsequence representing a combination of four nucleotides (A, C, G, and T) [47]. It is noteworthy that the selection of k value depends on the user choice and each of the subsequence is considered as a k-mer and we have discussed it in Eq. (1). To capture the nucleotide composition in a sequence, we count the occurrences of each k-mer in the sequence. This generates a frequency vector that represents the abundance of different k-mers in the sequence [48]. The frequency of each k-mer serves as a feature, and together, these features form a numerical representation of the sequence, capturing its local sequence patterns [49]. The k-mer technique, advantageous due to its ability to provide a compact and informative representation of sequences, allowing for efficient storage and computation, finds widespread usage across various bioinformatics applications, including sequence classification, prediction tasks, and sequence alignment [50].

$$\int_{k=1}^{k-mer} = \int_{1}^{k-mer} \int_{2}^{k-mer} \int_{3}^{k-mer} \int_{4^k}^{k-mer} \tag{1}$$

Here, we selected $k = 4$ as the parameter for the k-mer technique to extract a comprehensive amount of information from 5mU sequences. For instance, a given RNA sequence 'ACCUGUACU' is divided into 'ACCU', 'CCUG', 'CUGU', 'UGUA', 'GUAC', and 'UACU' through the 4-mer representation. This process allows us to capture overlapping subsequences of length four, providing a rich and detailed representation of the sequence's local patterns.

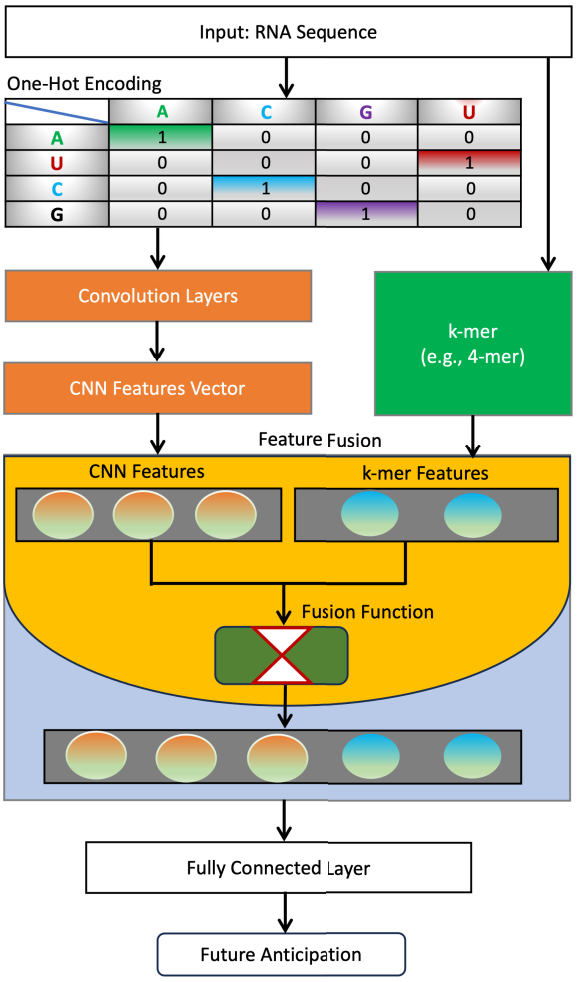

FIGURE 1. An illustration of the architecture of the proposed 5-Methyluridine (m5U) model, emphasizing the feature fusion mechanism and outlining the overall procedure of the proposed method.

By setting $k = 4$, we aim to maximize the information obtained from the sequences, enabling more accurate and meaningful analyses in our investigation. Nevertheless, each RNA sequence is transformed into a 4k-dimensional vector, where $k = 4$ in our case, leading to a $4 \times 4 \times 4 \times 4 = 256$-dimensional representation, as indicated by Eq. 2. This 256-dimensional vector captures the abundance of each 4-mer in the sequence, providing a comprehensive and detailed numerical representation of the RNA sequence, which is instrumental in our analysis and enables us to effectively leverage the k-mer technique for improved results. In sequel of this process, each RNA sequence is transformed into a 4k-dimensional vector, resulting in a long sequence with a length of 256, as demonstrated in Eq. (3).

$$\int_{k=1}^{4-mer} = \int_{1}^{4-mer} \int_{2}^{4-mer} \cdots \int_{256}^{4-mer} \tag{2}$$

$$\int_{k=1}^{4-mer} = \int_{1}(AAAA) \int_{2}(AAAC) \cdots \int_{256}(TTTT) \tag{3}$$

III. THE ROLE OF DEEP LEARNING AS A FUNDAMENTAL COMPONENT IN THE PROPOSED DEEP-M5U MODEL

The deep learning model based on a CNN has emerged as the most widely recognized and promising approach for addressing numerous biological classification and prediction problems, including sequence classification and prediction tasks [46], [47], [48]. The proposed Deep-m5U model adopts a sophisticated architecture, leveraging a series of CNN layers to extract generalized feature maps from the input data [46]. These feature maps capture essential patterns and information vital for accurate m5U site prediction. Then, concatenates the CNN features with tetra-nucleotide composition features. Subsequently, the model utilizes dense layers for classification, enabling it to make informed decisions based on the learned features [47]. This design ensures an effective and comprehensive learning process, ultimately leading to enhanced performance and precise predictions in the task of m5U site identification [48]. Thus, the proposed model exhibits increased reliability and accuracy, especially in scenarios with limited available training data and concerns about overfitting. This remarkable performance is attributed to the model's adeptness in effectively generalizing from a smaller dataset, making it a robust solution for m5U site prediction even under constraints of limited training data.

In the context of the proposed network, the one-hot encoding technique was employed to transform the RNA sequences into a format suitable for input data, as the RNA sequences are composed of four different nucleotides: Adenine (A), Cytosine (C), Guanine (G), and Uracil (U) [51]. The one-hot encoding process represents each nucleotide as a binary vector of length four, with a single "1" at the position corresponding to the nucleotide and "0"s elsewhere. Specifically, the one-hot encoding representations for the four nucleotides are as Adenine (A): (1, 0, 0, 0), Cytosine (C): (0, 0, 1, 0), Guanine (G): (0, 0, 0, 1), and Uracil (U): (0, 1, 0, 0), respectively. For instance, if we have an RNA sequence "ACGU," the corresponding one-hot encoding would be the concatenation of the one-hot vectors for each nucleotide: (1, 0, 0, 0)(0, 0, 1, 0)(0, 1, 0, 0)(0, 0, 0, 1), resulting in ([[1, 0, 0, 0], [0, 0, 1, 0], [0, 0, 0, 1],[0, 1, 0, 0]]). Thus leveraging the one-hot encoding scheme, each RNA sequence is transformed into a numerical representation that can be effectively processed by the proposed Deep-m5U network.

The benchmark dataset consists of samples, each containing RNA sequences with a fixed length of 41 nucleotides. To feed these sequences into the CNN network for m5U site classification [49], the input shape was represented as a 41×4 matrix. This matrix format enables the CNN to process the nucleotide information effectively. This impies that proposed Deep-m5U, is constructed with multiple layers, encompassing convolution layerss, normalization layers, and fully connected layers. These various layers contribute to the model's ability to learn complex patterns and features from the input data. Moreover, to optimize the model's performance, hyper-parameters have been fine-tuned using

the grid search algorithm, that systematically explores different combinations of hyper-parameter values to identify the optimal configuration for the Deep-m5U model.

The convolution layer plays a crucial role in extracting high-level features from the input data. It consists of multiple convolutional units, and their parameters are optimized through the backpropagation process. While, as an activation function for the convolution layer, the Rectified Linear Unit (ReLU) is utilized, which is widely adopted in deep learning architectures due to its ability to introduce non-linearity and alleviate the vanishing gradient problem [52]. After the convolution layer, a group normalization layer is applied, which serves as an effective alternative to batch normalization, especially when dealing with small batch sizes, as it normalizes the activations within each group, promoting stable and efficient training [53]. The normalization is applied as a regularization technique in groups, and for this study, a group size of two is chosen. Subsequently, a dropout layer with a dropout probability of 0.75 is incorporated into the proposed Deep-m5U model, as illustrated in the bird's-eye view of Figure 1. During the training process, the dropout layer is crucial in preventing overfitting by randomly deactivating certain neurons, thereby encouraging the model to rely on a more diverse set of connections and reducing the risk of memorizing specific patterns. Subsequent to the dropout layer, a fully connected layer is applied, using ReLU as the activation function. ReLU introduces non-linearity to the model, enabling it to capture complex relationships within the data [54]. To further address overfitting and promote better generalization, L2 regularization is employed on both the bias and weight terms with a regularization parameter set to $\{1 \times 10\text{exp-}2\}$. L2 regularization penalizes the model for having larger weights, encouraging the network to prioritize smaller, more evenly distributed weights [55]. The final layer in the architecture utilizes the sigmoid activation function, assigning probabilities to the outputs, allowing them to be mapped as either an m5U site or a non-m5U site [56]. The one-dimensional convolution layer is mathematically defined by Eq. (4) [57], which operates on the input RNA sequence denoted as X.

$$Cov(X)_{j,k} = ReLU\left(\sum_{s=0}^{Z-1}\sum_{n=0}^{I-1} W_{sn}^{k} X_j + s, n\right) \tag{4}$$

where the convolution filter is represented by the index k, and j signifies the index of the output position. Each W_k corresponds to a convolution filter, which is a weight matrix with dimensions $Z \times I$. Here, Z represents the size of the filter, and I denotes the number of input channels or features.

The ReLU activation function employed in the architecture is mathematically represented by Eq. (5).

$$ReLU(x) = \begin{cases} x & \text{If } x > 0 \\ 0 & \text{If } x \leq 0 \end{cases} \tag{5}$$

In this equation, the function ReLU takes the input x and returns the maximum value between 0 and x. The ReLU activation function introduces non-linearity to the model replaces negative values with zeros, effectively deactivating certain neurons in the network, which aids in preventing the vanishing gradient problem and promotes more effective learning during training [54].

The fully connected layer, combined with the dropout operation m_k having a probability p sampled from the Bernoulli distribution, is mathematically represented in Eq. (6).

$$d = ReLU\left(w_{d+1}\sum_{k=1}^{d} m_k w_k z_k\right) \tag{6}$$

where, z_k is a $1 \times d$ dimensional feature vector representing the output from the previous layer, w_k is the weight associated with z_k and w_{d+1} is the additive bias term. The operation m_k refers to the dropout process, which stochastically deactivates certain neurons in the fully connected layer during training, with a probability p sampled from the Bernoulli distribution. The output of the fully connected layer is obtained by applying the ReLU activation function to the linear combination of the weighted inputs z_k with w_k and the bias term w_{d+1}.

The sigmoid activation function, depicted in Eq. (7), utilizes the input x for constructing the iRNA-methyl model.

$$Sigmoid(x) = \frac{1}{1 + e^{-x}} \tag{7}$$

The sigmoid function applies transformation to the input x, mapping it to the value between 0 and 1, and is commonly used for binary classification tasks, like m5U site prediction in this context, where the output represents the probability of a sample belonging to a particular class. The Adam optimizer, with a learning rate of 0.00021, was chosen for training the proposed model. Adam is an adaptive learning rate optimization algorithm that efficiently updates the model's weights during the training process [58]. For the loss function, binary cross-entropy was employed. This loss function measures the discrepancy between the probability distributions of the actual class and the predicted class probabilities, making it suitable for binary classification tasks [59]. The model was trained for a maximum of 40-epochs with a batch size of 64 such that during each epoch, the model processes the training data in batches of 64 samples before updating the weights based on the optimizer's rules.

IV. EXPERIMENTAL SETUP

In this study, we utilize the Python programming language as the foundation to implement our proposed methodology for predicting m5U modification sites within mature mRNA and full transcript tRNA sequences. The BioPython package serves as our tool of choice for the extraction of data from FASTA files, ensuring a robust and efficient data acquisition process [60]. To facilitate the transformation of sequence data into suitable input for our deep learning model, we employ fundamental Python programming techniques for data encoding. For the development of our predictive

model for feature extraction and m5U site identification, we turn to industry-standard frameworks: TensorFlow [61] and Keras [62]. These libraries are renowned for their capabilities in constructing sophisticated deep-learning models. Our choice of these frameworks underscores our commitment to achieving accuracy and efficiency in the prediction of m5U modification sites.

V. PERFORMANCE EVALUATION CRITERIA

The performance metrics play a critical role in evaluating the model's performance and its capacity to accurately classify m5U sites. Consequently, the effectiveness of our model was thoroughly assessed using five key performance metrics: overall accuracy (ACC) [19], [20], Sensitivity (SN) [22], [23], [63], Specificity (SP) [28], [37], [41], [42], Matthews's correlation coefficient (MCC) [38], [39], [40], [42], [64], and Area under the Receiver Operating Characteristic (AUC) [65], [66], [67]. The overall ACC represents the ratio of correctly classified samples to the total number of samples, providing an overall measure of the model's performance as shown in Eq. (8) [19], [20]. The SN measures the proportion of actual m5U sites that are correctly identified by the model, indicating its ability to detect positive cases accurately, as shown in Eq. (9) [22], [23], [63]. The SP quantifies the model's capability to correctly classify non-m5U sites, reflecting its accuracy in identifying negative cases, as presented in Eq. (10) [28], [37], [41], [42]. The MCC provides a comprehensive evaluation of the model's performance by considering both true positive and true negative predictions, as it is particularly useful for imbalanced datasets and is presented in Eq. (11) [38], [39], [40], [42], [64]. Finally, the Area under the AUC assesses the model's ability to distinguish between positive and negative samples across different probability thresholds, offering a summary of its discriminatory power. It represents the area under the curve of the Receiver Operating Characteristic (ROC) curve. The ROC curve is a plot of the true positive rate (Sensitivity) against the false positive rate (1 - Specificity) as the discrimination threshold varies and can be represented by Eq. (12) [65], [66], [67]. The AUC value ranges from 0 to 1, where a higher AUC indicates better discriminatory power and performance of the model in distinguishing between positive and negative samples. An AUC of 0.5 indicates random guessing, while an AUC of 1.0 represents a perfect classifier.

$$ACC = \frac{TN + TP}{TP + FN + TN + FP} \tag{8}$$

$$SN = \frac{TP}{TP + FN} \tag{9}$$

$$SP = \frac{TN}{FP + TN} \tag{10}$$

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TN + FP)(TN + FN)(TP + FP)(TP + FN)}} \tag{11}$$

$$AUC = \int_0^1 SN\left(SP^{-1}d(x)\right)dx \tag{12}$$

where the terms true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN) represent the total count of different instances [48], [68], [69], [70]. To ascertain the sensitivity and specificity of the outcomes, a threshold of 0.5 was utilized. To evaluate the sensitivity and specificity of the results, a threshold value of 0.5 was employed.

VI. RESULTS AND DISCUSSION

We conducted a comprehensive evaluation of the performance of our proposed model, Deep-m5U, on two well-established benchmark datasets. These datasets consisted of a training dataset and an independent testing dataset, each serving a crucial role in assessing the model's generalization capabilities. On the training dataset, which comprised complete transcriptions, our model demonstrated remarkable proficiency across multiple performance metrics. It achieved an impressive sensitivity of 86.63%, indicating its ability to correctly identify positive instances. Additionally, the model exhibited a high specificity of 93.74%, reflecting its aptitude for accurately recognizing negative instances. Moreover, the model's overall accuracy on the training dataset was 91.26%, reaffirming its competence in correctly classifying both positive and negative samples. The Matthews correlation coefficient, a valuable metric for imbalanced datasets, achieved a noteworthy value of 0.807, further corroborating the model's robustness. Furthermore, the area under the receiver reached 0.967, signifying excellent discriminative power and a strong ability to differentiate between the two classes.

In the context of the mature mRNA mode, our model showcased commendable performance, achieving 93.48% sensitivity, 97.76% specificity, 95.63% accuracy, a Matthews correlation coefficient of 0.913, and an area under the receiver of 0.990. On a distinct note, the Deep-m5U model underwent evaluation using an independent test set, wherein it displayed differing results. For independent dataset, the full transcript mode, the model achieved 87.00% sensitivity, 91.10% specificity, 89.43% accuracy, a Matthews correlation coefficient of 0.781, and an AUC ROC of 0.953. Moreover, for independent dataset, the mature mRNA model, the corresponding metrics were 90.61% sensitivity, 95.54% specificity, 93.08% accuracy, a Matthews correlation coefficient (MCC) of 0.862, and an AUC of 0.959. A comparison of the outcomes of the Deep-m5U model with fully trascript and matured mRNA are presented in Table 2. Following that, a comprehensive analysis is carried out to compare the performance of the Convolutional Neural Network against conventional learning algorithms. This evaluation encompasses the use of both a benchmark dataset and an independent dataset to ensure robustness and generalizability. Eventually, we critically assess the outcomes produced by the Deep-m5U model and present a comparative study against previously established models.

TABLE 2. Performance evaluation of the proposed Deep-m5U pertaining to the two datasets.

Dataset	Testing Method	SN (%)	SP(%)	ACC (%)	MCC	AUC
Full transcript	Cross validation	86.63	93.74	91.26	0.807	0.967
	Independent Data	87.00	91.10	89.43	0.781	0.953
Mature mRNA	Cross validation	93.48	97.76	95.63	0.913	0.990
	Independent Data	90.61	95.54	93.08	0.862	0.959

A. MODEL HYPER-PARAMETERS

A Convolutional Neural Network model is characterized by several pivotal parameters known as hyper-parameters, whose appropriate tuning significantly impacts the model's performance. Critical aspects requiring configuration during model setup encompass hidden layers, learning rate, number of neurons, and activation functions. Among the plethora of techniques proposed for hyper-parameter tuning, the grid search approach stands as a widely embraced method for hyper-parameter optimization, as referenced in works [24], [42]. In this work, we have diligently employed the grid search approach to identify the optimal configuration values for the model's hyper-parameters. The grid search approach for hyper-parameter tuning due to its efficiency in systematically exploring the hyper-parameter space [71]. Thereby, evaluating a predefined set of hyper-parameter values arranged in a grid, the facilitate to thoroughly assess the model's performance across various configurations. Consequently, through systematic exploration of various hyper-parameter combinations, we have successfully determined the most effective setup. The resulting optimal configuration values of these hyper-parameters are comprehensively listed in Table 3.

TABLE 3. The optimum hyper-parameters utilized for the proposed Deep-m5U method.

Parameters	Optimum Values
Convolution Layer	2
Filter of Convolution Layer	8,16
Stride	1
Padding	same
Epochs	50
Filter Size	7,9
Batch size	32
L-2 Regularizer	le-3
Dropout	0.75
Optimizer	Adam Method
Learning rate	0.002

1) PERFORMANCE EVALUATION OF DEEP-M5U WITH BENCHMARK AND INDEPENDENT TESTING DATASETS

When contrasted with existing state-of-the-art methodologies, our novel computational approach, Deep-m5U, emerges as a remarkably dependable solution. The outcomes presented in Table 4 exemplify how the Deep-m5U approach effectively enhances sensitivity, specificity, accuracy, and Matthews correlation coefficient. Through rigorous experimentation, we have conclusively established the clear superiority of our prediction approach, surpassing the performance of current methods by a substantial margin. This remarkable success can be attributed primarily to the inherent strengths of the convolutional neural network features with tetra-nucleotide composition features employed in our model.

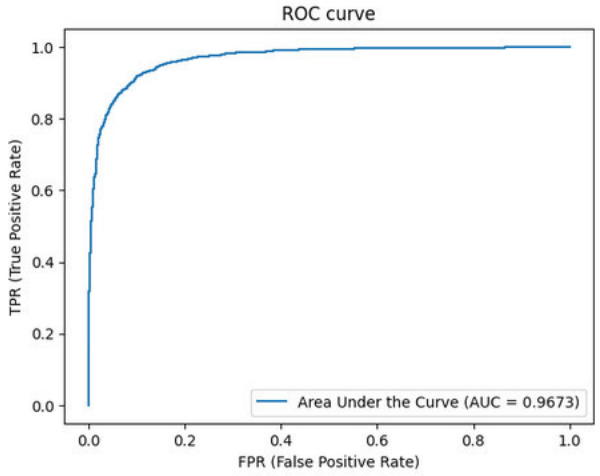

FIGURE 2. The full-transcript AUC on training dataset.

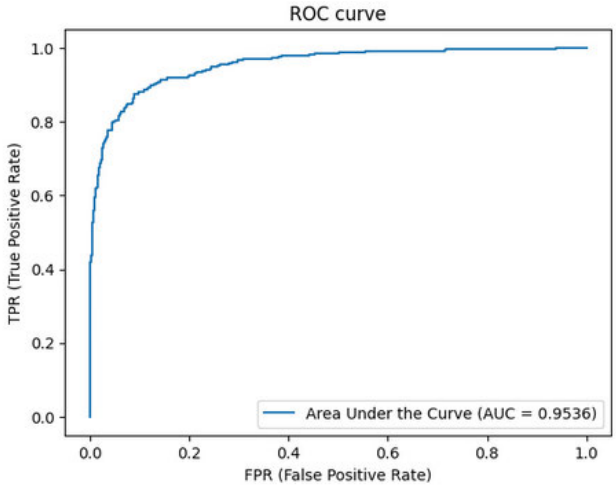

FIGURE 3. The full-transcript AUC on testing dataset.

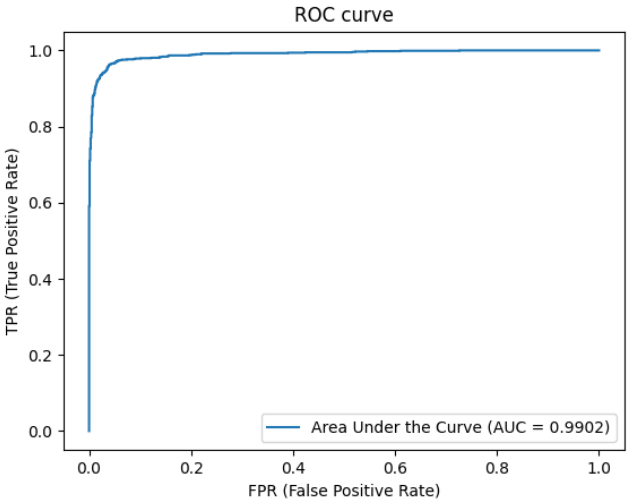

FIGURE 4. The mature mRNA AUC on training dataset.

Based on our current understanding, Deep-m5U stands out as the most reliable computational technique for precisely locating m5U sites within the human transcriptome. Consequently, we conducted a comprehensive comparison of Deep-m5U and m5UPred in their capacity to accurately identify m5U sites. Through evaluation on both the full-transcript and mature mRNA datasets, it was evident that m5UPred only managed to achieve an accuracy of 88.32% and 89.91%, respectively, in detecting m5U sites. In contrast, the Deep-m5U model displayed significantly higher accuracy in pinpointing m5U sites, reaching 91.26% for the full-transcript dataset and an impressive 95.63% for the mature mRNA dataset. Detailed performance comparisons on the training dataset can be found in Table 4. Our proposed Deep-m5U model convincingly outperformed the existing computational model, as demonstrated in Table 5.

2) PERFORMANCE COMPARISON OF DEEP-M5U WITH OTHER CUTTING-EDGE MODELS USING TRAINING DATASETS AND INDEPENDENT DATASETS

In comparison to state-of-the-art methods, our proposed Deep-m5U computational approach exhibits superior robustness in achieving success. As evidenced by the data presented in Table 4, our prediction method, Deep-m5U, yields substantial improvements in sensitivity, specificity, accuracy, and MCC. The experimental outcomes unambiguously establish the significant superiority of our proposed Deep-m5U method over the existing approach. This remarkable accomplishment can be primarily attributed to the effective utilization of a convolutional neural network in our model. To the best of our knowledge, Deep-m5U stands as the accurate computational method for effectively identifying m5U sites within the human transcriptome. Thus, we conducted a thorough performance comparison between Deep-m5U and m5UPred to assess their efficacy in m5U site identification. The evaluation outcomes on training datasets indicated that m5UPred achieved an accuracy of 89.04%

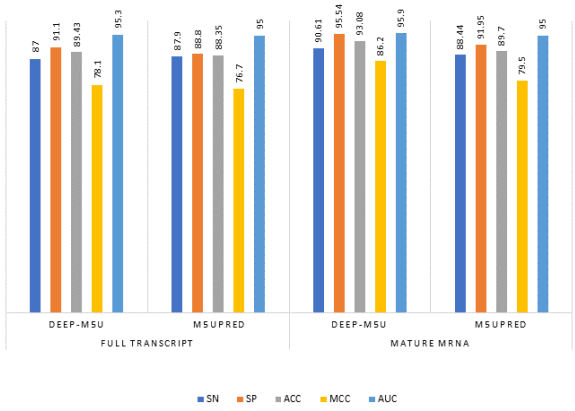

FIGURE 5. The proposed method training results comparison with existing state-of-the-art method.

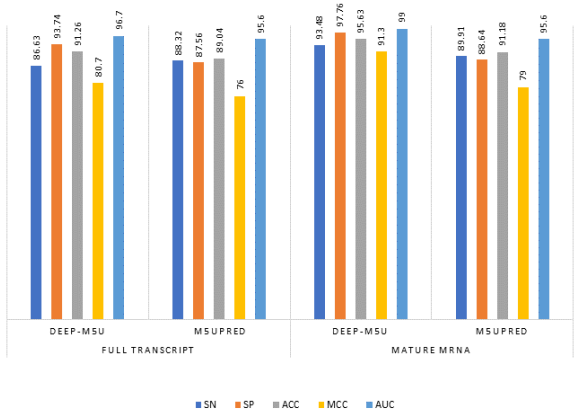

FIGURE 6. The proposed method independent test results comparison with existing state-of-the-art method.

and 91.18% for identifying m5U sites in the full-transcript and mature mRNA datasets, respectively. Conversely, the Deep-m5U model displayed superior accuracy on training dataset, achieving 91.26% for the full-transcript dataset and an impressive 95.63% for the mature mRNA dataset in m5U site identification. For detailed insights into the performance comparison on the training dataset, please refer to Table 4 and Figure 5.

In a similar vein, we performed a comprehensive performance comparison between Deep-m5U and m5UPred on independent datasets to assess their efficacy in identifying m5U sites. As evident from the results presented in Table 5, our proposed Deep-m5U model consistently outperformed the existing m5UPred models across all evaluation parameters. Remarkably, our model achieved an impressive accuracy of 89.43% on the Full transcription dataset and an even higher accuracy of 93.08% on the Mature mRNA dataset. Detailed performance comparisons on the independent datasets are meticulously documented in Table 5 and Figure 6.

VII. CONCLUSION

In this paper, we introduced "Deep-m5U," a novel computational approach that combines the strengths of Convolutional Neural Networks and tetra-nucleotide composition to

TABLE 4. Performance evaluation of the proposed method compared to existing methods using training datasets.

Dataset	Testing Method	SN (%)	SP(%)	ACC (%)	MCC	AUC
Full transcript	Deep-m5U	86.63	93.74	91.26	0.807	0.967
	m5uPred	88.32%	87.59%	89.04	0.76	0.956
Mature mRNA	Deep-m5U	93.48	97.76	95.63	0.913	0.990
	m5uPred	89.91%	88.64%	91.18	0.79	0.956

TABLE 5. Performance evaluation of the proposed method compared to existing methods using independent datasets.

Dataset	Testing Method	SN (%)	SP(%)	ACC (%)	MCC	AUC
Full transcript	Deep-m5U	87.00	91.10	89.43	0.781	0.953
	m5uPred	87.90	88.80	88.35	0.767	0.95
Mature mRNA	Deep-m5U	90.61	95.54	93.08	0.862	0.959
	m5uPred	88.44	91.95	89.70	0.795	0.95

accurately identify 5-Methyluridine (m5U) modification sites in RNA molecules. The m5U modification plays a crucial role in essential cellular processes, making the identification of these sites vital for understanding molecular mechanisms and regulatory functions in disease contexts. The proposed Deep-m5U model uniquely combines CNNs, which are proficient in detecting protein-coding regions and capturing relevant motifs, with the tetra-nucleotide composition to capture global compositional characteristics, thereby enabling the model to extract both local and global features, contributing to its robust performance. One critical aspect of Deep-m5U is the utilization of one-hot encoding that transforms RNA sequences into numerical inputs, facilitating the learning process and enhancing prediction accuracy. We conducted assessments on two benchmark datasets: the full transcript and mature mRNA datasets. Remarkably, Deep-m5U demonstrated outstanding performance with accuracies of 91.26% and 95.63%, respectively, surpassing current state-of-the-art methods. The model's high AUC ROC values on the training (0.997) and testing (0.953) datasets for full transcript, as well as on independent datasets (0.99 and 0.959) for mature mRNA, further validate its efficacy and reliability. The Deep-m5U model's ability to identify m5U sites with superior accuracy and efficiency makes it a valuable tool for drug discovery and academic research, providing crucial insights into molecular processes and regulatory functions. It is crucial to note that the performance of computational models is significantly influenced by the datasets used. The quality of these datasets varies based on factors such as diversity, size, and the representations of both training and testing data, ultimately impacting the model's performance. In our current approach, we applied the m5U model to two datasets of considerable size and evaluated its performance against state-of-the-art approaches like miCLIP-Seq [9] and m5uPred [12]. We further conducted evaluations using various training and testing datasets to comprehensively assess the model's capabilities. However, for ensuring the accuracy and authenticity of our proposed model, we plan to extend it by incorporating more comprehensive and diverse training datasets. This extension has the potential to significantly enhance the model's robustness and improve its generalization capabilities in practical applications.

ACKNOWLEGMENT

The authors are thankful to Deanship of Scientific Research and under the supervision of the Science and Engineering Research Centre at Najran University for funding this work under the Research Centers Funding program grant code (NU/RCP/SERC/12/5). This work was also suppoted by the Science Foundation Ireland under Grant 21/SPP/3756.

REFERENCES

[1] P. Feng and W. Chen, "IRNA-m5U: A sequence based predictor for identifying 5-methyluridine modification sites in Saccharomyces cerevisiae," *Methods*, vol. 203, pp. 28–31, Jul. 2022.

[2] Z. Song, D. Huang, B. Song, K. Chen, Y. Song, G. Liu, J. Su, J. P. D. Magalhães, D. J. Rigden, and J. Meng, "Attention-based multi-label neural networks for integrated prediction and interpretation of twelve widely occurring RNA modifications," *Nature Commun.*, vol. 12, no. 1, p. 4011, Jun. 2021.

[3] P. N. Pratanwanich, F. Yao, Y. Chen, C. W. Q. Koh, Y. K. Wan, C. Hendra, P. Poon, Y. T. Goh, P. M. L. Yap, J. Y. Chooi, W. J. Chng, S. B. Ng, A. Thiery, W. S. S. Goh, and J. Göke, "Identification of differential RNA modifications from nanopore direct RNA sequencing with xPore," *Nature Biotechnol.*, vol. 39, no. 11, pp. 1394–1402, Nov. 2021.

[4] P. Nombela, B. Miguel-López, and S. Blanco, "The role of m^6 A, m^5 C and ψ RNA modifications in cancer: Novel therapeutic opportunities," *Mol. Cancer*, vol. 20, no. 1, pp. 1–30, 2021.

[5] M. Tahir, H. Tayara, M. Hayat, and K. T. Chong, "KDeepBind: Prediction of RNA-proteins binding sites using convolution neural network and k-gram features," *Chemometric Intell. Lab. Syst.*, vol. 208, Jan. 2021, Art. no. 104217.

[6] Y. Huang, X. J. Shen, Q. Zou, S. P. Wang, S. M. Tang, and G. Z. Zhang, "Biological functions of micro RNAs: A review," *J. Physiol. Biochem.*, vol. 67, pp. 129–139, Mar. 2011.

[7] P. Haruehanroengra, Y. Y. Zheng, Y. Zhou, Y. Huang, and J. Sheng, "RNA modifications and cancer," *RNA Biol.*, vol. 17, no. 11, pp. 1560–1575, Nov. 2020.

[8] J. A. Abbott, C. S. Francklyn, and S. M. Robey-Bond, "Transfer RNA and human disease," *Frontiers Genet.*, vol. 5, p. 158, Jun. 2014.

[9] J.-M. Carter, W. Emmett, I. R. Mozos, A. Kotter, M. Helm, J. Ule, and S. Hussain, "FICC-seq: A method for enzyme-specified profiling of methyl-5-uridine in cellular RNA," *Nucleic Acids Res.*, vol. 47, no. 19, pp. e113–e113, Nov. 2019.

[10] R. Jacob, S. Zander, and T. Gutschner, "The dark side of the epitranscriptome: Chemical modifications in long non-coding RNAs," *Int. J. Mol. Sci.*, vol. 18, no. 11, p. 2387, Nov. 2017.

[11] F. Basta, F. Fasola, K. Triantafyllias, and A. Schwarting, "Systemic lupus erythematosus (SLE) therapy: The old and the new," *Rheumatol. Therapy*, vol. 7, no. 3, pp. 433–446, 2020.

[12] J. Jiang, B. Song, Y. Tang, K. Chen, Z. Wei, and J. Meng, "M5UPred: A web server for the prediction of RNA 5-Methyluridine sites from sequences," *Mol. Therapy-Nucleic Acids*, vol. 22, pp. 742–747, Dec. 2020.

[13] A.-K. Minnaert, H. Vanluchene, R. Verbeke, I. Lentacker, S. C. De Smedt, K. Raemdonck, N. N. Sanders, and K. Remaut, "Strategies for controlling the innate immune activity of conventional and self-amplifying mRNA therapeutics: Getting the message across," *Adv. Drug Del. Rev.*, vol. 176, Sep. 2021, Art. no. 113900.

[14] N. Pouyanfar, S. Z. Harofte, M. Soltani, S. Siavashy, E. Asadian, F. Ghorbani-Bidkorbeh, R. Keçili, and C. M. Hussain, "Artificial intelligence-based microfluidic platforms for the sensitive detection of environmental pollutants: Recent advances and prospects," *Trends Environ. Anal. Chem.*, vol. 34, Jun. 2022, Art. no. e00160.

[15] C. Zhang and G. Jia, "Reversible RNA modification N1-methyladenosine (m1A) in mRNA and tRNA," *Genomics, Proteomics Bioinf.*, vol. 16, no. 3, pp. 155–161, Jun. 2018.

[16] Q. Zou, P. Xing, L. Wei, and B. Liu, "Gene2vec: Gene subsequence embedding for prediction of mammalian N6-methyladenosine sites from mRNA," *RNA*, vol. 25, no. 2, pp. 205–218, Feb. 2019.

[17] M. U. Rehman, H. Tayara, Q. Zou, and K. T. Chong, "i6mA-caps: A capsulenet-based framework for identifying DNA N6-methyladenine sites," *Bioinformatics*, vol. 38, no. 16, pp. 3885–3891, 2022.

[18] W. Alam, S. D. Ali, H. Tayara, and K. T. Chong, "A CNN-based RNA N6-methyladenosine site predictor for multiple species using heterogeneous features representation," *IEEE Access*, vol. 8, pp. 138203–138209, 2020.

[19] W. Chen, P. Feng, H. Ding, H. Lin, and K.-C. Chou, "IRNA-methyl: Identifying N6-methyladenosine sites using pseudo nucleotide composition," *Anal. Biochemistry*, vol. 490, pp. 26–33, Dec. 2015.

[20] Y. Zhou, P. Zeng, Y.-H. Li, Z. Zhang, and Q. Cui, "SRAMP: Prediction of mammalian N6-methyladenosine (m6A) sites based on sequence-derived features," *Nucleic Acids Res.*, vol. 44, no. 10, p. e91, Jun. 2016.

[21] J. Li, Y. Huang, X. Yang, Y. Zhou, and Y. Zhou, "RNAm5Cfinder: A web-server for predicting RNA 5-methylcytosine (m5C) sites based on random forest," *Sci. Rep.*, vol. 8, no. 1, p. 17299, Nov. 2018.

[22] S. Akbar, M. Hayat, M. Iqbal, and M. Tahir, "IRNA-PseTNC: Identification of RNA 5-methylcytosine sites using hybrid vector space of pseudo nucleotide composition," *Frontiers Comput. Sci.*, vol. 14, no. 2, pp. 451–460, Apr. 2020.

[23] X. Qiang, H. Chen, X. Ye, R. Su, and L. Wei, "M6AMRFS: Robust prediction of N6-methyladenosine sites with sequence-based features in multiple species," *Frontiers Genet.*, vol. 9, p. 495, Oct. 2018.

[24] W. Alam, H. Tayara, and K. T. Chong, "XG-ac4C: Identification of N4-acetylcytidine (ac4C) in mRNA using eXtreme gradient boosting with electron-ion interaction pseudopotentials," *Sci. Rep.*, vol. 10, no. 1, pp. 1–10, Dec. 2020.

[25] R. R. Irshad, S. Hussain, S. S. Sohail, A. S. Zamani, D. Ø. Madsen, A. A. Alattab, A. A. A. Ahmed, K. A. A. Norain, and O. A. S. Alsaiari, "A novel IoT-enabled healthcare monitoring framework and improved grey wolf optimization algorithm-based deep convolution neural network model for early diagnosis of lung cancer," *Sensors*, vol. 23, no. 6, p. 2932, Mar. 2023.

[26] J. R. Oaks, C. W. Linkem, and J. Sukumaran, "Implications of uniformly distributed, empirically informed priors for phylogeographical model selection: A reply to hickerson et al.," *Evolution*, vol. 68, no. 12, pp. 3607–3617, Dec. 2014.

[27] Z. Li, J. Mao, D. Huang, B. Song, and J. Meng, "RNADSN: Transfer-learning 5-Methyluridine (m5U) modification on mRNAs from common features of tRNA," *Int. J. Mol. Sci.*, vol. 23, no. 21, p. 13493, Nov. 2022.

[28] C. Ao, X. Ye, T. Sakurai, Q. Zou, and L. Yu, "M5U-SVM: Identification of RNA 5-methyluridine modification sites based on multi-view features of physicochemical features and distributed representation," *BMC Biol.*, vol. 21, no. 1, p. 93, Apr. 2023.

[29] S. Salekin, M. Mostavi, Y.-C. Chiu, Y. Chen, J. Zhang, and Y. Huang, "Predicting sites of epitranscriptome modifications using unsupervised representation learning based on generative adversarial networks," *Frontiers Phys.*, vol. 8, p. 196, Jun. 2020.

[30] R. R. Irshad, S. S. Sohail, S. Hussain, D. Ø. Madsen, M. A. Ahmed, A. A. Alattab, O. A. S. Alsaiari, K. A. A. Norain, and A. A. A. Ahmed, "A multi-objective bee foraging learning-based particle swarm optimization algorithm for enhancing the security of healthcare data in cloud system," *IEEE Access*, p. 1, 2023.

[31] R. R. Irshad, S. Hussain, I. Hussain, I. Ahmad, A. Yousif, I. M. Alwayle, A. A. Alattab, K. M. Alalayah, J. G. Breslin, M. M. Badr, and J. J. P. C. Rodrigues, "An intelligent buffalo-based secure edge-enabled computing platform for heterogeneous IoT network in smart cities," *IEEE Access*, vol. 11, pp. 69282–69294, 2023.

[32] S. Abimannan, E.-S.-M. El-Alfy, Y.-S. Chang, S. Hussain, S. Shukla, and D. Satheesh, "Ensemble multifeatured deep learning models and applications: A survey," *IEEE Access*, vol. 11, pp. 107194–107217, 2023.

[33] H. Wang, S. Wang, Y. Zhang, S. Bi, and X. Zhu, "A brief review of machine learning methods for RNA methylation sites prediction," *Methods*, vol. 203, pp. 399–421, Jul. 2022.

[34] L. Yu, Y. Zhang, L. Xue, F. Liu, R. Jing, and J. Luo, "Evaluation and development of deep neural networks for RNA 5-methyluridine classifications using autoBioSeqpy," *Frontiers Microbiol.*, vol. 14, May 2023, Art. no. 1175925.

[35] S. Abimannan, E.-S.-M. El-Alfy, S. Hussain, Y.-S. Chang, S. Shukla, D. Satheesh, and J. G. Breslin, "Towards federated learning and multi-access edge computing for air quality monitoring: Literature review and assessment," *Sustainability*, vol. 15, no. 18, p. 13951, Sep. 2023.

[36] 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.

[37] H. Yang, H. Lv, H. Ding, W. Chen, and H. Lin, "IRNA-2OM: A sequence-based predictor for identifying 2'-O-Methylation sites inHomo sapiens," *J. Comput. Biol.*, vol. 25, no. 11, pp. 1266–1277, Nov. 2018.

[38] Y. Zhang, M. Wang, Z. Wang, Y. Liu, S. Xiong, and Q. Zou, "MetaSEM: Gene regulatory network inference from single-cell RNA data by meta-learning," *Int. J. Mol. Sci.*, vol. 24, no. 3, p. 2595, Jan. 2023.

[39] J. Jin, Y. Yu, R. Wang, X. Zeng, C. Pang, Y. Jiang, Z. Li, Y. Dai, R. Su, Q. Zou, K. Nakai, and L. Wei, "IDNA-ABF: Multi-scale deep biological language learning model for the interpretable prediction of DNA methylations," *Genome Biol.*, vol. 23, no. 1, pp. 1–23, Oct. 2022.

[40] C. Wang, Q. Zou, Y. Ju, and H. Shi, "Enhancer-FRL: Lmproved and robust identification of enhancers and their activities using feature representation learning," *IEEE/ACM Trans. Comput. Biol. Bioinf.*, vol. 20, no. 2, pp. 967–975, Mar. 2023.

[41] H. Tayara, M. Tahir, and K. T. Chong, "ISS-CNN: Ldentifying splicing sites using convolution neural network," *Chemometric Intell. Lab. Syst.*, vol. 188, pp. 63–69, May 2019.

[42] C. Ao, Q. Zou, and L. Yu, "RFhy-m2G: Identification of RNA N2-methylguanosine modification sites based on random forest and hybrid features," *Methods*, vol. 203, pp. 32–39, Jul. 2022.

[43] 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. BioSyst.*, vol. 12, no. 8, pp. 2587–2593, 2016.

[44] Z. Abbas, M. U. Rehman, H. Tayara, Q. Zou, and K. T. Chong, "XGBoost framework with feature selection for the prediction of RNA N5-methylcytosine sites," *Mol. Therapy*, vol. 31, no. 8, pp. 2543–2551, Aug. 2023.

[45] A. Senanayake, H. Gamaarachchi, D. Herath, and R. Ragel, "Deep selectnet: Deep neural network based selective sequencing for Oxford nanopore sequencing," *BMC Bioinf.*, vol. 24, no. 1, p. 31, 2023.

[46] W. Alam, H. Tayara, and K. T. Chong, "I4mC-deep: An intelligent predictor of N4-methylcytosine sites using a deep learning approach with chemical properties," *Genes*, vol. 12, no. 8, p. 1117, Jul. 2021.

[47] S. D. Ali, W. Alam, H. Tayara, and K. T. Chong, "Identification of functional piRNAs using a convolutional neural network," *IEEE/ACM Trans. Comput. Biol. Bioinf.*, vol. 19, no. 3, pp. 1661–1669, May 2022.

[48] H. Zulfiqar, Z.-J. Sun, Q.-L. Huang, S.-S. Yuan, H. Lv, F.-Y. Dao, H. Lin, and Y.-W. Li, "Deep-4mCW2 V: A sequence-based predictor to identify N4-methylcytosine sites in Escherichia coli," *Methods*, vol. 203, pp. 558–563, Jul. 2022.

[49] K. Niu, X. Luo, S. Zhang, Z. Teng, T. Zhang, and Y. Zhao, "IEnhancer-EBLSTM: Identifying enhancers and strengths by ensembles of bidirectional long short-term memory," *Frontiers Genet.*, vol. 12, Mar. 2021, Art. no. 665498.

[50] Z. Teng, L. Shi, H. Yu, C. Wu, and Z. Tian, "Measuring functional similarity of lncRNAs based on variable K-mer profiles of nucleotide sequences," *Methods*, vol. 212, pp. 21–30, Apr. 2023.
[51] Y. Wang, S. Tai, S. Zhang, N. Sheng, and X. Xie, "PromGER: Promoter prediction based on graph embedding and ensemble learning for eukaryotic sequence," *Genes*, vol. 14, no. 7, p. 1441, Jul. 2023.
[52] M. Tahir, M. Hayat, R. Khan, and K. T. Chong, "An effective deep learning-based architecture for prediction of N7-methylguanosine sites in health systems," *Electronics*, vol. 11, no. 12, p. 1917, Jun. 2022.
[53] L. Huang, J. Qin, Y. Zhou, F. Zhu, L. Liu, and L. Shao, "Normalization techniques in training DNNs: Methodology, analysis and application," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 8, pp. 10173–10196, Aug. 2023, doi: 10.1109/TPAMI.2023.3250241.
[54] S. Ankalaki and M. N. Thippeswamy, "A novel optimized parametric hyperbolic tangent swish activation function for 1D-CNN: Application of sensor-based human activity recognition and anomaly detection," *Multimedia Tools Appl.*, pp. 1–31, May 2023.
[55] A. M. J. N. Sman, "Lessons from the human mind: Enhancing resilience in deep learning models," Ph.D. dissertation, Dept. Comput. Sci., Florida State Univ., Tallahassee, FL, USA, 2023.
[56] A. Mehrish, N. Majumder, R. Bharadwaj, R. Mihalcea, and S. Poria, "A review of deep learning techniques for speech processing," *Inf. Fusion*, vol. 99, Nov. 2023, Art. no. 101869.
[57] L. Ruthotto and E. Haber, "Deep neural networks motivated by partial differential equations," *J. Math. Imag. Vis.*, vol. 62, no. 3, pp. 352–364, Apr. 2020.
[58] T. Niu, J. Wang, H. Lu, W. Yang, and P. Du, "Developing a deep learning framework with two-stage feature selection for multivariate financial time series forecasting," *Expert Syst. Appl.*, vol. 148, Jun. 2020, Art. no. 113237.
[59] Y. Ho and S. Wookey, "The real-world-weight cross-entropy loss function: Modeling the costs of mislabeling," *IEEE Access*, vol. 8, pp. 4806–4813, 2020.
[60] P. J. Cock, T. Antao, J. T. Chang, B. A. Chapman, C. J. Cox, A. Dalke, I. Friedberg, T. Hamelryck, F. Kauff, B. Wilczynski, and M. J. L. de Hoon, "Biopython: Freely available Python tools for computational molecular biology and bioinformatics," *Bioinformatics*, vol. 25, no. 11, pp. 1422–1423, 2009.
[61] M. Abadi et al., "TensorFlow: Large-scale machine learning on heterogeneous distributed systems," 2016, *arXiv:1603.04467*.
[62] F. Chollet. (2015). *Keras*. [Online]. Available: https://github.com/fchollet/keras
[63] Z. Abbas, H. Tayara, and K. T. Chong, "ENet-6 mA: Identification of 6 mA modification sites in plant genomes using ElasticNet and neural networks," *Int. J. Mol. Sci.*, vol. 23, no. 15, p. 8314, Jul. 2022.
[64] M. Kabir, M. Arif, F. Ali, S. Ahmad, Z. N. K. Swati, and D.-J. Yu, "Prediction of membrane protein types by exploring local discriminative information from evolutionary profiles," *Anal. Biochemistry*, vols. 564–565, pp. 123–132, Jan. 2019.
[65] F. Ali, M. Kabir, M. Arif, Z. N. Khan Swati, Z. U. Khan, M. Ullah, and D.-J. Yu, "DBPPred-PDSD: Machine learning approach for prediction of DNA-binding proteins using discrete wavelet transform and optimized integrated features space," *Chemometric Intell. Lab. Syst.*, vol. 182, pp. 21–30, Nov. 2018.
[66] 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.
[67] S. Singh, Y. Yang, B. Póczos, and J. Ma, "Predicting enhancer-promoter interaction from genomic sequence with deep neural networks," *Quant. Biol.*, vol. 7, no. 2, pp. 122–137, Jun. 2019.
[68] 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.
[69] H. Lv, F. Dao, and H. Lin, "DeepKla: An attention mechanism-based deep neural network for protein lysine lactylation site prediction," *iMeta*, vol. 1, no. 1, p. e11, Mar. 2022.
[70] F.-Y. Dao, H. Lv, W. Su, Z.-J. Sun, Q.-L. Huang, and H. Lin, "IDHS-deep: An integrated tool for predicting DNase I hypersensitive sites by deep neural network," *Briefings Bioinf.*, vol. 22, no. 5, p. bbab047, Sep. 2021.
[71] J. Viehweg, K. Worthmann, and P. Mäder, "Parameterizing echo state networks for multi-step time series prediction," *Neurocomputing*, vol. 522, pp. 214–228, Feb. 2023.

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, and currently focused on bioinformatics applications using deep learning.

MUHAMMAD TAHIR received the M.I.T. degree from Gomal University, Dera Ismail Khan, Pakistan, in 2005, the M.S. (C.S.) 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, in 2017. He was a Postdoctoral Researcher with the Advanced Electronics and Information Research Center, Jeonbuk National University, Jeonju, South Korea. Since November 2010, he has been an Assistant Professor with the Department of Computer Science, AWKUM. He is currently a Postdoctoral Fellow with the University of Maniotba, Winnipeg, MB, Canada. His main research interests include bioinformatics, machine learning, and deep learning.

SHAHID HUSSAIN received the B.S. degree in mathematics and the M.Sc. degree in computer science from the University of Peshawar, in 2002 and 2005, respectively, and the M.S. and Ph.D. degrees in computer engineering from Jeonbuk National University, South Korea, in 2016 and 2020, respectively. He achieved Jeonbuk National University Presidential Award for academic excellence during the Ph.D. studies. He was a Postdoctoral Researcher with the Gwangju Institute of Science and Technology (GIST), South Korea, in 2020, and the University of Galway (UoG), Ireland, from 2020 to 2022. He is currently a Senior Postdoctoral Researcher with the School of Business, Innovative Value Institute (IVI), National University of Ireland Maynooth (NUIM), Ireland. His research interests include smart grid, energy management, electric vehicles, smart grid infrastructure, optimization algorithms, microgrid operations, distributed energy resources, peer-to-peer energy trading, machine learning in medical applications (e.g., prediction and risk analysis of osteoporosis) using fuzzy logic, game theory, ontology, AI, and blockchain approaches and technologies.

SARAH GUL received the Ph.D. and Postdoc degrees in molecular medicine from the University of Ulm, Germany, in 2013 and 2014, respectively. She is currently working as an Assistant Professor and focal Person Quailty Enhancement at the Department of Biological Sciences, Faculty of Basic and Applied Sciences, International Islamic University Islamabad, Pakistan. Her research interests include cancer genetics, molecular medicine, and solving biological problems using AI and machine learning.

MAQSOOD HAYAT received the M.C.S. degree from Gomal University, Dera Ismail Khan, Pakistan, in 2004, the M.S. degree in software and system engineering from Mohammad Ali Jinnah University (MAJU), Islamabad, in 2009, and the Ph.D. degree from the Department of Computer and Information Sciences, Pakistan Institute of Engineering and Applied Sciences, Islamabad, Pakistan. Since August 2012, he has been a Professor. His main research interests include machine learning, pattern recognition, evolutionary computing, and its application in bioinformatics.

REYAZUR RASHID IRSHAD received the B.Sc. degree from Aligarh Muslim University, Aligarh, India, in 2000, and the master's degree in computer application from Indira Gandhi University, New Delhi, India, in 2010. He is currently pursuing the Ph.D. degree with JJT University, Rajasthan. He is a Lecturer with the Department of Computer Science, Najran University, Saudi Arabia. He has published many articles in reputed journals and has attended some conferences. His research interest includes web-based applications.

FABIANO PALLONETTO received the M.Sc. degree in computer science from Pisa University and the Ph.D. degree in engineering from University College Dublin. He is currently a Professor in management information systems with the National University of Ireland Maynooth (NUIM), Ireland, and a member of the Energy Institute, University College Dublin. His research interests include data analytics, digital business, intelligent energy systems, electric mobility, building flexibility, demand response, and data architectures. He is also working on the integration of electric vehicles in the power system and energy flexibility for buildings and smart districts.