NCC - An Efficient Deep Learning Architecture for Non-Coding RNA Classification

Konstantinos Vasilas*, Evangelos Makris†, Christos Pavlatos‡, Ilias Maglogiannis*
*Department of Digital Systems, University of Piraeus, 18534 Piraeus, Greece
Email: imaglo@unipi.gr, vasilas.cei@gmail.com
† School of Electrical and Computer Engineering, National Technical University of Athens, 9 Iroon Polytechniou St., 15780 Athens, Greece
Email: vmakris@mail.ntua.gr
‡ Hellenic Air Force Academy, Dekelia Air Base, Acharnes, 13671 Athens, Greece
Email: christos.pavlatos@hafa.haf.gr

***Abstract*—In this paper an efficient deep-learning architecture is proposed, aiming to classify a significant category of RNA, the non-coding RNAs (ncRNAs). These RNAs participate in various biological processes and play an important role in gene regulation as well. Because of their diverse nature, the task of classifying them is a hard one in the bioinformatics domain. All these parameters inspire this work, and specifically, the design of a neural network classifier called NCC. This deep neural network is appropriately trained to identify patterns in ncRNAs leveraging well-known datasets, which are publicly available. Additionally, a ten times larger dataset than the available ones is created for better training and testing. In terms of performance, the suggested model showcases a 6% enhancement in precision compared to prior state-of-the-art systems, with an accuracy level of 92.69%, in the existing dataset. In the larger one, its accuracy rate exceeded 98% as well as all the related tools compared, pointing to high prediction capability, that can act as a base for further findings in the ncRNAs analysis and the genomics field in general.**

***Index Terms*—Non-Coding RNA, Deep Learning, Neural Network, Classification.**

I. Introduction

NOn coding RNAs (ncRNAs) is a type of RNA that contributes to various essential molecular procedures. When these aspects arose, many researchers identified many potential challenges, and their identification has become of main interest in the field of biology and bioinformatics. Due to these significant functions, the accurate prediction of the families associated with different ncRNAs was published in the literature using experimental and computational methods. Conventional experimental methods such as in [1] were proposed, but the fact that they are time-consuming, labor intensive, and expensive lead the research to computational ones. The two many categories are the Sequence or Secondary Structure-based Methods and the Homologous Sequence Alignment Methods. The first methods analyze the 1D or 2D structure of ncRNAs to classify them accordingly. The main disadvantage of these methods, is that their accuracy is bounded with the predicted secondary structures, which is a standalone hard task to accomplished. The second approach as presented in [2] align ncRNAs with their homologues to identify common characteristics and according to those, predict their families. This method performs very well in many cases, but is bounded by an accurate secondary structure annotation for the sequences and is not capable of handling pseudoknotted structures. NCC archiecture is directed toward the classification of ncRNA sequences by means of deep learning. RNA sequences can be submitted to the model and no further information is required. The only requirement to transform input RNA sequences into one hot encoded is the one mentioned. The architecture consists of a convolution layer with a pooling layer to downsize the data. Next, a bi-directional RNN layer is used to analyze possible patterns forward and backward. Finally, a fully connected layer integrates all elements produced and generates the final prediction. The primary aim of NCC is to establish a deep learning-based architecture where high accuracy and faster training results can be achieved. The primary focus of this paper is to assist scientists with the design of future non coding RNA (ncRNA) classification algorithms by providing a solid background knowledge of surrounding information. Considering the state-of-the-art approaches adopted in the domain, problems to be solved, and the steps already accomplished the work intends to provide the biologists and bioinformaticians with the most needed information and skills to create more intelligent, more precise, and faster functioning ncRNA classifiers.

A. RNA

Ribonucleic acid, or RNA for short, is a molecule involved in biological activities. It is involved in the control of information, gene expression, and protein synthesis. RNA is a single-stranded molecule made of nucleotides. A base (adenine, cytosine, guanine, or uracil), a sugar known as ribose, and a phosphate group make up each nucleotide. In short, transcription is the process by which DNA is converted into RNA, which assists in converting the code into proteins. The cell's ability to perform essential duties for the organism's general health and optimal operation depends on many forms of RNA, including messenger RNA (mRNA), transfer RNA (tRNA), ribosomal RNA (rRNA), and small nuclear RNA (snRNA).

- From DNA, messenger RNA (mRNA) is created. transfers information to the ribosomes found in the cytoplasm

from the nucleus. It is translated at the ribosomes, where protein synthesis occurs.
- Transfer RNA (tRNA) is a molecule that carries amino acids to the ribosomes during protein synthesis. Using mRNA codons, which are three nucleotide sequences that encode different amino acids, to match acids, this operation is accomplished. Then, these acids are incorporated into the expanding chain.
- A component of ribosomes, which are cellular structures in charge of protein synthesis, is ribosomal RNA (rRNA). In addition to supporting ribosomes, rRNA aids in accelerating chemical events that result in the synthesis of proteins.
- A class of RNA molecules that support cellular functions includes small nuclear RNA (snRNA). Participating in RNA splicing, which entails cutting out coding sections from mRNA called introns and putting together coding sequences called exons, is one of its roles.

In addition to the RNA types already discussed, there exist other RNA molecules that play vital regulatory roles in cells. Among these are small interfering RNAs (siRNA) and microRNAs (miRNA). Their main function is to block genes or degrade mRNA molecules in order to control the expression of genes. When it comes to information transfer and gene expression regulation within cells, RNA in general plays a part. This molecule is essential for maintaining life and is actively involved in many important chemical reactions in nature.

B. Non-coding RNA

An important type of RNA that has been identified is ncRNAs ([3]), which play a key role and should not be overlooked whenever cellular mechanisms are discussed. Even though the original idea surrounding RNA revolved around the transfer of information for the creation of proteins, the reality of ncRNAs came to be that they were non-protein coding themselves, yet impacted protein activities along with gene regulation to be involved in other biological functions too. ncRNAs are categorized in regard to their size and function. MicroRNAs (miRNAs) and small interfering RNAs (siRNAs) are examples of small non-coding RNAs (ncRNAs), which are typically shorter than 200 nucleotides. Both of these RNAs are essential for RNA interference, which is a vital step in the silencing of genes following transcription. Long non-coding RNAs (lncRNAs) are a different subgroup that are longer than 200 nucleotides and have a variety of regulatory roles, including altering the structure of chromatin and regulating transcriptional activity. Moreover, transfer RNAs (tRNAs) and ribosomal RNAs (rRNAs), which are widely recognized for their roles in protein synthesis, are also categorized as non-coding RNAs. This demonstrates that the genome contains coding sequences. 13 classes of ncRNA families are registered in Rfam [25], which have been employed in this work as well as numerous other studies. 10% of RNAcentral [5] sequences, as reported in [6], do not belong to any known Rfam family and may be members of new families. Specific examples of the classes of ncRNAs that make up this number are the piRNAs and siRNAs, which, with one minor exception, comprise 5% of this total. In summary, neither of these classes is structurally relevant for inclusion in Rfam for effective structure modeling. To understand the function of an RNA molecule and design RNA molecules with specific functions, it is essential to predict its secondary structure. Numerous computational tools, including Knotify [7] and its variations [8], [9], [10], Knotty [11], IPknot [12], and RNAcon [13], are capable of predicting secondary structure of RNAs. These tools use a variety of techniques to tackle this task with respect to the sequence of interest. The three-dimensional structure of an RNA molecule is crucial for its function. The tertiary structure of RNA is made up of more complex folding, in which different parts fold in complex ways, as opposed to the basic structure of nucleotides or the secondary structure of simple loops and helices. Interactions between the RNA molecule's component parts, such as base pairs which are widely spaced in the molecule, have an impact on this folding process. The form that this folding yields determines the function of the RNA, which includes its participation in the control of gene expression, protein synthesis, and other cellular processes. All these different structures of RNA are usually used to train the models for the classification of the non-coding RNAs, which is the task that this study is focused on. In the literature, which is analyzed in the following Section, many systems utilize the primary, secondary or tertiary structure of the molecule, or a combination of them, to enhance their performance.

II. Related Work

An important issue in genomics is the classification of ncRNAs because of their contribution to a variety of biological processes. A number of computational techniques have been developed recently to enhance the precision of this task. Some of these technologies stand out because of their distinct advantages and are presented in this section. RNAcon is a tool that leverages structure and sequence alignment data to classify non-coding RNAs. Because of its dual methodology, RNAcon is capable to classify and identify several kinds of non-coding RNAs with remarkable accuracy, especially when it comes to secondary structures and conserved RNA motifs. A graph-based depiction of RNA sequences is presented by GraPPLE (Graph-based Prediction of ncRNA using Pairwise Labeled Edges) [16], in which nucleotides are described as nodes and their interactions as edges. GraPPLE's effectively captures complex structural links seen in RNA sequences, achieving high classification accuracy, particularly for newly discovered or poorly defined ncRNA families. NCodR [14] is a multi-class support vector machine classifier for identifying non-coding RNAs in plants, using distinct regions in the distribution of AU and minimun folding energy as additional attributes. Another classifier has been proposed, called MncR. This approach evaluates different approaches by utilizing primary sequences and secondary structures, using different neural network architectures. The ncRFP system [17], or non-coding RNA Feature Predictor, extracts and chooses pertinent characteristics for ncRNA classification. Large-scale

genomic research frequently deal with high-dimensional data, which ncRFP excels at managing through the use of a random forest classifier. The tool's excellent prediction accuracy is a result of its capacity to include various sequence-based and structural characteristics. Deep learning methods are also used for tackling this task. nRC (non-coding RNA Classifier) [18] for example, is able to discriminate between coding and non-coding RNAs. High sensitivity and specificity are demonstrated by this method. It integrates sequence-based characteristics with deep neural networks, especially in the differentiation of closely related RNA families. In another work, Residual networks, or ResNets, are used by ncDLRES (non-coding RNA Deep Learning Residual Network) [19] to classify ncRNAs. By addressing the degradation, this deep learning strategy improves classification level. ncDLRES is well-suited for genome-wide classification tasks because of its exceptional efficacy in handling sizable and intricate datasets. Similarly, ncDENSE (non-coding RNA Dense Network) [20] identifies complex patterns in RNA sequences using dense neural networks. The model can identify novel ncRNAs, even in cases where the available data for training are deficient. The accuracy with which ncDENSE can reliably classify data under a range of circumstances. Finally, incorporating a variety of deep learning models, including recurrent neural networks (RNNs) and convolutional neural networks (CNNs), BioDeepfuse [21] is a state-of-the-art framework for classifying non-coding RNAs. BioDeepfuse captures long-range relationships and local sequence patterns within RNA sequences by merging these models. A review of the other similar systems, which leverage machine learning techniques for this task, is also shown in [22].

III. Implementing the Prediction Mechanism.

The paper's implementation of the nncRNA classification model is the main topic of this part. This section, which is broken down into multiple parts, describes the procedures used to prepare the data, create the model architecture, train the model, and assess it. This paper aims to create a ncRNA classifier that uses the primary structure as its input. The errors of secondary structure prediction methods (graph characteristics) are not included in the classification flow, since the focus is solely on the primary structure of the RNA. Thus, the core contribution of this paper is focused on the development of a ncRNA classifier with the following elements:

1) The model's input is the primary structure of RNA.
2) The proposed model should be simple, as ncRFP [17], for short training periods.
3) High level of accuracy and to all prediction-related metrics.

A. Data preparation

The preparation of the data is conducted in two stages. The first stage is dedicated to padding and cutting the sequences when necessary, to homogenize the input data. The second one applies to all RNAs the one-hot encoding to prepare a format easy to interpret by the neural network.

1) Data preprocessing: Padding and cutting are common methods for the preparation of sequences ready for use in deep learning models. This is quite common in problems including sequential data tasks like speech recognition and natural language processing (NLP). By verifing a consistent length for each sequence, the model's validity and efficiency is ensured in terms of the computational aspect and homogeneity. To create a dataset with sequences of equal length, additional components can be used. Usually, zero padding or filler values are used for this task. The model is designed in such a way that features are learned from sequences of all lengths, which helps reduce overfitting risk. In this case, cutting is applied for longest sequences, and specifically, the final parts in the sequence are eliminated. In that way, better memory usage is achieved, while at the same time, the model is not too complex and difficult to train. By inspecting Figure 1 the majority of the

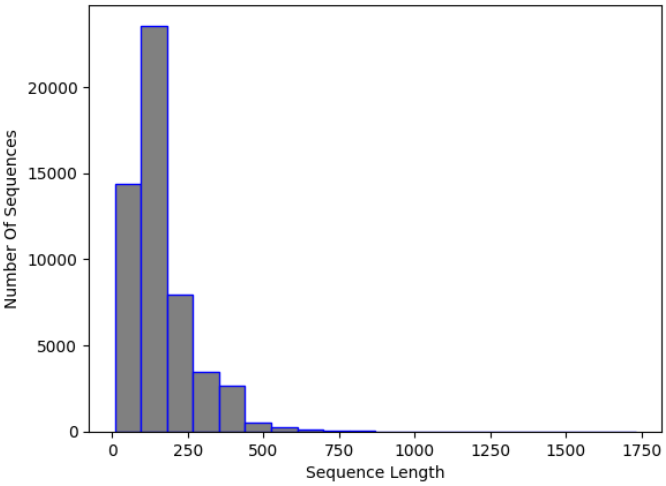

Fig. 1. NCC dataset, RNA Sequences length distribution

sequence has less than 500 nucleotides. So, it has been decided that the input sequence for NCC will be 500 lengths. This means that sequences made up of less than 500 nucleotides will be artificially padded with 0s, while sequences with more than 500 will be truncated, running the risk of ignoring some important sequence features.

2) One-hot encoding: In order to prepare categorical data for feeding to our ML classifier, One-Hot Encoding, which is a procedure that transforms the input into a numerical representation, is used. Leveraging this method, a binary vector is made for every dataset category.

TABLE I
RNA Bases in One hot encoding format

Base	4 digits	8 digits
A	1000	1000 0010
T/U	0100	0100 0001
G	0010	0010 1000
C	0001	0001 0100
X	0000	0000 0000

One-Hot Encoding's main benefit is the prevention of a misleading hierarchical order, which may happen if the representation consists only of numerical values, such as 1 for

category A, 2 for Category B, etc. Algorithms may incorrectly interpret the data in the integer format to suggest that, contrary to what is true, Category C is greater than B and B is greater than A. One-Hot Encoding uses binary vectors to maintain the uniqueness of each category without requiring any sort of order. One limitation of numerous categories for a categorical variable is that One-Hot Encoding might greatly expand the dataset's feature space. This extra dimension raises the possibility of overfitting in machine learning models as well as computational complexity. As a result, it is frequently applied sparingly and occasionally in combination with dimensionality reduction strategies. Before feeding the RNA sequences into the deep learning model, one-hot encoding has been used to encode RNA sequences, since the RNA bases are not inherently related. There are 16 possible characters to represent, as explained in Table II, but most of them are either absent or very rare in the dataset. To this end, to encode the data, the four RNA bases are described accordingly, while additional IUPAC characters are represented by 'X', which is also the character used for padding. Table I depicts the encoding of characters to 4 and 8 digits. This approach was selected over the full IUPAC 16-character encoding because the NCC and nRC datasets include only a very limited number of the remaining characters listed in Table II.

B. NCC Model Architecture

For the scope of this study 25 unique models were evaluated in the NCC dataset. A set of experiments that involved dense networks and convolutional networks as standalone networks and in a sequential way, was conducted. But finally, it was the recurrent models that were fitted with a bi-directional neural network (BiRNN) [24], that performed the best in accuracy among all types of networks tested. Recent investigations have demonstrated that Bi-directional RNN outperforms its counterparts such as standard RNN in learning sequential data by moving through the sequence in both directions. This is important for understanding the RNA case because in most cases it is determined by the neighboring nucleotides on either side of it. Nonetheless, a huge amount of context or patterns can be easily overlooked if typical RNNs are solely applied because RNNs operate in one direction. A BiRNN is a two-directional approach where the RNA sequence is estimated by looking at the nucleotides in the center and both ends. Its flexibility to be able to see the sequence dependencies in all possible places gives a better appreciation of how the RNA is structured and its functioning. The unique classification of BiRNN's gives it an upper hand over other sequence order tasks such as RNA classification. The model with the highest prediction capability and tolerable loss margins while maintaining a fairly simple architecture, is presented. It comprises a one-dimensional convolutional neural network, a one-dimensional max pooling layer, a bidirectional recurrent layer, and a fully connected layer. The neural network is presented in Figure 2.

TABLE II
RNA IUPAC SYSTEM'S ENCODING FORMAT

IUPAC Code	Meaning	Complement	Encoding
A	A	T	1000 0000 0000 0000
C	C	G	0100 0000 0000 0000
G	G	C	0010 0000 0000 0000
T/U	T	A	0001 0000 0000 0000
M	A or C	K	0000 1000 0000 0000
R	A or G	Y	0000 0100 0000 0000
W	A or T	W	0000 0010 0000 0000
S	C or G	S	0000 0001 0000 0000
Y	C or T	R	0000 0000 1000 0000
K	G or T	M	0000 0000 0100 0000
V	A or C or G	B	0000 0000 0010 0000
H	A or C or T	D	0000 0000 0001 0000
D	A or G or T	H	0000 0000 0000 1000
B	C or G or T	V	0000 0000 0000 0100
N	G or A or T or C	N	0000 0000 0000 0010
X	None	-	0000 0000 0000 0000

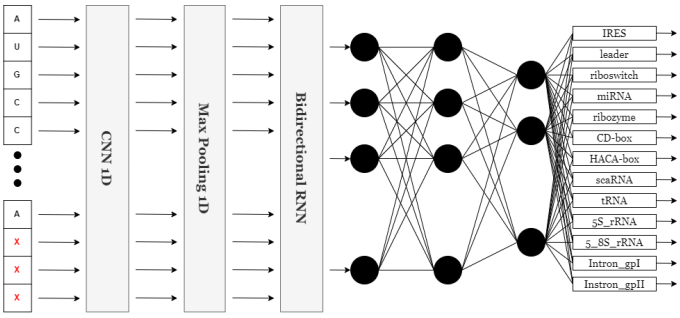

Fig. 2. NCC Neural Network Architecture

The convolutional layer scans for essential features in the input data, while the max pooling layer reduces the size of the feature map by downsampling. The bidirectional recurrent layer captures long-range dependencies within the input, and this is combined in the fully connected dense layer to output an ultimate prediction. The model reaches the highest accuracy with this neural network, while keeping an acceptable loss on the dataset in NCC. Since these experiments have shown its application across a wide variety of the classification of sequential data in NCC. The presented structure was implemented in Keras [23], representing a neural network capable of categorizing sequence information into 13 types by identifying 13 units and an 'softmax' activation function at its very end. The model then takes an input layer of a sequence with a dimensionality of either (500, 4) or (500, 8), with 500 being the length of the padded and truncated sequence, while 4 or 8 is the number of features depending on the number of features used in one-hot encoding. That translates to sequences of length 500 with 4 or 8 features each. This is followed by a one-dimensional convolutional layer, which uses 32 filters of size 9 with ReLU for the activation function to extract meaningful features from the sequence data. The max pooling layer reduces dimensionality by pooling over windows of size 4, therefore condensing the data while retaining key features and reducing computational complexity. Next, this is passed through a bidirectional GRU layer with 128 units to capture contextual dependency in both ways. Thus, it contains a dropout of 0.3 and initialization using randomized initializers for kernels and recurrent weights, but with zeros as biases.

The output from the GRU layer is then flattened into a single-dimensional format through a flattening layer, thereby preparing it for the fully connected dense layers. Finally, it ends with a dense layer composed of 13 units and the 'softmax' activation function, which gives the probability distribution across the 13 classes. This study highlights both accuracy and efficiency during training and testing. The streamlined architecture facilitates rapid processing, allowing models to be trained and tested efficiently without compromising accuracy. The balance of speed and precision is a key focus of this research.

C. Data Collection

Rfam database [25] has been used to compile small non-coding RNA sequences, which are stored in a plain text file in the standard FASTA format. Fasta is a widely accepted text-based format in bioinformatics for representing sequences from DNA or RNA and for representing amino acid sequences of proteins. The FASTA-file consists of sequence records with sequence definition lines and sequence data respectively. Each definition line begins with a greater-than symbol (">") followed by some unique identifier, often a descriptive name or code. The sequence data are single-letter codes representing individual nucleotides or amino acids. Because the focus of this chapter is classification, only class information is included in the definition line of each sequence.

For the creation of a dataset of small ncRNAs, the Rfam database [25] was primarily utilized. The representation of the sequences is the FASTA format, a well-known and acceptable format in the field of bioinformatics. Each record has a definition line that begins with a symbol greater than (">") accompanied by a distinct sequence marker – in our case the class of the RNA. In the next line, there is the primary structure of the molecule, i.e. the characters that represent the nucleotides. For better understanding, an illustrative set of examples is shown below.

```
>IRES
ATACCTTTCTCGGCCTTTTGGCTAAGATCAAGTGTAGTATCTGTTCTTAT...
>tRNA
GCACCACTCTGGCCTTTTGGCTTAGATCAAGTGTAGTATCTGTTCTTATT...
>tRNA
ATACCTTTCTCGGCCTTTTGGCTAAGATCAAGTGTAGTATCTGTTTTTAT...
>riboswitch
ATTACTTCTCAGCCTTTTGGCTAAGATCAAGTGTAATAAATCTCATTGTG...
>HACA-box
CCAGCTCTCTTTGCCTTTTGGCTTAGATCAAGTGTAGTATCTGTTCTTTT...
>tRNA
ACAGCTGATGCCGCAGCTACACTATGTATTAATCGGATTTTTGAACTTGG...
```

D. Final dataset (NCC dataset)

To create the final dataset, files belonging to the same family were merged to create a single .fasta file for each family. Due to the limited data regarding the IRES non-coding RNA family, a second data source was used, called IRESbase [26], which is dedicated to this family. In table III the source and count of RNAs of each family are presented.

TABLE III
SEQUENCES GROUPED BY FAMILY AND DATABASE

RNA Family	Source	# of Sequences
IRES	Rfam	1472
IRES	IRESbase	1328
leader	Rfam	31662
riboswitch	Rfam	69465
miRNA	Rfam	387173
ribozyme	Rfam	220007
CD-box	Rfam	132915
HACA-box	Rfam	36938
scaRNA	Rfam	2962
tRNA	Rfam	1432442
5S_rRNA	Rfam	140644
5_8S_rRNA	Rfam	4940
Intron_gpI	Rfam	2611
Intron_gpII	Rfam	15729

The final dataset has around 2,500–5,000 sequences per family, give or take a few. The distribution of data can be seen in Figure 3, while the number of sequences per class is presented in Table IV.

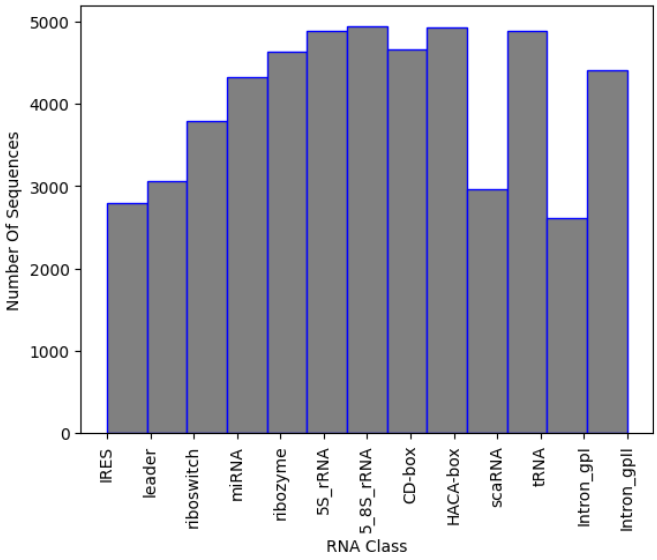

Fig. 3. Sequences per class in NCC dataset

TABLE IV
SEQUENCES PER FAMILY IN NCC DATASET

RNA Family	Number of Sequences
IRES	2800
leader	3061
riboswitch	3791
miRNA	4317
ribozyme	4630
CD-box	4661
HACA-box	4931
scaRNA	2962
tRNA	4882
5S_rRNA	4882
5_8S_rRNA	4940
Intron_gpI	2611
Intron_gpII	4409

Analysis of the RNA sequence length distribution, as depicted in Figure 4, reveals a correlation between the sequence

length and the class of RNA. This relationship can potentially improve the performance of the classifiers. For instance, as shown in Figure 5, miRNA sequences are typically shorter than 200 nucleotides, whereas ribozyme RNA sequences often exceed 800 nucleotides. These values indicate that incorporating the length into classification models could be beneficial to distinguish the classes. Understanding the variation in the lengths of the RNA sequences of different classes enhances our knowledge of their structure, while at the same time, contributing to the development of more accurate and efficient RNA classifiers.

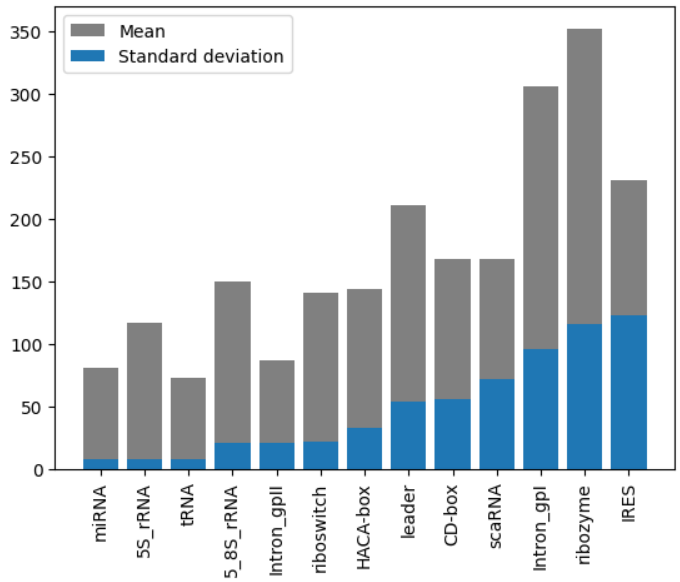

Fig. 4. Length mean and STD, grouped by class and sorted by STD of NCC dataset

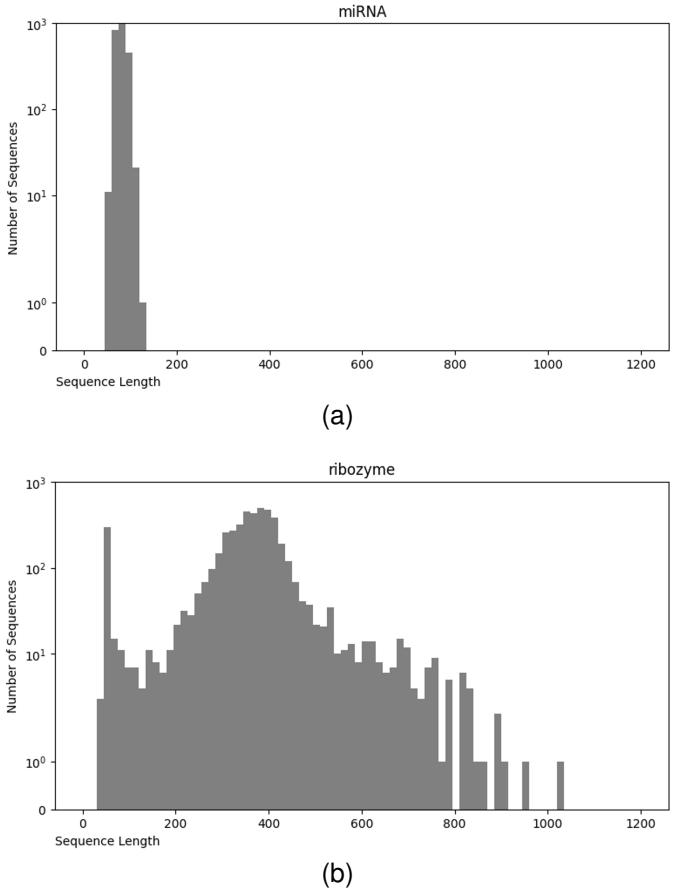

Fig. 5. Sequence length distribution of miRNA (a) and ribozyme (b)

E. nRC dataset

The dataset used in the nRC [18] was specifically curated for classifying non-coding RNA (ncRNA) sequences. This dataset was assembled using sequences from the Rfam database and includes 13 different ncRNA classes: miRNA, 5S rRNA, 5.8S rRNA, ribozymes, CD-box, HACA-box, scaRNA, tRNA, Intron gpI, Intron gpII, IRES, leader, and riboswitch. To ensure a balanced set of data and maintain statistical significance, a process used in previous works was applied. The CD-HIT tool [27] was utilized to select 20% non-redundant sequences, thereby avoiding redundancy that could bias the classification results. For most ncRNA classes, 500 sequences were randomly chosen, except for the IRES class, which only had 320 available sequences. This resulted in a total of 6320 ncRNA sequences in the dataset. In addition, a second dataset consisting of 2600 sequences, with 200 sequences from each class, was downloaded from Rfam for testing the nRC tool. Since this dataset is normally used for benchmarking against other classification architectures, it is also adopted for training and testing the proposed model to ensure an equitable comparison, as illustrated in Figure 6.

1) Comparative analysis of NCC and nRC: The NCC dataset used in this work is significantly larger, including more than an order of magnitude more sequences than the earlier dataset nRC. Indeed, the distribution of RNA classes is rather similar between the NCC dataset and the nRC dataset, with this resemblance being very apparent for some RNA classes, such as IRES sequences, ribozymes, and Intron_gpI, which are generally longer in both datasets, though minor variations exist, as depicted in Figures 7 and 8. However, a notable difference arises in the leader family, as highlighted in Figure 9, which provides a side-by-side comparison of the sequence length distribution for the leader RNA class in both sets.

F. Preparing and Benchmarking the Model

The Adam optimization algorithm [28] was employed with a learning rate of 0.001. Adam is a widely used method for training neural networks as it leverages the advantages of Stochastic Gradient Descent (SGD) and RMSProp, making it well-suited for handling sparse gradients in noisy environments. ”Adam,” short for ”Adaptive Moment Estimation,” leverages squared gradients to adapt the learning rate, similar to RMSProp, while also using the moving average of gradients, akin to SGD with momentum, rather than relying solely on the raw gradients. This synergy enables dynamic learning rate adjustment and gradient smoothing, facilitating efficient convergence toward the global minimum in the optimization

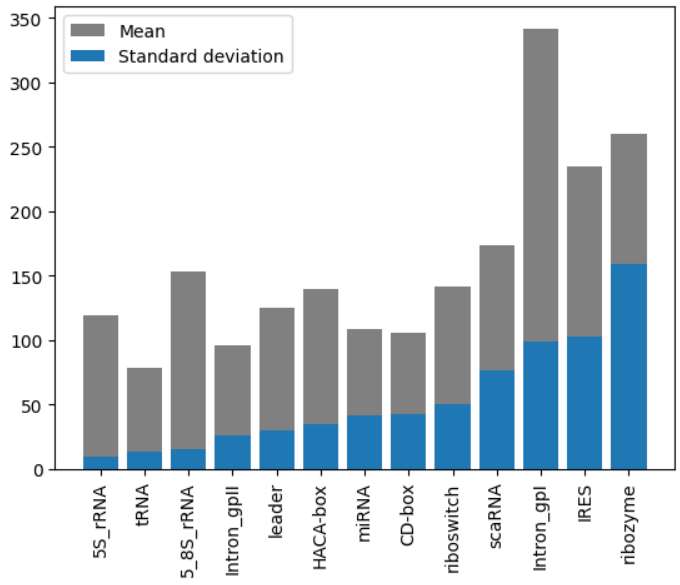

Fig. 6. Length mean and STD, grouped by class and sorted by STD of nRC dataset

process. Adam is particularly valued for its memory efficiency, computational cost-effectiveness, and specifically with large datasets and many parameters. Categorical Cross-Entropy is a loss function commonly used for multi-class classification problems in machine learning. It quantifies the difference between two probability distributions, one being the actual or true distribution of the real labels, and the other, a predicted distribution generated by the model. The loss calculates the negative logarithm of the probability related to the correct class. This method is especially effective when the model's output is a probability distribution across multiple classes, as seen in neural networks with a softmax activation function in the output layer. Categorical Cross-Entropy is essential for models requiring accurate probability distributions, particularly in multi-class classification. This loss function is calculated by the formula below:

$$L = -\sum_{i=1}^{M} y_i \log(p_i) \tag{1}$$

And for multi-class classification:

$$L = -\sum_{i=1}^{M} y_{o,i} \log(p_{o,i}) \tag{2}$$

Where:

- L is the loss.
- M is the number of classes.
- y_i is a binary indicator (0 or 1) if class label i is the correct classification for the observation.
- p_i is the predicted probability of the observation belonging to class i.
- $y_{o,i}$ is a binary indicator (0 or 1) if class i is the correct classification for observation o.
- $p_{o,i}$ is the predicted probability of observation o being of class i.

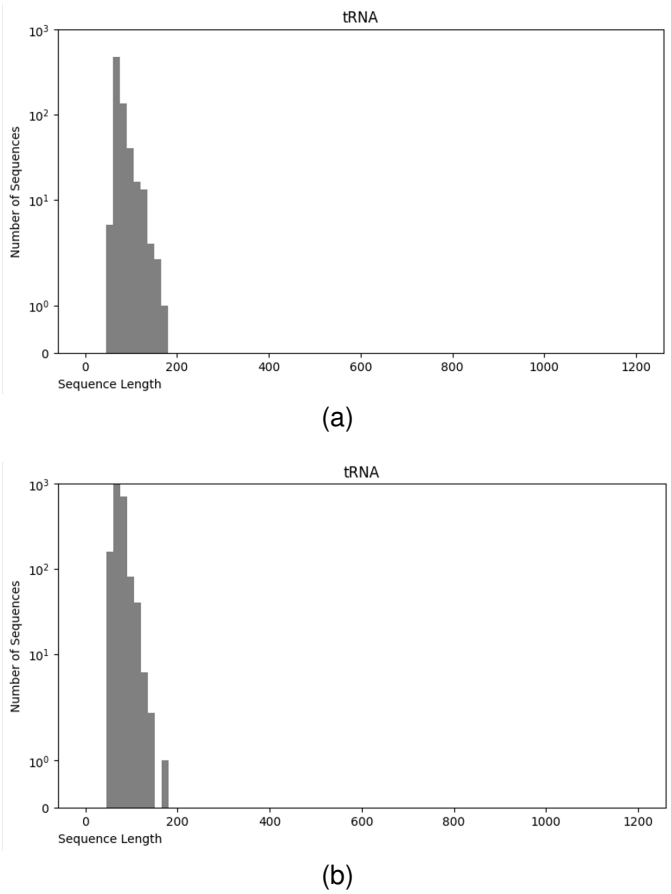

Fig. 7. Length distribution of leader RNA class in nRC (a) and NCC (b) datasets

To efficiently train the model, three different sets of epochs: 20, 50, and 100 were applied. Analyzing the data in Figure 10, it is obvious that after the first 20 epochs, the model's accuracy reached a plateau, exhibiting minimal to no additional improvement. This pattern is also mirrored in the loss function metrics. A detailed analysis of the training process is shown in Table V.

TABLE V
NCC MODEL TRAINING PARAMETERS

Number of Epochs	20/50/100
Batch size	128
Steps per Epoch	Default = 277
Optimizer	Adam
loss function	Categorical Cross Entropy
Shuffle Train data	True

1) Performance indicators: To evaluate the effectiveness of each prediction method for ncRNA classes, a set of well-established metrics is presented. Specifically, those of accuracy, sensitivity, precision, F1-score, Matthews correlation coefficient (MCC), and the confusion matrix were chosen for this task.

- **Accuracy** is the percentage of correctly classified cases.
- **Sensitivity**, also known as **Recall**, is the ratio of accurately anticipated positive cases to all actual positive cases.

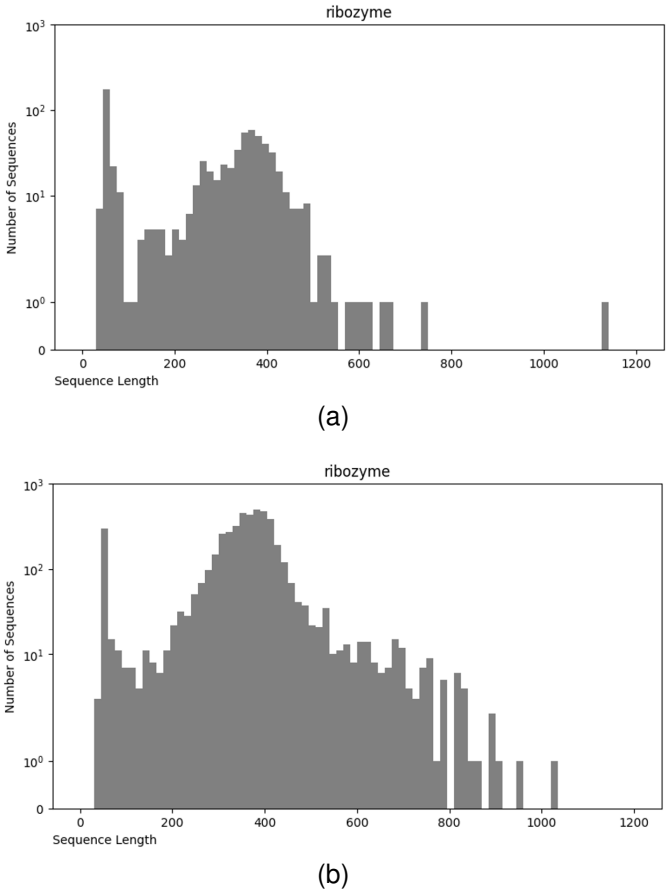

Fig. 8. Sequence length distribution of leader RNA class in nRC (a) and NCC (b) datasets

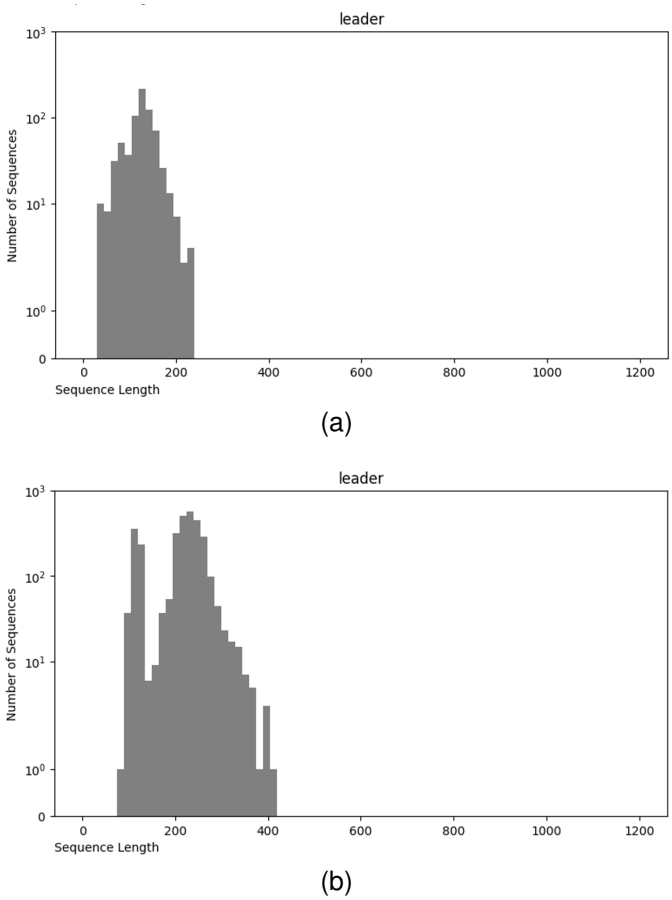

Fig. 9. Sequence length distribution of leader RNA class in nRC (a) and NCC (b) datasets

- **Precision** is the ratio of correctly predicted positive cases to the total predicted positive cases.
- The **F1-score** is the harmonic mean of a model's precision and recall.
- The **Matthews correlation coefficient (MCC)** is an effective metric for unbalanced classes, well-known for evaluating this task.

In multi-class classification, accuracy is calculated as a proportion of correctly classified instances in relation to all the instances in the data set. For other measures such as precision, recall, or F1-score, the scores for a metric are first obtained individually in each class and later averaged through macro-averaging. An example of precision computation can be seen in the formulas 3 and 4.

$$Precision_{ClassA} = \frac{TP_{ClassA}}{TP_{ClassA} + FP_{ClassA}} \tag{3}$$

$$Precision_{Macro-Average} = \frac{Precision_{ClassA} + ... + Precision_{ClassN}}{N} \tag{4}$$

In the multi-class, the Matthews correlation coefficient (MCC) can be defined in terms of a confusion matrix C for K classes. To simplify the definition, consider the following intermediate variables:

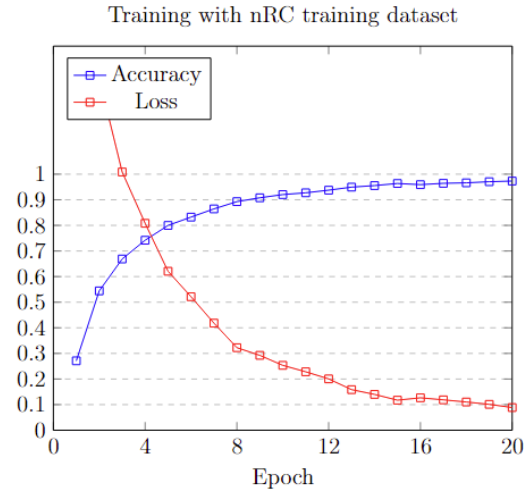

Fig. 10. Training accuracy and loss on each epoch

- t_k : the number of times class k truly occurred

$$t_k = \sum_{i}^{K} C_{ik} \tag{5}$$

- p_k : the number of times class k was predicted

$$p_k = \sum_{i}^{K} C_{ki} \tag{6}$$

- c : the total number of samples correctly predicted

$$c = \sum_{k}^{K} C_{kk} \tag{7}$$

- s : the total number of samples

$$s = \sum_{i}^{K} \sum_{j}^{K} C_{ij} \tag{8}$$

Then the multiclass MCC is defined as:

$$MCC = \frac{c \times s - \sum_{k}^{K} p_k \times t_k}{\sqrt{(s^2 - \sum_{k}^{K} p_k^2) \times (s^2 - \sum_{k}^{K} t_k^2)}} \tag{9}$$

It is required to remove any bias in the multi-task classification process, so each measure is computed for every class label independently. For a class C_k, any instance belonging to that class is marked as positive, while any instance belonging to any other class is marked as negative.

2) Results on NCC dataset: The initial step involves dividing the NCC dataset into training and testing sets. The model is trained using 35427 sequences with the parameters shown in Table V, while the remaining 17450 are used for testing purposes. Table VI shows the testing metrics using the NCC dataset with 4 and 8 digits RNA Base encoding. In both cases, the systems are very efficient, approaching almost perfect the ground truth values.

TABLE VI
RESULTS ON THE NCC DATASET FOR THE TWO PROPOSED VARIATIONS (4 AND 8 DIGITS)

One-hot Enc	Accuracy	Sensitivity	Precision	F-Score	MCC
8 Digits	**0.98922**	0.98651	0.98789	0.98718	0.98828
4 Digits	**0.99054**	0.98864	0.98886	0.98873	0.98971

3) Results on nRC dataset: To accurately assess the efficiency of the proposed NCC model, a comparison with state-of-the-art non-coding classifiers is presented. Training and testing the systems on the same dataset is crucial for a fair and precise comparison, ensuring the results are directly comparable without data-variation influences. In this section, a comparative analysis of key metrics for the state-of-the-art classifiers is presented. In particular, the NCC model has not been explicitly optimized for this dataset but was trained and fine-tuned using the NCC dataset created during this study to ensure an unbiased comparison. This approach allows for a thorough analysis of the NCC model's prediction capability relative to existing non-coding RNA classification methodologies.

TABLE VII
MODELS COMPARISON.

Model/Method	Accuracy	Sensitivity	Precision	F-score	MCC
RNAcon	0.3737	0.3787	0.4500	0.3605	0.3341
GraPPLE	0.6487	0.6684	0.7325	0.7050	0.6857
nRC	0.6960	0.6889	0.6878	0.6878	0.6627
ncRFP	0.7972	0.7878	0.7904	0.7883	0.7714
ncDLRES	0.8430	0.8344	0.8419	0.8407	0.8335
ncDENSE	0.8687	0.8677	0.8703	0.8667	0.8574
NCC 4d	**0.9269**	**0.9269**	**0.9286**	**0.9268**	**0.9210**
NCC 8d	**0.9292**	**0.9292**	**0.9311**	**0.9293**	**0.9234**

Table VII displays the efficiency of the examined systems in five metrics. The terms 4d and 8d refer to encoding with 4 digits and 8 digits per RNA nucleotide, respectively. From this data, it can be concluded that the NCC tool surpasses the other systems in all metrics. Specifically, it shows a slight improvement in accuracy, outperforming the next best model, ncDENSE, by approximately 0.06.

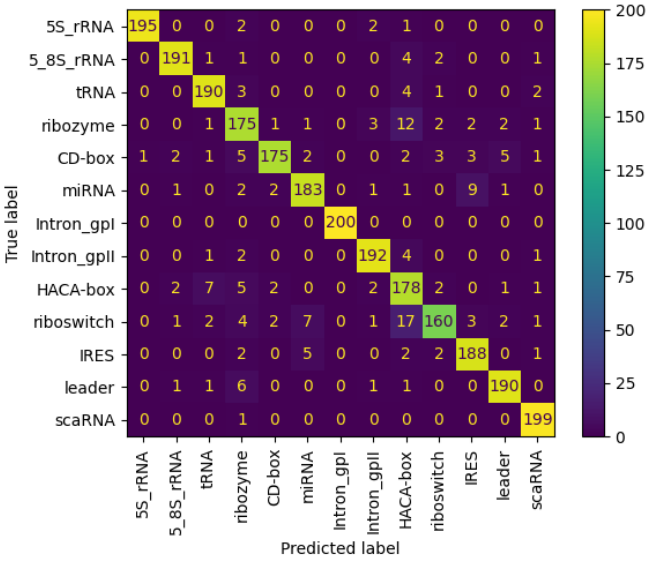

Fig. 11. Confusion Matrix of NCC model train and tested with nRC dataset

This demonstrates the NCC tool's superior capability in accurately classifying non-coding RNA sequences. Additionally, Figure 11 shows the confusion matrix for the nRC dataset, in regards to the proposed NCC model.

IV. CONCLUSIONS AND FUTURE WORK

This work focuses on the development and evaluation of a tool designed to classify ncRNAs, aiming to enhance our knowledge of the functions of RNA and the role of ncRNAs in various chemical processes. An illustrative aspect of the research is the creation of a new dataset, the NCC dataset, which is wider than the nRC, previously utilized, by ten times, providing a robust foundation for preparing a machine learning-based classifier. The research involves a comprehensive examination of the NCC and nRC sets, which reveals the inherent complexity of ncRNAs. A significant contribution of this research is the creation of a novel deep learning model, called NCC, specifically adjusted for the

classification of ncRNA. It uniquely encodes RNA sequences directly, thereby avoiding potential inaccuracies associated with RNA secondary structure prediction tools. The proposed model has shown the highest prediction capability, marginally surpassing state-of-the-art systems in accuracy, tested on the same dataset. In particular, the model achieved accuracy rates as high as 98% with the larger created dataset, indicating substantial improvements in this challenging classification task. The findings highlight the effectiveness of the model and the advances made, while also addressing the current limitations of RNA databases, noting their underdevelopment and significant imbalance among RNA classes, with specific classes being underrepresented. This imbalance underscores the ongoing need for more sophisticated and precise methodologies in the field of ncRNA classification. Active research and interest in this area suggest that future developments will likely bring about new methodologies that will further improve classification capabilities.

Although the NCC model did not take advantage of secondary structure and graph properties, their potential significance should not be overlooked. While these structures and their information were not utilized in the proposed model, they may provide essential knowledge and improve the model's prediction capability. Investigating their incorporation into future versions may reveal hidden patterns and other notable discoveries in the data, potentially resulting in more robust classifications. Given their importance within the wider scope of the field, further investigation and consideration of these factors are warranted.

REFERENCES

[1] Hüttenhofer, A. & Vogel, J. Experimental approaches to identify non-coding RNAs. *Nucleic Acids Research*. **34**, 635-646 (2006,1), https://doi.org/10.1093/nar/gkj469.

[2] Da-Fei, F. & F., D. Progressive sequence alignment as a prerequisitetto correct phylogenetic trees. *Journal Of Molecular Evolution*. **25**, 351-360 (1987), https://doi.org/10.1007/BF02603120.

[3] Mattick, J. & Makunin, I. Non-coding RNA. *Human Molecular Genetics*. **15**, R17-R29 (2006,4), https://doi.org/10.1093/hmg/ddl046.

[4] Griffiths-Jones, S., Bateman, A., Marshall, M., Khanna, A. & Eddy, S. Rfam: an RNA family database. *Nucleic Acids Research*. **31**, 439-441 (2003,1), https://doi.org/10.1093/nar/gkg006.

[5] RNAcentral Consortium RNAcentral 2021: secondary structure integration, improved sequence search and new member databases. *Nucleic Acids Research*. **49**, D212-D220 (2020,10), https://doi.org/10.1093/nar/gkaa921.

[6] Kalvari, I., Argasinska, J., Quinones-Olvera, N., Nawrocki, E., Rivas, E., Eddy, S., Bateman, A., Finn, R. & Petrov, A. Rfam 13.0: shifting to a genome-centric resource for non-coding RNA families. *Nucleic Acids Research*. **46**, D335-D342 (2017,11), https://doi.org/10.1093/nar/gkx1038.

[7] Andrikos, C., Makris, E., Kolaitis, A., Rassias, G., Pavlatos, C. & Tsanakas, P. Knotify: An Efficient Parallel Platform for RNA Pseudoknot Prediction Using Syntactic Pattern Recognition. *Methods And Protocols*. **5** (2022), https://www.mdpi.com/2409-9279/5/1/14.

[8] Makris, E., Kolaitis, A., Andrikos, C., Moulos, V., Tsanakas, P. & Pavlatos, C. Knotify+: Toward the Prediction of RNA H-Type Pseudoknots, Including Bulges and Internal Loops. *Biomolecules*. **13** (2023), https://www.mdpi.com/2218-273X/13/2/308.

[9] Koroulis, C., Makris, E., Kolaitis, A., Tsanakas, P. & Pavlatos, C. Syntactic Pattern Recognition for the Prediction of L-Type Pseudoknots in RNA. *Applied Sciences*. **13** (2023), https://www.mdpi.com/2076-3417/13/8/5168.

[10] Makris, E., Kolaitis, A., Andrikos, C., Moulos, V., Tsanakas, P. & Pavlatos, C. An Intelligent Grammar-Based Platform for RNA H-type Pseudoknot Prediction. *Artificial Intelligence Applications And Innovations. AIAI 2022 IFIP WG 12.5 International Workshops*. pp. 174-186 (2022).

[11] Jabbari, H., Wark, I., Montemagno, C. & Will, S. Knotty: efficient and accurate prediction of complex RNA pseudoknot structures. *Bioinformatics*. **34**, 3849-3856 (2018,6).

[12] Sato, K., Kato, Y., Hamada, M., Akutsu, T. & Asai, K. IPknot: fast and accurate prediction of RNA secondary structures with pseudoknots using integer programming. *Bioinformatics*. **27**, i85-i93 (2011,6), https://doi.org/10.1093/bioinformatics/btr215.

[13] Panwar, B., Arora, A. & Raghava, G. Prediction and classification of ncRNAs using structural information. *BMC Genomics*. **15** (2014), https://doi.org/10.1186/1471-2164-15-127.

[14] Nithin, C., Mukherjee, S., Basak, J. & Bahadur, R. NCodR: A multi-class support vector machine classification to distinguish non-coding RNAs in Viridiplantae. *Quantitative Plant Biology*. **3** pp. e23 (2022).

[15] Dunkel, H., Wehrmann, H., Jensen, L., Kuss, A. & Simm, S. MncR: Late Integration Machine Learning Model for Classification of ncRNA Classes Using Sequence and Structural Encoding. *International Journal Of Molecular Sciences*. **24** (2023), https://www.mdpi.com/1422-0067/24/10/8884.

[16] Childs, L., Nikoloski, Z., May, P. & Walther, D. Identification and classification of ncRNA molecules using graph properties. *Nucleic Acids Research*. **37**, e66-e66 (2009,4), https://doi.org/10.1093/nar/gkp206.

[17] Wang, L., Zheng, S., Zhang, H., Qiu, Z., Zhong, X., Liu, H. & Liu, Y. ncRFP: A Novel end-to-end Method for Non-Coding RNAs Family Prediction Based on Deep Learning. *IEEE/ACM Transactions On Computational Biology And Bioinformatics*. **18**, 784-789 (2021).

[18] Fiannaca, A., La Rosa, M., La Paglia, L., Rizzo, R. & Urso, A. nRC: non-coding RNA Classifier based on structural features. *BioData Mining*. **10** (2017), https://doi.org/10.1186/s13040-017-0148-2.

[19] Wang, L., Zhong, X., Wang, S. & Liu, Y. ncDLRES: a novel method for non-coding RNAs family prediction based on dynamic LSTM and ResNet. *BMC Bioinformatics*. **22** (2021,9), https://doi.org/10.1186/s12859-021-04365-4.

[20] Chen, K., Zhu, X., Hao, L., Wang, J., Liu, Z. & Liu, Y. ncDENSE: a novel computational method based on a deep learning framework for non-coding RNAs family prediction. *BMC Genomics*. (2022,12)

[21] Anderson P. Avila Santos & Carvalho, A. BioDeepfuse: a hybrid deep learning approach with integrated feature extraction techniques for enhanced non-coding RNA classification. *RNA Biology*. **21**, 1-12 (2024).

[22] Zhao, Q., Zhao, Z., Fan, X., Yuan, Z., Mao, Q. & Yao, Y. Review of machine learning methods for RNA secondary structure prediction. *PLoS Computational Biology*. **17** pp. e1009291 (2021,8).

[23] Chollet, F. & Others Keras. (https://keras.io,2015).

[24] Schuster, M. & Paliwal, K. Bidirectional recurrent neural networks. *Signal Processing, IEEE Transactions On*. **45** pp. 2673 - 2681 (1997,12).

[25] Griffiths-Jones, S., Bateman, A., Marshall, M., Khanna, A. & Eddy, S. Rfam: an RNA family database. *Nucleic Acids Research*. **31**, 439-441 (2003,1), https://doi.org/10.1093/nar/gkg006.

[26] Zhao, J., Li, Y., Wang, C., Zhang, H., Zhang, H., Jiang, B., Guo, X. & Song, X. IRESbase: A Comprehensive Database of Experimentally Validated Internal Ribosome Entry Sites. *Genomics, Proteomics & Bioinformatics*. **18**, 129-139 (2020), https://www.sciencedirect.com/science/article/pii/S1672022920300577, Special Issue:Bioinformatics Commons—2020.

[27] Fu, L., Niu, B., Zhu, Z., Wu, S. & Li, W. CD-HIT: accelerated for clustering the next-generation sequencing data. *Bioinformatics*. **28**, 3150-3152 (2012,10), https://doi.org/10.1093/bioinformatics/bts565.

[28] Kingma, D. & Ba, J. Adam: A Method for Stochastic Optimization. *International Conference On Learning Representations*. (2014,12).