Received September 26, 2020, accepted November 1, 2020, date of publication November 16, 2020,
date of current version February 2, 2021.

Digital Object Identifier 10.1109/ACCESS.2020.3037642

A Robust and Precise ConvNet for Small Non-Coding RNA Classification (RPC-snRC)

MUHAMMAD NABEEL ASIM[1,2], MUHAMMAD IMRAN MALIK[3], CHRISTOPH ZEHE[4], JOHAN TRYGG[5,6], ANDREAS DENGEL[1,2], AND SHERAZ AHMED[1]

[1]German Research Center for Artificial Intelligence (DFKI), 67663 Kaiserslautern, Germany
[2]Department of Computer Science, TU Kaiserslautern, 67663 Kaiserslautern, Germany
[3]National Center for Artificial Intelligence (NCAI), National University of Sciences and Technology, Islamabad 44000, Pakistan
[4]Sartorius Corporate Research, Sartorius Stedim Cellca GmbH, 89081 Ulm, Germany
[5]Computational Life Science Cluster (CLiC), Umeå University, 901 87 Umeå, Sweden
[6]Sartorius Corporate Research, Sartorius Stedim Data Analytics, 90333 Umeå, Sweden

Corresponding author: Muhammad Nabeel Asim (muhammad_nabeel.asim@dfki.de)

This work was supported by the Sartorius Artificial Intelligence Laboratory (SAIL).

ABSTRACT Small non-coding RNAs (ncRNAs) are attracting increasing attention as they are now considered potentially valuable resources in the development of new drugs intended to cure several human diseases. A prerequisite for the development of drugs targeting ncRNAs or the related pathways is the identification and correct classification of such ncRNAs. State-of-the-art small ncRNA classification methodologies use secondary structural features as input. However, such feature extraction approaches only take global characteristics into account and completely ignore co-relative effects of local structures. Furthermore, secondary structure based approaches incorporate high dimensional feature space which is computationally expensive. The present paper proposes a novel Robust and Precise ConvNet (RPC-snRC) methodology which classifies small ncRNAs into relevant families by utilizing their primary sequence. RPC-snRC methodology learns hierarchical representation of features by utilizing positioning and information on the occurrence of nucleotides. To avoid exploding and vanishing gradient problems, we use an approach similar to DenseNet in which gradient can flow straight from subsequent layers to previous layers. In order to assess the effectiveness of deeper architectures for small ncRNA classification, we also adapted two ResNet architectures having a different number of layers. Experimental results on a benchmark small ncRNA dataset show that the proposed methodology does not only outperform existing small ncRNA classification approaches with a significant performance margin of 10% but it also gives better results than adapted ResNet architectures. To reproduce the results Source code and data set is available at https://github.com/muas16/small-non-coding-RNA-classification

INDEX TERMS RNA sequence analysis, small non-coding RNA classification, DenseNet, ResNet.

I. INTRODUCTION

Besides serving as coding template in the expression of proteins, RNA has a plethora of additional biological functions and plays a key role in several diseases such as Alzheimer, cardiovascular, Cancer, and type 2 diabetes [1], [2]. RNA can be classified into protein coding or non-coding, where about 3% of total RNA is coding for proteins (so called messenger RNA = mRNA) and the remaining 97% known as non-coding (ncRNA) or functional RNA [3]. While the function of mRNAs is well known and has been studied extensively, ncRNAs were considered useless for quite some time. However, with the progress in biological research, it was discovered that the majority ncRNAs are involved in many essential biological processes such as dosage compensation, genomic imprinting, and cell differentiation [4], [5]. Over time, the analysis of ncRNAs has become even more interesting because of their importance in understanding the phenomena behind human health and disease [4].

ncRNAs differ from each other in terms of length, conformation, and biological function. As shown in Figure 1, ncRNAs are typically classified into small non-coding RNAs (sncRNA) and long non-coding RNAs (lncRNA). The lncRNAs are larger than 200 bp in size [5] and are further

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

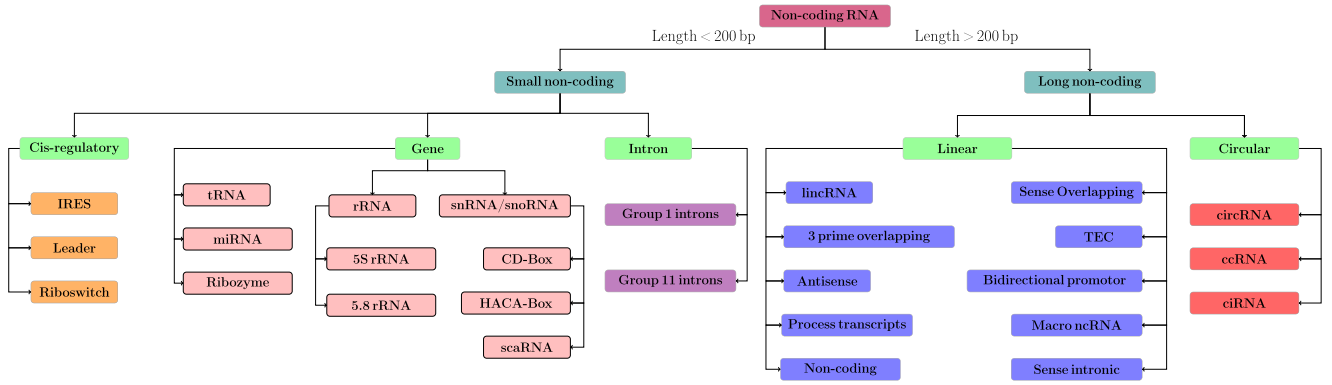


FIGURE 1. Overall taxonomy of non-coding RNA families, adapted from [4]. Where the magenta colored square at top layer represents non-coding RNA, dark green squares at second layer namely small non coding and long non coding refer to the subclasses of non-coding RNA. Similarly, light green squares are the major subclasses of small, and large non-coding RNA. At last level, yellow squares show the types of Cis-regulatory sequences, light pink show the kinds of Gene sequences, purple shows the subtypes of Intron sequences. On the other hand, navy blue reveals the kinds of Linear long non-coding RNA sequences, and Circular subclasses are shown by Red squares.

divided into linear RNAs and circular RNAs. Linear RNAs have been demonstrated to play a role in gene transcription and translation [6]. Circular RNAs are involved in gene regulation and strongly connected with complex human diseases like lung cancer and are considered important in identification and treatment of tumors [7], [8]. Classification of RNA through manual experimentation is time consuming and expensive [9]. Extensive research has been done to differentiate protein coding RNA from non-coding RNA. Especially, in order to discriminate long non-coding RNA (lncRNA) from mRNA or to classify long non-coding RNAs (lncRNA) into their corresponding families, diverse machine and deep learning based methodologies have been proposed [6], [7], [10]–[13].

Small ncRNAs have a length of around 20-30 bp and are involved in translation, splicing, and regulation of genes [14]. Primarily, small ncRNAs are classified into 13 subclasses where each subclass has distinct medical and biological significance. For instance, scaRNAs, most of which are functionally and structurally identical to snoRNAs can guide modifications in pseudo uridylation and methylation. miRNAs are involved in post transcriptional gene expression regulation and RNA silencing. They target almost 60% of all human genes and play an important role in several biological processes like cell differentiation, proliferation, and death [15]–[18]. Studies have demonstrated that miRNAs are also involved in diverse and complex human diseases such as cancer, autoimmune, cardiovascular, and neurodegenerative diseases [19]. Similarly, Ribosomal RNA (rRNA) plays an essential role in protein synthesis and its characteristics are considered very valuable for the development of antibiotics. 5.8S ribosomal RNA actively participates in protein translocation [20], forms covalent connection with tumour suppressor proteins [21], can be used to detect miRNA [22] and to understand other rRNA pathways and processes in the cell [23]. Although the function is 5 S ribosomal RNA has not been discovered yet, it has been shown that its deletion substantially reduces protein synthesis and has detrimental effects on cell fitness [24].

Classification of small non-coding RNAs (sncRNAs) is of high importance because of their large number and diverse functions. It can support biologists and clinicians to get a better understanding for the role of sncRNAs in biological processes and the development of various diseases. For example, classification of sncRNAs is important in developing strategies for cancer therapeutics [4]. To the best of our knowledge, two computer based (in particular, deep learning based) approaches have shown the best classification results for small non-coding RNA to date. The first approach,"nRC", proposed by Fiannaca *et al.* [25] comprises three fundamental tasks including estimation of secondary structures from Rfam dataset (publicly available benchmark dataset containing 8920 samples belonging to 13 sncRNA subclasses), extraction of common substructures, and classification into 13 known ncRNA classes using a convolutional neural network (CNN). This approach achieved 81% ncRNA classification accuracy.

The second approach, proposed by Rossi *et al.* [26] extracts secondary structural features from the same Rfam database. However, rather than using a simple convolutional neural network, the tool utilizes Graph based convolutional architecture for the extraction of discriminative features and classification. This approach is setting the standard with 85% accurate classification of small non-coding RNA sequences.

As described above, both state-of-the-art small non-coding RNA classification approaches use secondary structure of RNA sequences as input and extract discriminative features by utilizing convolution layers or graph based methodologies. However, feature extraction methods based on secondary structures usually only consider the global characteristics while ignoring the mutual influence of local structures [27]. Such methods are usually neglecting important

information that might have been available in the primary sequences and was potentially lost while developing the secondary structures, on which the final classification is based. Furthermore, secondary structure based methods integrate high-dimensional feature space which is computationally inefficient [27].

In the present paper, we propose to use primary RNA sequences directly, instead of extracting discriminative features from secondary structures. We present a Robust and Precise Convolutional neural network for a small noncoding RNA Classification (RPC-snRC) system. The proposed system is based on an end to end small non-coding RNA classification methodology which uses a set of deep convolutional layers for the extraction of discriminative features by utilizing positioning and information on the occurrence of various nucleotides in non-coding RNA sequences. To evaluate the integrity of the proposed methodology, we performed experiments using the publicly available benchmark dataset provided by Fiannaca *et al.* [25]. The proposed system clearly outperforms state-of-the-art method (by Rossi *et al.* [26]) by a fair 10% margin in terms of different performance metrics including accuracy, precision, recall and $F1$-measure. In addition, extensive experimentation is performed with different sequence k-mers (1-mer, 3-mers) and representation schemes including one hot vector, random embedding initialization, and pre-trained prot2vec embeddings. From this it could be concluded weather deep architectures performs better at atom level or word level and which kind of feature representation is better for discriminative feature extraction. Moreover, to further analyze the idea of utilizing primary RNA sequences, we performed experiments with two adapted deep ResNet architectures which vary in terms of hyperparameters. Both of these architectures also outperformed state-of-the-art deep learning approaches thereby validating the idea of utilizing the primary RNA sequence for classification.

II. RELATED WORK

Non coding Ribonucleic Acid (ncRNA) has been classified into a range of distinct classes or families which vary in function and composition. The interest to develop sophisticated methods for ncRNA classification has rocketed over the period since knowing the family of ncRNA is substantial for drug targeting and understanding growth of various complex diseases. Non-coding RNA classification is a vast domain where classification at different levels of ncRNA(shown in figure 1) has been performed. Mainly, researchers have been focusing to 1) distinguish non-coding RNA from coding RNA, 2) categorize ncRNA into long and small non-coding RNA, 3) segregate non-coding RNA into its subtypes such as circular RNA, and to 4) classify small non-coding RNA into its 13 subclasses. Classification at each level facilitates distinct biological advantages. In order to discover more classes of non-coding RNA, researchers have also developed clustering-based computational methodologies. Although the main focus of this article is small non-coding RNA classification, however considering the importance of ncRNA classification, this section provides an overview of state-of-the-art non-coding RNA classification approaches at diverse levels. It also sheds light on clustering-based approaches for non-coding RNA identification

In the last decade, researchers were more inclined towards the development of computational methodologies which can discriminate between non-coding RNA and coding RNA. Washietl *et al.* [28] proposed a method, namely RNAz, based on Support Vector Machines (SVM) to classify ncRNAs. The RNAz combined sequence analysis approach with structure prediction. Primarily two components consensus secondary structure and thermodynamic stability were used. RNAz also integrated multifold sequence alignment and pairwise alignment of ncRNA sequence with extremely high sensitivity and specificity. They utilized RFAM genomic database [29] containing ncRNAs of humans, mice, zebrafish, and rats. RNAz exploited RNA folding of least free energy and computed z-scores by performing regression through SVM. Input parameters of proposed approach were number of alignment sequences, structure conservation index (SCI), and the mean of MFE z-score [30] of diverse sequences present in alignment excluding gaps. It also utilized the functionality of program namely RNAALIFOLD [31] which was primarily developed to estimate secondary structure from aligned sequence. RNAz used a folding algorithm to predict the secondary structure of RNA's through implementing dynamic, and robust programming algorithms. They reported that when SCI was almost zero, it indicated that consensus structure was not found by the RNAALIFOLD, contrarily perfect conserved structures had the SCI of almost 1. RNAz produced decent results for genomic annotation performed at large scale. Likewise, Liu *et al.* [32] presented a method based on SVM namely Coding or non-coding (CONC) to classify ncRNAs. It integrated multiple sequence alignment and used the databases FANTOM3 [33], NONCODE [34], and RNAdb [35] for experimentation. This method utilized composition of amino acid, exposed residues estimated percentage, peptide length, compositional entropy, found homologs from mentioned databases searches, alignment entropy, and estimated content of secondary structure.

In order to raise the performance of ncRNA classification further, few researchers explored ensemble approaches considering the effectiveness of decision trees. For instance, Lertampaiporn *et al.* [36] came up with a hybrid tool for the task of ncRNAs classification. They combined an ensemble of several decision trees and random forest with logstic regression model to discriminate short, and long ncRNA sequences. This tool includes naive feature SCORE which was computed by logistic regression through the combination of five features, i.e., structure, robustness, sequence, modularity, and coding potential. For experimentation, it used multiple datasets including, RefSeq [37], Rfam [29], lncRNAdb [38], and genome database "GenBank" of NCBI. In the proposed methodology, a set of 369 features were

extracted to predict ncRNAs. Amongst these features, discriminative features were acquired through feature selection based on correlation and genetic algorithm. While logistic regression was utilized to locate relationships among features, sequence similarity was facilitated by fundamental local alignment finder (BLAST) [39]. Random forest acted as primary classifier. Ensemble of several decision trees in random forest was capable to acquire heterogeneity of ncRNA subfamilies. This methodology was robust as it exploited composite features which raised the classifier performance. This approach was used to classify known ncRNAs, and also unknown ncRNAs. Similarly, Achawanantakun *et al.* [40] presented a method namely lncRNA-ID based on balanced decision trees to identify long ncRNAs. This method utilized multiple sequence alignment and LncRNADisease database [41] for experimentation.

Furthermore, researchers also experimented with unsupervised methodologies for ncRNAs identification. For example, Saito *et al.* [42] presented a methdology, namely EnsembleClust, for hierarchical clustering of ncRNAs. This methodology enabled the discovery of new ncRNA families [42] and aided to investigate functional diversity of ncRNAs. EnsembleClust implemented an unsupervised approach which utilized unlabelled data to construct clusters of ncRNAs on the basis of structural alignment results. As the computation of structural alignment was extremely expensive, approximate algorithms were utilized which considered all possible secondary structures and sequence alignments. In addition, for the sake of accurate clustering, a robust measure was used which considered primary sequences, and secondary structures. EnsembleClust produced better performance when compared with previous approaches such as FOLDALIGN [43], Stem kernel [44], and LocARNA [45]. Moreover, Miladi *et al.* [46] came up with an approach, RNAscClust, to identify ncRNAs. RNAscClust was used to combine RNA sequences through structure conservation, and graph oriented motifs [46]. This approach used structural similarities in order to group paralogous RNAs. RNAscClust enbaled clustering of humongous occurrences. Sequences were transformed into a graph, where every nucleotide was taken as graph vertices represented with the labels A, U, G, C in form of base pair connections, and the edges were representing encoded backbone. The structures were compared with one another through graph kernels. This method considered the changes of base pairs-which were never encountered by previous clustering approaches. For experimentation, Rfam database having ncRNA sequences was used. Authors reported that the proposed method managed to facilitate accurate clustering which made it possible to align large clusters efficiently.

Considering the promising performance of deep neural network for diverse natural language processing tasks, researchers employed Convolutional Neural Networks (CNNs) to classify ncRNAs. For example, Aoki and Sakakibara [47] proposed a methodology CNNClust to make the clusters of ncRNAs. This technique integrated pair wise alignment of ncRNA sequences. CNN was trained using positional weigh matrices of underlying sequence motifs. Two kinds of neural word embeddings, one hot encoding and word2vec, were used by CNNClust. Information of secondary structures and read mapping were also utilized in CNNClust. Matrix of similarity score was computed for each pair of RNA sequences and clustering was performed to group highly similar structures. CNNClust categorizes ncRNA into either positive or negative class. When both ncRNA sequences belong to the same class then it was classified as positive otherwise negative. Several new kinds of ncRNA such as microRNA, tRNA, and snoRNA were discovered through this approach. For experimentation, authors used Rfam, HUGO gene nomenclature committee (HGNC), and Genomic tRNA (GtRNAdb) datbabses. Similarly, Fiannaca *et al.* [25] presented an approach, nRC, for classification of ncRNAs. This approach used features of secondary structures and incorporated alignment of multiple sequences to categorize 13 known ncRNA classes using a CNN. The nRC utilizes IPKnot43 which is capable of predicting secondary structures and generate an accurate graph based on multifarious topologies of non-coding RNA sequences. IPKnot43 yields a graph database as it generates an undirected label graph for every input transcript. Considering the ideology that graphs having similar substructure usually belong to same RNA family, common subgraph extraction is exploited with a minimum threshold to locate frequent substructures which represent features of diverse small non-coding RNA subclasses. In order to locate subgraphs, nRC utilizes Molecular Substructure Miner (MoSS) which produces common subgraphs using a depth-first search. In this way, nRC only considers close common subgraphs. Lastly, nRC leverages the power of a CNN containing two convolutional and two fully connected layers.

Likewise, few researchers experimented with Recurrent Neural Networks (RNN) to classify ncRNAs. Baek *et al.* [11] presented a methodology namely lncRNAnet for the ncRNA classification. LncRNAnet identified long ncRNA through next generation sequencing [11] and deep learning. They used both CNN and RNN. While RNN was exploited to model RNA sequences, CNN was used to spot stop condons in order to locate an idicator of open reading frame. LncRNAnet showed decent performance while classifying short length RNA sequences. It learned intrinsic features through RNN for modelling RNA sequences. Authors performed experimentation on GENCODE, ENSEMBL and Human and Vertebrate Analysis and Annotation (HAVANA) databases. They reported that the proposed methodology produced robust performance regardless of variable sequence length, and helped to identify latest lncRNA from large transcriptome data. Moreover, Park *et al.* [43] proposed a methodology based on deep RNN for the task of ncRNA classification. The proposed method utilized the features of secondary structures to identify ncRNA and incorporated pairwise alingment of sequences. Authors used fRNAdb, NON-CODE, and NCBI datasets for extensive experimentation.

TABLE 1. Summary of the previous work for non-coding RNA classification and clustering in terms of exploited technique, alignment of sequences information, and type of features used as an input. In Table three methodologies namely Ensemble clust, RNAscClust, SHARAKU and CNN clust makes groups of non coding RNAs according to their structural similarities. LNCRNAid and LncRNANet performs long non coding RNA identification. Two machine learning based methodologies namely RNAZ, and CONC differentiates between coding and non coding RNA sequences. Hybrid random forest methodology is used to discriminate between small non coding and long non coding RNA sequences. Deep RNN methodology identifies microRNAs.Two deep learning methodologies namely nRC and RNAGCN methodologies performs classification of small non coding RNA.

Method	Database	Alignment	Features	Technique	Categorization type
RNAz [28]	Rfam	Pairwise and multiple sequence alignment	Thermodynamic stability measure, consensus secondary structure	SVM	Classification into coding or non coding RNA
CONC [32]	RNAdb, NONCODE, FANTOM	multiple sequence alignment	Amino acid composition, peptide length, predicted secondary structure content, predicted percentage of exposed residues, compositional entropy, number of homologs from database searches and alignment entropy	SVM	Classification into coding or non coding RNA
Hybrid random forest [36]	Rfam, RefSeq, NCBI GenBank genome database and lncRNAdb database	multiple sequence alignment	sequence, structure, structural robustness, modularity and coding potential	Random forest	Classification into small non coding or long non coding RNA
Deep RNN [48]	NCBI, fRNAdb, NON-CODE	pairwise sequence alignment	Secondary sequence features	RNN	Micro RNAs Identification
lncRNAID [11]	LncRNADisease database	profile hidden Markov model (profile HMM)-based alignment	open reading frame (ORF), protein conservation and ribosome interaction	Random forest	Long non coding RNA identification
lncRNAnet [11]	GENCODE, ENSEMBL and Human and Vertebrate Analysis and Annotation group databases	multiple sequence alignment	Open reading frame (ORF) indicator	RNN	Long non coding RNA identification
EnsembleClust [42]	ENSEMBLE	Pairwise sequence alignment	structural alignments score	Hierarchical Clustering	Clustering of non coding RNA
RNAscCLust [46]	Rfam	-	structure conservation and graph-based motifs	Hierarchical Clustering	Clustering of non coding RNA
SHARAKU [49]	NCBI Reference sequence database, ENSEMBLE database and next generation sequencing output	Pairwise sequence alignment	Similarity score matrix	Random forest	Clustering of non coding RNA
CNNClust [47]	Rfam, HUGO gene nomenclature committee (HGNC) databases, Ensembl and genomic tRNA database	Pairwise sequence alignment	Derived position weight matrices of sequence motifs	CNN	Clustering of non coding RNA
Deep next generation sequencing [50]	NONCODE, NCBI, lncRNA	Pairwise sequence alignment	Protein coding features	Deep next generation sequencing	Classification into coding or non coding RNA
circ-Deep [7]	CircRNADb	-	RCM features, conservation features	CNN and LSTM	Long non coding Circular RNA classification
nRC [51]	Rfam	Multiple sequence alignment	Secondary structure features	CNN	Classification of small non coding RNA
RNAGCN [26]	Rfam	Multiple sequence alignment	Secondary structure features	graph convolutional network	Classification of small non coding RNA

Contrarily deep sequencing has also been employed for ncRNA classification. For instance, Tsuchiya *et al.* [49] presented an approach SHARAKU based on deep sequencing for ncRNA classification. SHARAKU incorporated an algorithm which aligned read mapping profiles of ncRNAs next generation data containing sequences. This system also implemented a program for the alignment of read mapping profile which used decomposition for the sake of folding and aligning RNA sequences at the same time [52]. Profiles of read mapping allowed the detection of common patterns. Secondary structure and sequence information were acquired concurrently in this approach. The proposed approach helped to locate ncRNAs specifically combined in brain. The authors used NCBI, ENSEMBLE, and next generation sequencing output databases as reference. SHARAKU managed to achieve better performance than deepBlockAlign [53]. Likewise, Weikard *et al.* [50] presented a method based on next generation deep sequencing for the task of classifying ncRNA. The proposed method utilized features of protein coding to discriminate among coding and non-coding RNAs. It incorporated alignment of pairwise sequences and used lncRNA, NONCODE, NCBI datasbases for experimentation.

In order to improve the performance of small non coding RNA classification, more recently, Rossi *et al.* [26] proposed Graph convolutional neural network based methodology which also takes secondary structural features as input. This methodology uses Graph convolutions for the extraction of discriminative features from the secondary structural features. According to our best knowledge, this is the latest methodology which has excluded manual feature engineering and produced state-of-the-art performance for small non-coding RNA classification. Table 1 summarizes the state-of-the-art work for RNA sequence classification.

In this article, we proposed a RPC-snRC methodology which takes input RNA sequence data and utilises convolutional layers for the extraction of discriminative features which are eventually passed to dense layers for classification. Note that our methodology does not require any alignment or manual feature extraction technique as it provides an end to end deep learning system which takes RNA sequences as input and provides class label as output.

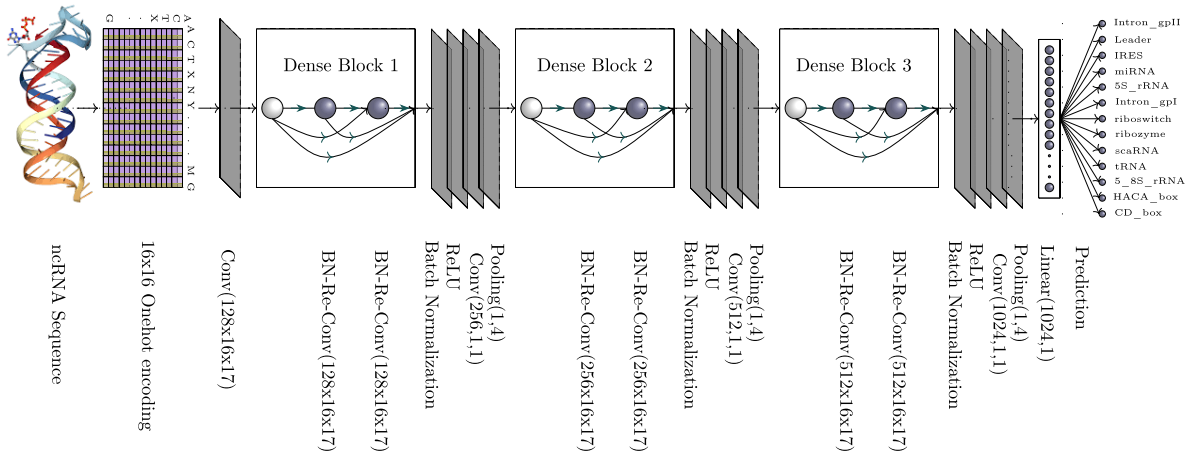


FIGURE 2. Proposed RPC-snRC methodology for small non-coding RNA classification. In figure, (128,16,18) indicates there are 128 kernels, each of width 16 and length 18 in a convolutional layer and (1,4) indicates kernel width and length are set to 1 and 4 respectively in a pooling layer. Others have the similar meaning.

III. MATERIALS AND METHODS

Following the success of deep learning methodologies in the ImageNet Large Scale Visual Recognition Challenge (ILSVRC),[1] researchers became more interested to employ deep learning for diverse computer vision, natural language processing, and bioinformatics tasks [54]–[58]. Generally, the aim was to develop deeper architectures with proper gradient flow among the layers which could learn better hierarchical representation of features.

Deoxyribonucleic acid (DNA) and ribonucleic acid (RNA) sequences are often treated in the same way as traditional text in natural language processing [59]. A term K-mers is used for DNA and RNA sequences where a group of three or four nucleotides are combined to form a word known as 3-mers or 4-mers. However, there is a debate about which atom-level (single nucleotide known as character or K-mers known as word) would be the most effective representation for DNA and RNA sequence analysis tasks. Furthermore, researchers are also working on proteomic and genomic data to provide biomedical pretrained neural word embeddings for different k-mers. This is because neural word embeddings, known as continuous representations of features or words, have played an important role to improve the performance of various NLP tasks. In this regard, Asgari and Mofrad [60] recently provided pretrained neural word embeddings for proteins and genes. However, there are again several open questions about the impact of utilizing pretrained neural word embeddings for DNA, and RNA sequence analysis, e,g., will deep architectures learn better features using pretrained word embeddings of proteins and genes?

The present paper presents a robust and precise convnet based system for small non-coding RNA classification. The proposed system takes direct RNA sequence data as input and utilizes convolutional layers for the extraction of discriminative features which are eventually passed to dense layers for classification. This system does not require any alignment or manual feature extraction as it provides an end to end deep learning based system which takes primary RNA sequences as input and provides class labels as output. Furthermore, to provide answers to the questions above, we have performed detailed experimentation on small non-coding RNA classification dataset with the proposed system and also with two adapted Res-Net architectures. Parameter details of the adapted Res-Net architectures are summarized in Table 2.

A. PROPOSED METHODOLOGY

This section briefly describes the proposed methodology of RPC-snRC for classification of small non-coding RNA. We develop a deep classifier in which a phenomenon similar to DenseNet is used to enable proper flow of gradient between the layers. RPC-snRC utilizes a set of convolutional layers for extraction of discriminative features from the primary sequences of small non-coding RNA. Discriminative features are then fed to dense layers for classification of sequences into a set of predefined classes.

Figure 2 illustrates the architecture of the proposed methodology along with noteworthy model parameters. The proposed RPC-snRC methodology is based on three dense modules. Each dense module contains the same number of layers; however, output units get doubled in every following dense module. Each dense module first performs batch normalization on the given input and then applies ReLu activation to introduce non-linearity followed by convolution operation to extract discriminative features. Finally, it repeats the discussed operations one more time in order to better learn hierarchical representation of data. Each dense module is followed by a transition layer which performs batch normalization, ReLu activation, convolution with the filter size 1×1, and max pooling with the size of 4 to retain discriminative features and discard useless ones. Dense architecture was proposed by Huang *et al.* [58], and has been widely utilized for various applications of computer vision. We utilize this architecture for sequence data which is one dimensional and

[1] http://www.image-net.org/challenges/LSVRC/

TABLE 2. Architecture summary of Res18-nRC and Res50-nRC: In both architectures, before res modules, there is a convolutional layer through which ncRNA samples are passed. Both architectures have 4 res modules, while each module of Res18-nRC has 2 basic blocks, where each basic block has two convolutional layers, but Res50-nRC architecture has variable bottleneck blocks in each res module which are mentioned by a number outside the matrix brackets, i.e., first res module has 3 bottleneck blocks and second has 4. In fist matrix (64,17) 64 represents number of feature maps and 17 shows the kernel size.

Layer_Name	Res18_nRC		Res50_nRC	
	Output Size	Parameters detail	Parameters detail	Output Size
Conv-1	64×1182	(64,3), s=1,p=1		64 x 1182
Conv-2	64×1182	$\begin{bmatrix}(64,17)\\(64,17)\end{bmatrix}\times 2, p=8$	$\begin{bmatrix}(64,1) & ,p=0\\(64,17) & ,p=8\\(256,1) & ,p=0\end{bmatrix}\times 3$	256 x 1182
Pool-1	64×591	(2, 2)		256 x 591
Conv-3	128×296	$\begin{bmatrix}(128,17)\\(128,17)\end{bmatrix}\times 2, s=2, p=8$	$\begin{bmatrix}(128,1) & ,p=0\\(128,17) & ,p=8\\(512,1) & ,p=0\end{bmatrix}\times 4, s=2$	512 x 296
Pool-2	128×148	(2, 2)		512 x 148
Conv-4	256×74	$\begin{bmatrix}(256,17)\\(256,17)\end{bmatrix}\times 2, s=2, p=8$	$\begin{bmatrix}(256,1) & ,p=0\\(256,17) & ,p=8\\(1024,1) & ,p=0\end{bmatrix}\times 6, s=2$	1024 x 74
Pool-3	256×37	(2, 2)		1024 x 37
Conv-5	512×19	$\begin{bmatrix}(512,17)\\(512,17)\end{bmatrix}\times 2, s=2, p=8$	$\begin{bmatrix}(512,1) & ,p=0\\(512,17) & ,p=8\\(2048,1) & ,p=0\end{bmatrix}\times 3, s=2$	2048 x 19
Pool-4	512 x 9	(2, 2)		2048 x 9
Output	13	Flatten-4608	Flatten-18432	13

entirely different from visual data. Integral components of the proposed methodology such as DenseNet, Dense connectivity, Composite function, Pooling layers, Growth rate and Bottleneck layers which are adapted to cope one dimensional data, are discussed below.

1) DenseNets

Consider a small non-coding RNA sample S_0 that is passed through a convolutional network. The network consists of L layers, each of which performs a non-linear conversion $H_L(\cdot)$, where L indicates the layer. $H_L(\cdot)$ may be a composite function for operations like batch normalization [61], rectified linear units (RELU) [62], Pooling [63], or Convolution (Conv). We refer to the L^{th} layer output as x_L.

a: DENSE CONNECTIVITY

State-of-the-art feed-forward convolutional networks attach the L^{th} layer output as an input to the $(L+1)^{th}$ layer, which produces the following transition layer $x_L = H_L(x_L-1)$ [54]. ResNets [57] along with skip connection strategy use an identity function to bypass non-linear transformations shown in equation 1

$$X_L = H_L\left(X_{L-1}\right) + x_{L-1} \tag{1}$$

ResNets benefit is that the gradient can flow straight from subsequent layers to previous layers through the identity function. However, the identity function and output of H_L are mixed by summation which can hinder the flow of data in the network.

We utiliZe Densenet a distinct connectivity model to further enhance the information flow between layers. In this model L^{th} layer gets all previous layers ' feature maps, $x_0, \cdots; x_{L-1}$, as input.

$$X_L = H_L\left([x_0, x_1, \cdots; x_{L-1}]\right) \tag{2}$$

In equation 2, $x_0, \cdots; x_{L-1}$ relates to the concatenation of the feature maps in the $0, \cdots, L-1$ layers

b: COMPOSITE FUNCTION

Following He *et al.* [57], we define $H_L(\cdot)$ as a composite function of three successive operations: Batch Normalization

(BN) [61], accompanied by Activation function named as rectified linear unit (ReLU) [62] and a convolution (Conv) layer.

c: TRANSITION LAYERS

We refer to the layers between blocks that perform convolution and pooling operations as transition layers. The procedure of concatenation used in equation 2 is not applicable if size of feature maps is variable. In our architecture we split the network into various tightly linked dense blocks to make the same size of feature maps. Down sampling is performed through transition layers which consist of batch normalization layer and a convolution layer of kernel size 1, followed by an average pooling layer of kernel size 4.

d: GROWTH RATE

If each composite function $H_L(\cdot)$ produces N feature maps, then L^{th} layer will have $N_0 + N \times (L-1)$ input feature-maps, where N_0 denotes number of channels in the input layer. We refer to the N hyper parameter as the network's growth rate.

B. VALIDATION METHOD AND EVALUATION CRITERIA

We perform experimentation on a small non-coding RNA classification dataset manually tagged by Antonino *et al.* [25]. This is the only benchmark dataset which is publicly available. It consists of 8920 samples that belong to 13 different ncRNA classes: miRNA, ribozymes, 5S rRNA, 5_8S_rRNA, HACA-box, CD-box, tRNA, scaRNA, IRES, Intron_gpI, Intron_gpII, riboswitch, and leader. This dataset is quite balanced as almost every class has 700 samples except the ires class which contains 520 samples. Detailed statistics of this dataset are shown in table 3.

TABLE 3. Characteristics of Non-coding RNA classification dataset, where Max-seq length and Min-seq length illustrate maximum and minimum length of neucleotides in each class.

Classes	No.of Samples	Max-seq length	Min-seq length
IRES	520	630	53
Intron_gpI	700	1182	133
leader	700	237	38
scaRNA	700	445	78
S5_rRNA	700	199	61
miRNA	700	631	52
tRNA	700	177	47
riboswitch	700	399	44
ribozyme	700	1136	41
S8_rRNA	700	290	50
CD-box	700	404	54
HACA-box	700	508	59
Intron_gpII	700	241	48

The dataset has benchmark defined split with 6320 training and 2600 test samples belonging to 13 classes of ncRNA. In the test set, each class has 200 samples, whereas in training set, each class has 500 samples except the IRES class which has 320 samples available for training. A well known statistical cross validation method namely leave one out cross validation is used to better analyze behaviour of the proposed model. We have used the training set for training and validation of the proposed model while the test set is only used for the final evaluation of the model. Furthermore, the training set is split into 5 equal parts, 4 parts are used to train the model and the 5^{th} part is used to validate the trained model. For dual evaluation, the trained model is also evaluated on the test data set which was held out separate. The process of training and dual evaluation is repeated five times where every time the test set remains the same but every next fold is taken as validation set. Final results are computed by taking the average of 5 results which are produced by the proposed model at each fold.

1) EVALUATION METRICS

The proposed system is evaluated using four different evaluation metrics namely Accuracy, Precision, Recall, and F_1 measure. All four evaluation metrics compute scores by utilizing four parameters, i.e., true positives, true negatives, false positives, and false negatives, as shown in Table 4.

TABLE 4. Confusion Matrix where True Positive illustrates the count of correctly predicted positive class values, e.g., if both the actual and predicted class labels will be yes then it will be considered as true prediction of positive class label. Similarly, True Negative is accurate prediction of negative class labels. False Positive denotes the count for wrongly predicted class labels, i.e., when actual class is 'no' but model predicts 'yes', similarly, False Negative is wrong prediction of 'no' class when actual class was 'yes'.

		Predicted Class	
		Class=yes	Class=no
Actual Class	Class=yes	True Positive	False Negative
	Class=no	False Positive	True Negative

2) ACCURACY

Accuracy is considered as a reasonable metric when dataset is symmetric-where values of false negatives and false positive are nearly equal. It computes the ratio of correctly predicted samples to the total samples.

$$Accuracy = \frac{T_p + T_n}{T_p + T_n + F_p + F_n} \tag{3}$$

3) PRECISION

Precision is the ratio between correctly predicted positive samples and total predicted positive samples.

$$Precision = \frac{T_p}{F_p + T_p} \tag{4}$$

4) RECALL

Recall is the ratio among correctly predicted positive samples and actual number of positive samples. It is also known as sensitivity. Recall is preferred when we are more concerned with false negatives. For instance, if a person having cancer is predicted as normal then the value of false negative gets high which eventually decreases the recall.

$$Recall = \frac{T_p}{F_n + T_p} \tag{5}$$

TABLE 5. Performance of the proposed RPC-snRC, Adapted (Res18-nRC, Res50-nRC), and state-of-the-art (nRC [51], and RNAGCN [26]) methodologies on the benchmark small non-coding RNA dataset.

	RPC-snRC				Res18-nRC				Res50-nRC			State-of-the-art	
Performance Measures	Character one-hot	3-mers one-hot	3-mers random embeddings	3-mers prot2vec embeddings	Character one-hot	3-mers one-hot	3-mers random embeddings	3-mers prot2vec embeddings	Character one-hot	3-mers random embeddings	3-mers prot2vec embeddings	nRC [51]	RNAGCN [26]
Accuracy	**0.9538**	0.9285	0.9327	0.9326	0.9169	0.8842	0.8880	0.9000	0.8680	0.8365	0.8915	0.7838	0.8573
Precision	**0.9539**	0.9312	0.9344	0.9322	0.9185	0.8859	0.8929	0.9000	0.8701	0.8377	0.8941	0.7780	–
Recall	**0.9538**	0.9285	0.9326	0.9326	0.9169	0.8842	0.8880	0.9000	0.8680	0.8365	0.8915	0.7830	–
F1-Score	**0.9536**	0.9286	0.9328	0.9319	0.9174	0.8842	0.8880	0.8987	0.8680	0.8357	0.8921	0.7790	0.8561

5) F_1 MEASURE

F_1 measure is harmonic average of precision and recall. It performs better than accuracy for imbalance class distributed dataset because it keeps track of both precision and recall.

$$F_1 Measure = 2 \times \frac{Precision \times Recall}{Precision + Recall} \tag{6}$$

IV. EXPERIMENTAL SETUP AND RESULTS

We implement the proposed RPC-snRC and ResNet based methodologies in Python using Pytorch [64]. Detailed parametric description about adapted ResNet based methodologies is summarized in Table 2. Cross entropy is used as a loss function with Adam [65] optimizer where learning rate is initialized from 0.001. In order to alleviate training time, an early stopping approach is used. High-performance NVIDIA GeForce GTX 1080Ti GPU is used for experimentation.

Results: This section briefly describes the performance of the proposed RPC-snRC classification system and two adapted ResNet architectures (res-net 18 layer, res-net 50 layer) for the task of ncRNA classification. It shows the impact of three sequence representation schemes while treating RNA sequence as a set of characters, and k-mers based word for both proposed and adapted methodologies. In the benchmark dataset maximum length of the sequence is 1180, so to make the length of sequences equal, we apply paddings for the sequences which has length less than 1180. To make all sequences of equal lengths, we apply padding where the size of sequence is less than 1180. Experimentation is performed in two different ways: First, RNA sequence is taken as a set of characters with two different representation schemes namely one hot vector encoding and random embedding initialization, which are separately fed to the proposed RPC-snRC system. Second, we generate 3-mers of the sequence by sliding a window of size three on the sequence. K-mers based sequence representation along with one hot vector encoding, random embedding initialization, and pretrained word embeddings provided by Asgari *et al.* [60] are fed to the proposed RPC-snRC system.

Table 5 compares the performance of state-of-the-art and adapted res-net based methodologies with the proposed RPC-snRC methodology for the task of small non-coding RNA classification. It also illustrates the performance of the proposed RPC-snRC methodology when RNA sequence is treated as set of characters, 3-mers based features with random, and pre-trained neural word embeddings. As is depicted by the Table 5 renowned methodology proposed by Antonio Fiannaca *et al.* [25] managed to achieve the performance figures of 78%, 77%, 78%, and 77% in terms of accuracy, precision, recall, and F 1 measure, respectively. This performance is outperformed by a recent Graph Convolutional Neural architecture based methodology given by Rossi *et al.* [26] as it marked state-of-the-art performance for small non-coding RNA classification with 85.7% accuracy. However, the adapted ResNet-18 and Res-Net-50 manage to produce the peak performance of 91%, and 89% by representing RNA sequences as character with one hot encoding and as 3-mers features with pre-trained prot2vec embedding, respectively. On the other hand, the proposed RPC-snRC classification system has significantly outperformed the state-of-the-art methodology as well as the two adapted ResNet architectures in all settings. While, RPC-snRC with 3-mers random embedding initialization and pre-trained neural word embeddings schemes has raised state-of-the-art performance almost by the figure of 8% in terms of F_1 measure, the RPC-snRC with character level features and one hot encoding manages to mark the peak performance at 95% thereby clearly outperforming all the other systems (previously existing systems and ResNet based systems adapted in this research).

In a nutshell, convolutional neural network based deep architectures have the ability to extract discriminative features directly from primary sequences of small non-coding RNA. This is depicted by the results where performances of the proposed and adapted methodologies are significantly higher than the state-of-the-art methodologies which take secondary structural features as input. Moreover, performance of ResNet based architectures is lower than the performance of the proposed RPC-snRC methodology because in ResNet models gradient does not flow properly from subsequent layers to previous layers [58]. It can also be inferred that ResNet model with 50 layers extracted some irrelevant and redundant features which slightly reduced its performance as compared to the performance of ResNet 18 layers model.

A. CLASS LEVEL PERFORMANCE COMPARISON OF PROPOSED RPC-snRC AND STATE-OF-THE-ART nRC METHODOLOGIES

In order to further compare the performance of the proposed RPC-snRC and the adapted ResNet based methodologies with the state-of-the-art methods, a class level performance comparison is performed in terms of accuracy confusion matrix. Accuracy confusion matrices of RPC-snRC, ResNet-18, and nRC methodologies on the test set of nCR dataset are shown in the Figure 3.RNAGCN [26] is the most recently reported method for small non-coding RNA classification, however, the authors have not provided class level results of their method. Therefore, we performed class level performance comparison of the proposed RPC-snRC

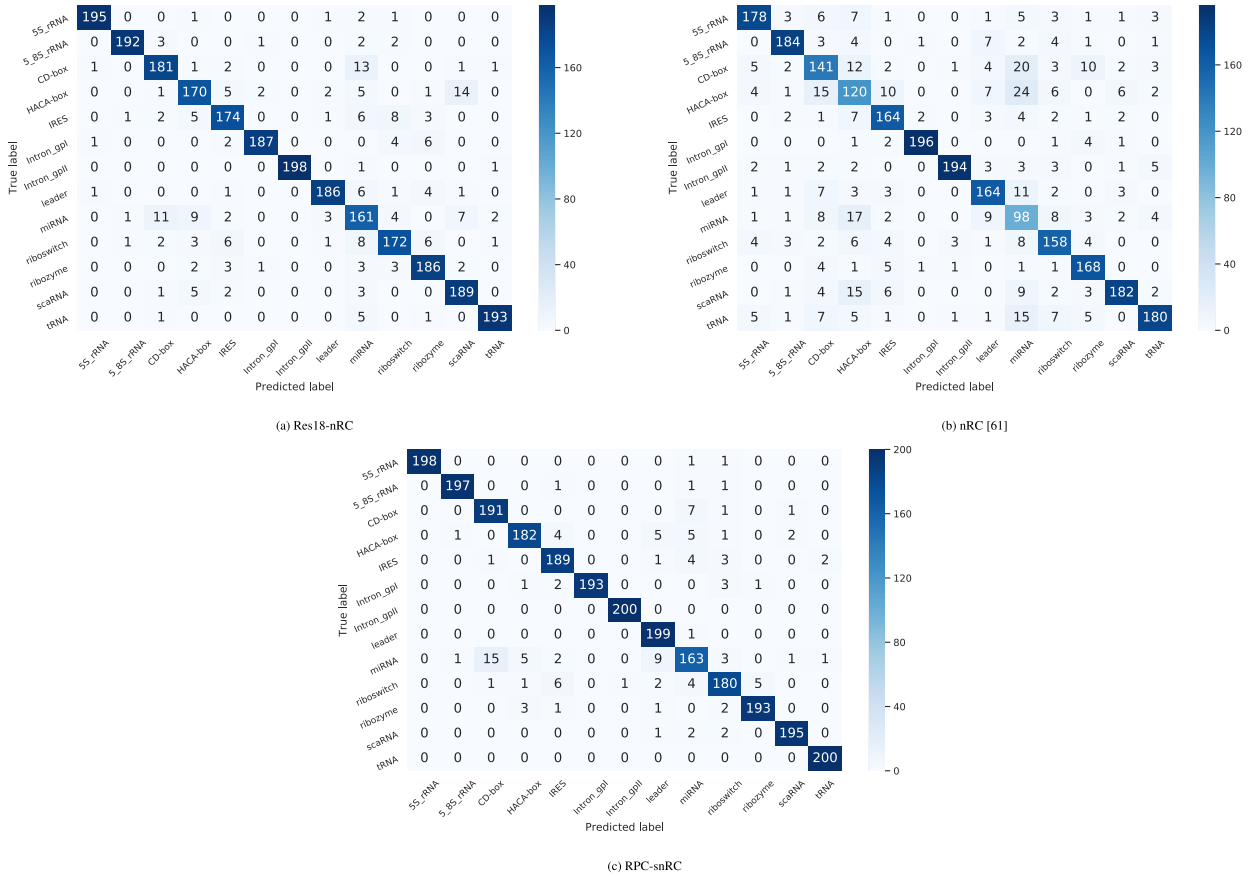
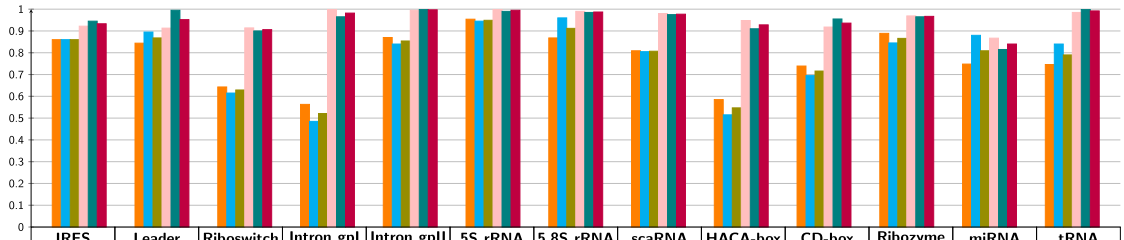


FIGURE 3. Accuracy Confusion matrix of the proposed RPC-snRC, Adapted Res18-nRC and state-of-the-art nRC [51] classification methodologies.

and adapted methodologies with nRC classification methodology. The proposed RPC-snRC and the adapted Res18snRC based methodologies produce highest performance with character level and one hot vector representation. So here we take confusion matrices of both methodologies with highest performance values. As depicted in Figure 3, RPCsnRC methodology correctly classifies all 200 samples of two classes namely Intron gpII and tRNA as compared to the state-of-the-art nRC methodology which manages to correctly classify only 180 samples of tRNA and 196 samples of Intron gpII class. Performance of Res18-snRC remains in between the performance of nrc and RPC-snRC methodologies as it correctly predicted 198 samples of Intron gpII and 193 samples of tRNA class. In addition, state-of-the-art nRC methodology fails to mark prominent performance as significant samples of almost every class are mistakenly classified in miRNA, HACA-box, CD-box, and IRES classes, while, only a few samples of each class are misclassified in the proposed RPC-snRC methodology.

Although, miRNA has shown the lowest performance amongst all classes in both methodologies, however, the proposed RPC-snRC still correctly classifies 163 samples out of the maximum possible 200 as compared to state-of-the-art nRC methodology which only manages to correctly classify 98 samples. Also, the proposed RPC-snRC methodology successfully classifies more than 190 samples in each of the nine classes, i.e., introl_gpll, tRNA, 5S_rRNA, 5_8S_rRNA, leader, scaRNA, ribozyme, introl_gpl, and CD-box. Whereas, the other classes achieve counts of 180's and 160's as shown by Figure3. In contrast to the state-of-the-art nRC methodology, only two classes intron_gpl, and intron_gpll correctly classify more than 190 samples. Similarly, the adapted Res18-nRc methodology was able to correctly predict more than 190 samples for 4 classes, namely 5S_rRNA, 5_8S_rRNA, introl_gpll, and tRNA.

Figure 4 shows individual class level performances of RPC-snRC and nRC classification methodologies over small ncRNA classification dataset in terms of precision, recall, and F_1 measure. Overall, for all classes, RPC-snRC methodology significantly outperforms the state-of-the-art nRC methodology in all three performance metrics with exception of the miRNA class, where nRC methodology manages to deliver better recall figure. Moreover, amongst all performance metrices, nRC classification methodology manages to sustain performance values of precision, recall and F_1 measure only for three classes(IRES, 5.8S rRNA, scaRNA), on the other hand, the performance of RPC-snRC classification methodology stays consistent for 7 classes namely ribozymes, 5_8S_rRNA, tRNA, scaRNA, Intron_gpII, and riboswitch. This unique behavior of RPC-snRC methodology



		IRES	Leader	Riboswitch	Intron_gpI	Intron_gpII	5S_rRNA	5.8S_rRNA	scaRNA	HACA-box	CD-box	Ribozyme	miRNA	tRNA
nRC	Precision	0.860	0.844	0.643	0.563	0.870	0.954	0.868	0.809	0.585	0.739	0.889	0.748	0.746
	Recall	0.860	0.895	0.615	0.485	0.840	0.945	0.960	0.805	0.515	0.695	0.845	0.880	0.840
	F1-Score	0.860	0.868	0.6291	0.521	0.854	0.949	0.912	0.807	0.547	0.716	0.866	0.809	0.790
Deep nRC	Precision	0.922	0.913	0.914	1.000	0.995	1.000	0.990	0.980	0.948	0.918	0.969	0.867	0.985
	Recall	0.945	0.995	0.900	0.965	1.000	0.990	0.985	0.975	0.910	0.955	0.965	0.815	1.000
	F1-Score	0.933	0.952	0.906	0.982	0.997	0.995	0.987	0.977	0.928	0.936	0.967	0.840	0.992

FIGURE 4. Individual class performance of proposed RPC-snRC methodology and state-of-the-art nRC [51] methodology on small ncRNA classification dataset.

shows that it suffers less from type I, and type II errors as compared to nRC methodology-performance of whom seems less stable at class level.

V. CONCLUSION

This article proposes a novel methodology, named RPC-snRC, which classifies small non-coding RNA sequences into their relevant families by utilizing positioning and occurrences information of various nucleotides. Experimental results have shown that the proposed RPC-snRC methodology is highly robust as it is neither biased towards false positive nor towards false negative predictions. Adapted Res18-snRC and Res50-snRC methodologies perform better than the state-of the-art small non-coding RNA classification methodologies. However, their performance is not as promising as with the proposed RPC-snRC methodology because in ResNet architectures gradient cannot flow properly from subsequent layers to previous layers. The proposed RPC-snRC methodology marks highest F1-score of 95% by representing character based features through one hot encoding as compared to state-of-the-art ncRNA, RNAGCN, and the adapted Res18-nRC, Res50-nRC classification methodologies which only manage to produce the performance of 77%, 85%, 91%, and 89% respectively. Moreover, in our experimentation, almost all methodologies perform better with one hot vector encoding than randomly initialized or pre-trained word embeddings. From these results, it can be concluded that character or atom level feature generates better performance as compared to k-mers based features. A compelling future line of our work would be exploring the impact of a hybrid methodology combining the benefits of both primary and secondary structural features.

REFERENCES

[1] W. Kraczkowska and P. P. Jagodziński, "The long non-coding RNA landscape of atherosclerotic plaques," *Mol. Diagnosis Therapy*, vol. 23, pp. 735–749, Oct. 2019.

[2] H. Ghasemi, Z. Sabati, H. Ghaedi, Z. Salehi, and B. Alipoor, "Circular RNAs in β-cell function and type 2 diabetes-related complications: A potential diagnostic and therapeutic approach," *Mol. Biol. Rep.*, vol. 46, pp. 5631–5643, Jul. 2019.

[3] F. Hubé and C. Francastel, "Coding and non-coding RNAs, the frontier has never been so blurred," *Frontiers Genet.*, vol. 9, p. 140, Apr. 2018.

[4] N. Amin, A. McGrath, and Y.-P. P. Chen, "Evaluation of deep learning in non-coding RNA classification," *Nature Mach. Intell.*, vol. 1, no. 5, p. 246, 2019.

[5] Y. Fang and M. J. Fullwood, "Roles, functions, and mechanisms of long non-coding RNAs in cancer," *Genomics, Proteomics Bioinf.*, vol. 14, no. 1, pp. 42–54, Feb. 2016.

[6] N. Yu, Z. Yu, and Y. Pan, "A deep learning method for lincRNA detection using auto-encoder algorithm," *BMC Bioinf.*, vol. 18, no. 15, p. 511, Dec. 2017.

[7] M. Chaabane, "End-to-end learning framework for circular RNA classification from other long non-coding RNAs using multi-modal deep learning," Ph.D. dissertation, 2018, Paper 2954.

[8] Y. Ma, X. Zhang, Y.-Z. Wang, H. Tian, and S. Xu, "Research progress of circular RNAs in lung cancer," *Cancer Biol. Therapy*, vol. 20, no. 2, pp. 123–129, Feb. 2019.

[9] T. Li, S. Wang, R. Wu, X. Zhou, D. Zhu, and Y. Zhang, "Identification of long non-protein coding RNAs in chicken skeletal muscle using next generation sequencing," *Genomics*, vol. 99, no. 5, pp. 292–298, May 2012.

[10] Y.-J. Kang, D.-C. Yang, L. Kong, M. Hou, Y.-Q. Meng, L. Wei, and G. Gao, "CPC2: A fast and accurate coding potential calculator based on sequence intrinsic features," *Nucleic Acids Res.*, vol. 45, no. 1, pp. W12–W16, Jul. 2017.

[11] J. Baek, B. Lee, S. Kwon, and S. Yoon, "LncRNANet: Long non-coding RNA identification using deep learning," *Bioinformatics*, vol. 34, no. 22, pp. 3889–3897, Nov. 2018.

[12] C. Yang, L. Yang, M. Zhou, H. Xie, C. Zhang, M. D. Wang, and H. Zhu, "LncADeep: An *ab initio* lncRNA identification and functional annotation tool based on deep learning," *Bioinformatics*, vol. 34, no. 22, pp. 3825–3834, 2018.

[13] S. Han, Y. Liang, Q. Ma, Y. Xu, Y. Zhang, W. Du, C. Wang, and Y. Li, "LncFinder: An integrated platform for long non-coding RNA identification utilizing sequence intrinsic composition, structural information and physicochemical property," *Briefings Bioinf.*, vol. 20, no. 6, pp. 2009–2027, Nov. 2019.

[14] G. Stefani and F. J. Slack, "Small non-coding RNAs in animal development," *Nature Rev. Mol. Cell Biol.*, vol. 9, no. 3, p. 219, 2008.

[15] M. Jovanovic and M. Hengartner, "miRNAs and apoptosis: RNAs to die for," *Oncogene*, vol. 25, no. 46, p. 6176, 2006.

[16] I. Büssing, F. J. Slack, and H. Großhans, "Let-7 microRNAs in development, stem cells and cancer," *Trends Mol. Med.*, vol. 14, no. 9, pp. 400–409, Sep. 2008.

[17] R. Schickel, B. Boyerinas, S. Park, and M. Peter, "MicroRNAs: Key players in the immune system, differentiation, tumorigenesis and cell death," *Oncogene*, vol. 27, no. 45, p. 5959, 2008.

[18] B. Hrdlickova, R. C. de Almeida, Z. Borek, and S. Withoff, "Genetic variation in the non-coding genome: Involvement of micro-RNAs and long non-coding RNAs in disease," *Biochimica et Biophysica Acta (BBA)-Mol. Basis Disease*, vol. 1842, no. 10, pp. 1910–1922, Oct. 2014.

[19] M. Esteller, "Non-coding RNAs in human disease," *Nature Rev. Genet.*, vol. 12, no. 12, p. 861, 2011.

[20] S. A. Elela and R. N. Nazar, "Role of the 5.8S rRNA in ribosome translocation," *Nucleic Acids Res.*, vol. 25, no. 9, pp. 1788–1794, May 1997.
[21] B. M. Fontoura, C. A. Atienza, E. A. Sorokina, T. Morimoto, and R. B. Carroll, "Cytoplasmic p53 polypeptide is associated with ribosomes," *Mol. Cellular Biol.*, vol. 17, no. 6, pp. 3146–3154, Jun. 1997.
[22] R. Shi and V. L. Chiang, "Facile means for quantifying microRNA expression by real-time PCR," *BioTechniques*, vol. 39, no. 4, pp. 519–525, Oct. 2005.
[23] D. Ammons, J. Rampersad, and G. E. Fox, "5S rRNA gene deletions cause an unexpectedly high fitness loss in escherichia coli," *Nucleic Acids Res.*, vol. 27, no. 2, pp. 637–642, Jan. 1999.
[24] R. N. Nazar, "The ribosomal 5.8 S RNA: Eukaryotic adaptation or processing variant?" *Can. J. Biochem. Cell Biol.*, vol. 62, no. 6, pp. 311–320, 1984.
[25] A. Fiannaca, M. L. Rosa, L. L. Paglia, R. Rizzo, and A. Urso, "NRC: Non-coding RNA classifier based on structural features," *BioData Mining*, vol. 10, no. 1, p. 27, Dec. 2017.
[26] E. Rossi, F. Monti, M. Bronstein, and P. Liu, "NcRNA classification with graph convolutional networks," 2019, *arXiv:1905.06515*. [Online]. Available: http://arxiv.org/abs/1905.06515
[27] X. Fu, W. Zhu, L. Cai, B. Liao, L. Peng, Y. Chen, and J. Yang, "Improved pre-miRNAs identification through mutual information of pre-miRNA sequences and structures," *Frontiers Genet.*, vol. 10, p. 119, Feb. 2019.
[28] S. Washietl, I. L. Hofacker, and P. F. Stadler, "Fast and reliable prediction of noncoding RNAs," *Proc. Nat. Acad. Sci. USA*, vol. 102, no. 7, pp. 2454–2459, 2005.
[29] S. Griffiths-Jones, A. Bateman, M. Marshall, A. Khanna, and S. R. Eddy, "Rfam: An RNA family database," *Nucleic Acids Res.*, vol. 31, no. 1, pp. 439–441, Jan. 2003.
[30] S. Washietl and I. L. Hofacker, "Consensus folding of aligned sequences as a new measure for the detection of functional RNAs by comparative genomics," *J. Mol. Biol.*, vol. 342, no. 1, pp. 19–30, Sep. 2004.
[31] I. L. Hofacker, M. Fekete, and P. F. Stadler, "Secondary structure prediction for aligned RNA sequences," *J. Mol. Biol.*, vol. 319, no. 5, pp. 1059–1066, Jun. 2002.
[32] J. Liu, J. Gough, and B. Rost, "Distinguishing protein-coding from noncoding RNAs through support vector machines," *PLoS Genet.*, vol. 2, no. 4, p. e29, Apr. 2006.
[33] P. Carninci, T. Kasukawa, S. Katayama, J. Gough, M. C. Frith, N. Maeda, R. Oyama, T. Ravasi, B. Lenhard, C. Wells, and R. Kodzius, "The transcriptional landscape of the mammalian genome," *Science*, vol. 309, no. 5740, pp. 1559–1563, Sep. 2005.
[34] C. Liu, B. Bai, G. Skogerbø, L. Cai, W. Deng, Y. Zhang, D. Bu, Y. Zhao, and R. Chen, "NONCODE: An integrated knowledge database of non-coding RNAs," *Nucleic Acids Res.*, vol. 33, pp. D112–D115, Dec. 2004.
[35] K. C. Pang, S. Stephen, P. G. Engström, K. Tajul-Arifin, W. Chen, C. Wahlestedt, B. Lenhard, Y. Hayashizaki, and J. S. Mattick, "RNAdb—A comprehensive mammalian noncoding RNA database," *Nucleic Acids Res.*, vol. 33, pp. D125–D130, Dec. 2004.
[36] S. Lertampaiporn, C. Thammarongtham, C. Nukoolkit, B. Kaewkamnerdpong, and M. Ruengjitchatchawalya, "Identification of non-coding RNAs with a new composite feature in the hybrid random forest ensemble algorithm," *Nucleic Acids Res.*, vol. 42, no. 11, p. e93, Jun. 2014.
[37] K. D. Pruitt, T. Tatusova, G. R. Brown, and D. R. Maglott, "NCBI reference sequences (RefSeq): Current status, new features and genome annotation policy," *Nucleic Acids Res.*, vol. 40, no. 1, pp. D130–D135, Jan. 2012.
[38] P. P. Amaral, M. B. Clark, D. K. Gascoigne, M. E. Dinger, and J. S. Mattick, "LncRNAdb: A reference database for long noncoding RNAs," *Nucleic Acids Res.*, vol. 39, pp. D146–D151, Jan. 2011.
[39] S. F. Altschul, W. Gish, W. Miller, E. W. Myers, and D. J. Lipman, "Basic local alignment search tool," *J. Mol. Biol.*, vol. 215, no. 3, pp. 403–410, Oct. 1990.
[40] R. Achawanantakun, J. Chen, Y. Sun, and Y. Zhang, "LncRNA-ID: Long non-coding RNA identification using balanced random forests," *Bioinformatics*, vol. 31, no. 24, pp. 3897–3905, 2015.
[41] G. Chen, Z. Wang, D. Wang, C. Qiu, M. Liu, X. Chen, Q. Zhang, G. Yan, and Q. Cui, "LncRNADisease: A database for long-non-coding RNA-associated diseases," *Nucleic Acids Res.*, vol. 41, no. 1, pp. D983–D986, Nov. 2012.
[42] Y. Saito, K. Sato, and Y. Sakakibara, "Fast and accurate clustering of noncoding RNAs using ensembles of sequence alignments and secondary structures," *BMC Bioinf.*, vol. 12, no. 1, p. 48, 2011.
[43] J. H. Havgaard, E. Torarinsson, and J. Gorodkin, "Fast pairwise structural RNA alignments by pruning of the dynamical programming matrix," *PLoS Comput. Biol.*, vol. 3, no. 10, p. e193, 2007.
[44] K. Sato, T. Mituyama, K. Asai, and Y. Sakakibara, "Directed acyclic graph kernels for structural RNA analysis," *BMC Bioinf.*, vol. 9, no. 1, p. 318, Dec. 2008.
[45] S. Will, K. Reiche, I. L. Hofacker, P. F. Stadler, and R. Backofen, "Inferring noncoding RNA families and classes by means of genome-scale structure-based clustering," *PLoS Comput. Biol.*, vol. 3, no. 4, p. e65, Apr. 2007.
[46] M. Miladi, A. Junge, F. Costa, S. E. Seemann, J. H. Havgaard, J. Gorodkin, and R. Backofen, "RNAscClust: Clustering RNA sequences using structure conservation and graph based motifs," *Bioinformatics*, vol. 33, no. 14, pp. 2089–2096, Jul. 2017.
[47] G. Aoki and Y. Sakakibara, "Convolutional neural networks for classification of alignments of non-coding RNA sequences," *Bioinformatics*, vol. 34, no. 13, pp. i237–i244, Jul. 2018.
[48] S. Park, S. Min, H.-S. Choi, and S. Yoon, "Deep recurrent neural network-based identification of precursor microRNAs," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 2891–2900.
[49] M. Tsuchiya, K. Amano, M. Abe, M. Seki, S. Hase, K. Sato, and Y. Sakakibara, "SHARAKU: An algorithm for aligning and clustering read mapping profiles of deep sequencing in non-coding RNA processing," *Bioinformatics*, vol. 32, no. 12, pp. i369–i377, Jun. 2016.
[50] R. Weikard, F. Hadlich, and C. Kuehn, "Identification of novel transcripts and noncoding RNAs in bovine skin by deep next generation sequencing," *BMC Genomics*, vol. 14, no. 1, p. 789, 2013.
[51] S. Memczak, M. Jens, A. Elefsinioti, F. Torti, J. Krueger, A. Rybak, L. Maier, S. D. Mackowiak, L. H. Gregersen, M. Munschauer, A. Loewer, U. Ziebold, M. Landthaler, C. Kocks, F. L. Noble, and N. Rajewsky, "Circular RNAs are a large class of animal RNAs with regulatory potency," *Nature*, vol. 495, no. 7441, p. 333, 2013.
[52] K. Sato, Y. Kato, T. Akutsu, K. Asai, and Y. Sakakibara, "DAFS: Simultaneous aligning and folding of RNA sequences via dual decomposition," *Bioinformatics*, vol. 28, no. 24, pp. 3218–3224, Dec. 2012.
[53] J. S. McCaskill, "The equilibrium partition function and base pair binding probabilities for RNA secondary structure," *Biopolymers*, vol. 29, nos. 6–7, pp. 1105–1119, May 1990.
[54] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2012, pp. 1097–1105.
[55] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2014, *arXiv:1409.1556*. [Online]. Available: http://arxiv.org/abs/1409.1556
[56] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 1–9.
[57] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.
[58] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 4700–4708.
[59] A. Fabijanska and S. Grabowski, "Viral genome deep classifier," *IEEE Access*, vol. 7, pp. 81297–81307, 2019.
[60] E. Asgari and M. R. K. Mofrad, "Continuous distributed representation of biological sequences for deep proteomics and genomics," *PLoS ONE*, vol. 10, no. 11, Nov. 2015, Art. no. e0141287.
[61] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," 2015, *arXiv:1502.03167*. [Online]. Available: http://arxiv.org/abs/1502.03167
[62] X. Glorot, A. Bordes, and Y. Bengio, "Deep sparse rectifier neural networks," in *Proc. 14th Int. Conf. Artif. Intell. Statist.*, 2011, pp. 315–323.
[63] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proc. IEEE*, vol. 86, no. 11, pp. 2278–2324, Nov. 1998.
[64] V. Subramanian, *Deep Learning With PyTorch: A Practical Approach to Building Neural Network Models Using PyTorch*. Birmingham, U.K.: Packt Publishing Ltd, 2018.
[65] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2014, *arXiv:1412.6980*. [Online]. Available: http://arxiv.org/abs/1412.6980