Life Science Informatics Publications

Research Journal of Life Sciences, Bioinformatics, Pharmaceutical and Chemical Sciences

Journal Home page http://www.rjlbpcs.com/

RJLBPCS

ISSN 2454-6348

Original Review Article **DOI: 10.26479/2019.0502.54**

CLASSIFICATION OF NON-CODING RNA - A REVIEW FROM MACHINE LEARNING PERSPECTIVE

Jasmit Singh[1], Shailendra Singh[1], Dharam Vir[2]

1. Computer Science and Engineering Department, Punjab Engineering College, Sector 12, Chandigarh, 160012, India.

2. Department of Otolaryngology, PGIMER, Chandigarh, India

ABSTRACT: Non-coding RNA (ncRNA) is the RNA that is not converted into protein rather produces functional RNA molecules. The type of non-coding RNAs include microRNA, snoRNA and many other small RNAs like siRNAs and microRNAs. Identifying non-coding RNA has emerged over the past decade as a hot trend in bioinformatics. Many techniques are developed for classification of non-coding RNA and it is appropriate to select a particular technique according to the situation and circumstances. In this article, several machine learning techniques on classification of non-coding RNA are discussed with their merits, limitations and application scope to aid people in selecting a suitable method and obtaining a reliable result. These techniques are compared on the basis of their performance and accuracy. The future scope is also provided.

KEYWORDS: ncRNA, Convolutional Neural Network, Random Forest, Recurrent Neural Network, Machine Learning.

Corresponding Author: Mr. Jasmit Singh* M.E.

Computer Science and Engineering Department, Punjab Engineering College, Sector 12, Chandigarh, 160012, India.

1.INTRODUCTION

Non-coding RNA (ncRNA) does not produce encoding proteins [1]. They do produce functional RNA [2]. Most non-coding RNAs provide functions that are comparatively generic in cells like tRNAs and rRNAs are used in mRNA translation and small nuclear RNAs are used in splicing. Earlier, Non-coding RNA (ncRNA) were seen as junk gene or transcriptional noise [3]. This

perception has changed over time. Now it is no longer seen as junk. Due to massive volumes of data used in human next-generation sequencing (NGS), finding the structure and function of ncRNAs is a difficult task [4]. The research in non-coding RNAs is now being taken with more interest, as many biological functions associated with it, have been found. ncRNAs are involved in many processes like gene regulation in mammals [5]. Non-coding RNA's close associativity with human disorders and diseases like cancer [6], makes it an important topic for research in healthcare. Over the last decade, classification of non-coding RNA has become a hot trend in bioinformatics. There are many non-coding RNA types like ribosomal RNAs (rRNAs), transfer RNAs (tRNAs), small RNAs such as siRNAs, microRNAs, snoRNAs, piRNAs, snRNAs, scaRNAs, exRNAs and long ncRNAs such as Xist and HOTAIR [7]. The classification of non-coding RNA is shown in Figure 1. Different non-coding RNAs are associated with different biological functions. So, classification of non-coding RNA is very important to explore the functionality of non-coding RNA. Many machine learning techniques are used for classification of non-coding RNA, like recurrent neural network (RNN), convolutional neural network (CNN), hierarchical clustering, random forest, support vector machine (SVM) and deep sequencing. Machine learning [8] is the study of statistical models and algorithms that helps the computers to efficiently perform a task without using direct instructions. Machine learning is a subset of artificial intelligence. Only inferences and models are used by machine learning. A model of sample data known as training data is used by machine learning algorithms that helps in making decisions or predictions without programming explicitly to complete the task [9]. They are used in many applications like computer vision, email filtering, network intruder's detection and classification of non-coding RNA. Machine learning is used where developing an algorithm of specific instructions to perform a task is difficult.

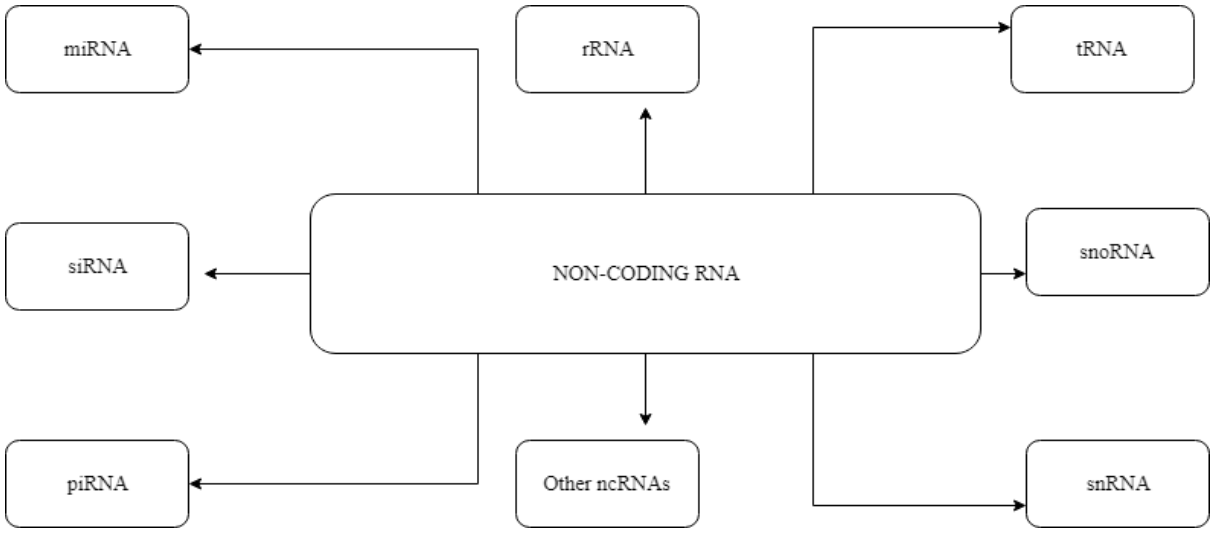

Figure 1: Classification of Non-Coding RNA

The main focus of the paper is to discuss machine learning techniques used in classification of non-coding RNA and their comparison based on performance and accuracy. The rest of the paper is as follows, the next section describes overview of non-coding RNA and its types. In the subsequent section, various machine learning algorithms used for classification of non-coding RNA are discussed along with their applications. Thereafter, comparison of these methods is presented. Finally, conclusion and future scope of the research are provided.

2.Background

The non-coding RNA was not taken with much interest in the past. It was treated as garbage data in the beginning. But with recent researches, it has been established that non-coding RNA do have biological functions and many diseases like Alzheimer and cancer are caused due to the non-coding RNA [6]. Over the last decade, several machine learning techniques were used for detection of non-coding RNA that incorporated both supervised and unsupervised learning. These machine learning techniques were used to identify new non-coding RNA. Many new non-coding RNAs were found with different biological function[6]. Support Vector Machine (SVM) is supervised learning model that is used for regression and classification. SVM is implemented by RNAz for classification of non-coding RNA. RNAz implements feature extraction of non-coding RNA [10]. This method uses multiple sequence alignment. An issue with SVM is that it can be used to find only two classes, it cannot be used to classify multiple types of non-coding RNAs at once. It can only be used for distinction between coding and non-coding RNA. SVM cannot identify new types of non-coding RNA which is the basic requirement for exploring their biological function. Another supervised learning technique that is used to identify non-coding RNA is hybrid random forest. This technique uses new feature SCORE built on function of logistic regression that combines five features-sequence, structure, modularity, structural robustness and coding potential [11]. Hybrid random forest uses genetic algorithm and correlation-based feature selection to capture features. The drawback of supervised learning is that it cannot work on unlabelled data. Most of the data of non-coding RNA is unlabelled as it is still unexplored. For unlabelled data, unsupervised learning methods are used that find clusters of data and then map the new data into these clusters [12]. To improve on the shortcomings of supervised learning, unsupervised learning techniques like clustering are used. Hierarchical clustering is unsupervised learning technique used for clustering unknown data. Hierarchical clustering is implemented by RNAscClust, which is used for clustering RNA sequences by means of graph-based motifs and structure conservation [13]. This method makes groups of paralogous RNAs according to their structural similarities. RNAscClust incorporates multiple alignment of RNA sequences. This method uses minimum free energy structures for every sequence as a prior information for the folding. Hierarchical clustering is also implemented by EnsembleClust. This method helps in identifying new subfamilies of non-coding RNA. It uses

structural alignment score as performance metrics [14]. This clustering technique uses all sequence alignment and secondary structures. This approach improves on techniques that were previously used like LocARNA and FOLDALIGN. Previous techniques calculated similarity based on only one optimal structural alignment's score. EnsembleClust provides balance between clustering quality and computational cost. Deep sequencing is another unsupervised learning technique, useful in sequencing the non-coding RNA sequences. Deep sequencing is implemented by SHARAKU. SHARAKU is a new algorithm that uses next generation sequencing data and aligns two read mapping profiles of non-coding RNAs [15]. It uses sequence information and secondary structure information simultaneously for the detection of non-coding RNA. Artificial neural networks also implement unsupervised learning. Convolutional neural networks are implemented by CNNClust. CNNClust is a machine learning technique that incorporates convolutional neural network for the detection of non-coding RNA [16]. CNNClust uses pairwise alignment and extracts position weight matrices of sequence motifs that are used for training of convolutional neural network. Recurrent neural networks are implemented by lncRNAnet. lncRNAnet is a machine learning technique to identify long non-coding RNA using deep learning [17]. It uses both recurrent neural networks (RNN) and convolutional neural network (CNN). It performs well for short length sequences. In all these techniques, the input is the non-coding RNA sequences generally in fasta format, taken from the databases of non-coding RNA. There are many databases of non-coding RNA available including RFam, Hugo Gene nomenclature committee (HGNC) database and Genomic tRNA database.

3. Machine Learning Techniques for Classification of Non-Coding RNA

In this section, the various machine learning techniques being used in classification of non-coding RNA and their applications are discussed. These techniques are used to categorize the dataset into different non-coding RNA types. There are three types of machine learning algorithms- supervised learning, unsupervised learning and reinforcement learning as presented in Figure 2.

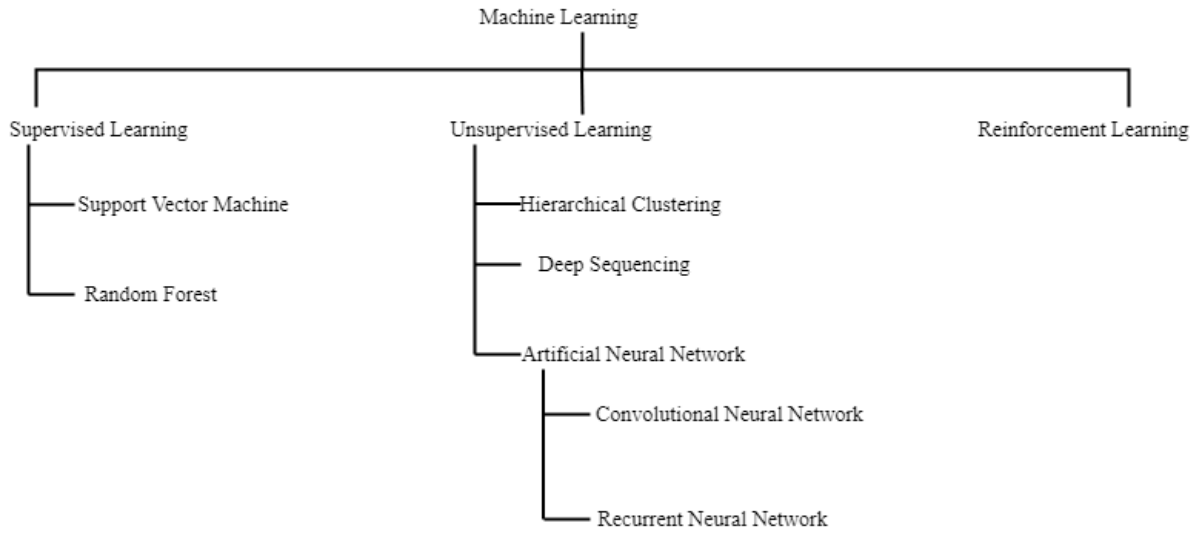

Figure 2: Classification of Machine Learning Techniques

3.1 Supervised Learning

Supervised learning algorithms create a mathematical model that contains both inputs and their anticipated outputs [18]. This collection of input-output data pairs is called as training data which consists of training example's set. Every training example has input-output pair consisting of one or more inputs and its corresponding desired output. Optimization of the objective function is done iteratively. The supervised learning algorithm uses a function that can predict the output corresponding to the new inputs. These new inputs are not in training data and are known as testing data [19]. The algorithm that improves the performance accuracy of the predicting outputs learns in performing the task [20]. Supervised learning includes regression and classification. Regression algorithms are used when output can have any numerical value within a range and classification algorithm are used when output range consists of limited set of values. Supervised learning is used in recommendation systems, ranking, visual identity tracking and speaker verification. Supervised learning can used to classify non-coding RNA into different classes of non-coding RNA. It works on labelled data from the non-coding dataset. The techniques that use supervised learning are SVM and Random forest.

3.1.1 Support Vector Machines (SVM)

Support vector machines (SVMs) are collection of supervised learning approaches, that are used for regression and classification. SVM classifies the training data examples into two categories of data. SVM training algorithm is a binary, non-probabilistic and linear classifier[12]. It can also achieve non-linear classification by means of kernel trick, i.e. indirectly plotting the inputs into high dimensional feature space. The methods that implements SVM for classification of non-coding RNA are provided in Table 1. Peter F. Stadler et al. [10] defined a method RNAz for detecting non-coding RNA which implements support vector machines. RNAz combines structure prediction and comparative sequence analysis. In RNAz, two basic components are used- thermodynamic stability measure and consensus secondary structure. This method incorporates both pairwise alignment and multiple sequence alignment of the non-coding RNA sequence with high specificity and high sensitivity. The database used in this machine learning technique is RFAM database[21], a comparative regulatory genomics database, i.e. the database of non-coding RNA of humans, rats, mice and zebrafish. RNAz uses minimum free energy (MFE) RNA folding. RNAz includes calculating z-scores using regression by SVM. The SVM is used for binary classification, i.e., it tells whether it is non-coding RNA or any other sequence. The input parameters that are used includes MFE z score's mean [22] of the different sequences in the alignment without gaps, the number of sequences in the alignment, the mean pairwise identity , and the structure conservation index (SCI) of the alignment. This method takes help of the program RNAALIFOLD[23] that was developed originally to guess secondary structure in aligned sequence. This technique uses a folding algorithm

for prediction of RNA's secondary structure by implementing the dynamic programming algorithms. When the SCI is close to zero, it indicates that consensus structure is not found by RNAALIFOLD, on the other hand, perfectly conserved structures set has SCI close to 1. RNAz provides good results for large scale genomic annotation. Jinfeng Liu et al. [24] stated a coding or non-coding (CONC) method for classifying non-coding RNA. CONC also implements SVM. It incorporates multiple sequence alignment. The databases used by this method are RNAdb[25], NONCODE[26] and FANTOM3[27]. It uses protein features for classifying non-coding RNA like amino acid composition, peptide length, predicted percentage of exposed residues, number of homologs from database searches, compositional entropy, predicted secondary structure content and alignment entropy. The limitation of SVM's implementations is that they can only work on labelled data to classify the non-coding RNA. These methods cannot work on unannotated data. Most of the data is unlabelled as many types of non-coding RNA are still to be identified.

Table 1: Studies related to ncRNA classification using SVM

Author, citation and year	Method	Features	Alignment	Database
Peter F. Stadler et al.[10], 2005	RNAz [10]	Thermodynamic stability measure, consensus secondary structure	Pairwise and multiple sequence alignment	RFAM
Jinfeng Liu et al. [24], 2006	CONC [24]	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	multiple sequence alignment	RNAdb, NONCODE, FANTOM3

3.1.2 Random Forest

Random forests are learning methods, used for regression and classification. They create many decision trees during the training process and store the class in the output. Random forest outputs mode of the classes in classification and individual tree's mean prediction in regression[28]. This technique also incorporates decision trees. The methods that implements random forest for classification of non-coding RNA are provided in Table 2. Marasri Ruengjitchatchawalya et al.[11] proposed a hybrid random forest tool for classifying non-coding RNA. It is a tool for classification based on hybrid random forest combined with a model of logistic regression to differentiate long as

well as short non-coding RNA sequences. This method includes a new feature SCORE, built on logistic regression function combining five features, i.e. sequence, structure, structural robustness, coding potential and modularity. Hybrid random forest is a classifier built on ensemble of multiple decision trees and random forest. This technique uses datasets including Rfam[21], RefSeq[29], NCBI GenBank genome database and lncRNAdb[30] database. In this method, a total of 369 features are extracted for prediction of non-coding RNA. Out of these features, genetic algorithm and correlation-based feature selection are used to capture features having good predictive power. Logistic regression model is used to find relationships between the features. The sequence similarity is found from basic local alignment search tool (BLAST) [31]. Random forest acts as main classifier with decision trees as base classifiers. The ensemble of trees in the random forest (RF) can capture the heterogeneity of non-coding RNA subfamilies. The model is robust and uses composite feature that improves the performance of the classifier. This technique is used for classification of both known and unknown non-coding RNA. Yanni Sun et al. [32] proposed a method lncRNA-ID for identifying long non-coding RNA using balanced random forests. This method incorporates multiple sequence alignment. The database used by this method is LncRNADisease database[33].

Table 2: Studies related to ncRNA classification using Random Forest

Author, citation and year	Method	Features	Alignment	Database
Marasri Ruengjitchatchawalya et al.[11], 2014	Hybrid random forest[11]	sequence, structure, structural robustness, modularity and coding potential	multiple sequence alignment	Rfam, RefSeq, NCBI GenBank genome database and lncRNAdb database.
Milad Miladi et al. [32], 2017	lncRNA-ID[32]	open reading frame (ORF), protein conservation and ribosome interaction	profile hidden Markov model (profile HMM)-based alignment	LncRNADisease database[33]

3.2 Unsupervised Learning

Unsupervised learning algorithms take a data set that comprises only inputs and then do clustering or grouping of the data points [34]. The algorithm learns from unlabelled data, data that is not classified or clustered. It identifies similarities between the data and make clusters based on the similarities. Unsupervised learning is used in density estimation in statistics [35]. Clustering is the allocation of datasets into subsets called clusters. The data of the same cluster is similar conferring to one or more criteria and data from different clusters are not similar. Every clustering technique derive different assumption from the data's structure that is demarcated by a similarity metric.

Similarity metrics is difference between the clusters and internal compactness is similarity between data of same cluster. Many methods use graph connectivity and estimated density as similarity metrics. The techniques that use unsupervised learning are hierarchical clustering, deep sequencing and artificial neural networks like convolutional neural network (CNN) and recurrent neural network (RNN).

3.2.1 Hierarchical Clustering

Hierarchical clustering is a category of cluster analysis that create a hierarchy of clusters. There are two kinds of hierarchical clustering – divisive and agglomerative [36]. Divisive clustering is a top-down approach, where all observations are started on clusters and splits are moved hierarchically, moving down the hierarchy. Agglomerative clustering is a bottom up approach, where every observation is started in its individual cluster and cluster's pairs are merged, moving up the hierarchy. The merging and splitting are done by greedy approach. The results of clustering are represented by dendrograms. The methods that implements hierarchical clustering for classification of non-coding RNA are provided in Table 3. Yasubumi Sakakibara et al. [14] proposed a method EnsembleClust for clustering non-coding RNA. EnsembleClust provides a innovative method for hierarchical clustering of non-coding RNA that could be used to identify new non-coding RNA families [14]. It helps in exploring the functional diversity of non-coding RNA. EnsembleClust implements an unsupervised learning algorithm [14]. This method takes the input as unlabelled data and build clusters of the non-coding RNA. Non-coding RNAs are clustered based on structural alignment scores. But computational cost of structural alignment is high, so approximate algorithms are employed. It uses all possible sequence alignments and secondary structures. There are many unknown non-coding RNAs, so clustering is very important to identify new types of non-coding RNA. This technique incorporates pair-wise alignment of non-coding RNA sequence. For accurate clustering, a reliable measure is used that would take into interpretation both secondary structures and primary sequences. The similarity score is obtained from these measures. Previously used methods were LocARNA [37] and FOLDALIGN [38] that calculated similarity based on one optimal structural alignment's score. The scoring function for the sequence alignment is designed using all the possible secondary structures. This method uses Waterman algorithm[39] for sequence alignment and McCaskill algorithm [40] for secondary structure. These both algorithms provide sequence alignment faster than the previously used Sankoff's algorithm[41]. EnsembleClust provides better performance than LocARNA v1.5.2 [37], FOLDALIGN v2.1.1 [38], and Stem kernel v216c [42]. With high accuracy, EnsembleClust provides balance between clustering quality and computational cost. Milad Miladi et al. [13] proposed a method RNAscClust for identification of non-coding RNA. RNAscClust is used for grouping RNA sequences using graph-based motifs and structure conservation [43]. This technique groups paralogous RNAs according to the structural similarities. RNAscClust incorporates

multiple alignment of RNA sequences. This technique finds minimum free energy structures for every sequence, as a preceding information for the folding. Then clustering of paralogs is done using graph kernel-based strategy identifying the common structural features. The RNA structures are set as graphs and graph kernels generate sparse feature vectors, creating a pairwise similarity notion. Increasing the degree of compensatory base pair changes in the alignments improves the clustering accuracy. Iterative clustering is used, creating more and more accurate feature vectors after every iteration. RNAscClust allows millions of occurrences to be clustered. The sequence is converted into graph where each nucleotide is taken as vertices with labels A, U, G, C with the base pair relations and backbone being encoded as edges. Base pair stacks are represented by adding base pair vertices adjacent to the existing base pair vertices. The structures are compared using graph kernels. This method also takes into account the base pair changes that were not being considered by many clustering approaches. RNAscCLust uses the RFam database of non-coding RNA. It uses neighbourhood subgraph pairwise distance kernel (NSPDK) [44] to extract sparse feature vectors. This method provides accurate clustering with linear runtime which makes alignment of large clusters possible.

Table 3: Studies related to ncRNA classification using Hierarchical Clustering

Author, citation and year	Method	Features	Alignment	Database
Yasubumi Sakakibara et al.[14], 2011	EnsembleClust [14]	structural alignments score	Pairwise sequence alignment	ENSEMBLE
Milad Miladi et al. [13], 2017	RNAscCLust [13]	structure conservation and graph-based motifs	-	RFAM

3.2.2 Artificial Neural Networks (ANNs)

Artificial neural networks (ANNs) are computing systems, encouraged by the biological neural network that constitutes brains of humans or animals [36]. ANN is a framework of several machine learning algorithms that function together and process many intricate data inputs. These systems learn by examples i.e. training data, without being programmed to perform the task with task-specific rules. ANN consists of collection of connected nodes, known as artificial neurons. The learning process is adjusted by the weights of neurons and edges. The neurons are combined into layers. Signals transmit from the first layer to the last layer, crossing the hidden layers in between. Artificial neural networks are used in speech recognition, computer vision, video games, medical diagnosis and social network filtering. Deep learning implements many hidden layers in an artificial neural network. Some of its applications are computer vision and speech recognition [37]. CNN and RNN are types of artificial neural networks that are used for classification of non-coding RNA.

3.2.2.1 Convolutional Neural Networks (CNNs)

Convolutional neural networks (CNNs) are class of deep neural networks and are mostly used in image processing. They incorporate a variation of multilayer perceptron and helps in minimizing pre-processing. CNNs are used in recommender systems, image recognition, image classification, natural language pre-processing, classification of ncRNA, medical image analysis etc. They comprise of input layer, output layer and intermediate layers i.e. hidden layers. The hidden layers contain RELU layer, pooling layers, pooling layers, convolutional layers, normalization layers and fully connected layers. The convolutional layer applies convolution operation to the input before transmitting to the next layer. The convolution operation tries to simulate a neuron's response to visual stimuli. The methods that implements CNN for classification of non-coding RNA are provided in Table 4. Yasubumi Sakakibara et al. [16] proposed a method CNNClust for clustering non-coding RNA. This technique incorporates pair-wise alignment of non-coding RNA sequences. Derived position weight matrices of sequence motifs are used for training of convolutional neural network. Two types of distributed representation are used by CNNClust i.e. one hot coding and word2vec. Secondary structure information and read mapping are also used. Similarity score matrix is calculated for every pair of the RNA sequences. The clustering is done to form clusters of similar structures. CNNClust clusters non-coding RNA into positive or negative class. When both sequences are of same class then it is a positive class, otherwise it is a negative class. Many new types of snoRNA, microRNA and tRNA are found by this method. The databases used by this method are Rfam, HUGO gene nomenclature committee (HGNC) databases, Ensembl and genomic tRNA database (GtRNAdb). Antonino Fiannaca et al. [47] proposed a method nRC for classification of non-coding RNA. This method uses secondary structure features for detection of non-coding RNA. It incorporates multiple sequence alignment. The database used by this method is Rfam.

Table 4: Studies related to ncRNA classification using CNN

Author, citation and year	Method	Features	Alignment	Database
Yasubumi Sakakibara et al.[16], 2018	CNNClust [16]	Derived position weight matrices of sequence motifs	pairwise sequence alignment	Rfam, HUGO gene nomenclature committee (HGNC) databases, Ensembl and genomic tRNA database (GtRNAdb)
Antonino Fiannaca et al.[47], 2017	nRC[47]	Secondary structure features	Multiple sequence alignment	Rfam

3.2.2.2 Recurrent Neural Network (RNN)

Recurrent neural network (RNN) is a type of artificial neural network where the connections among the nodes make a directed graph along a sequence [38]. RNNs also use the memory to transmit the sequences of the input. RNNs are used in speech recognition, handwriting recognition, grammar learning, robot control and human action recognition. These networks have many layers and each node in a layer is connected by directed connections to the node of the next layer. Every node has a real valued time-varying activation and every connection has a weight that has a real value. The nodes are either in input layer, output layer or hidden layer. The methods that implements RNN for classification of non-coding RNA are provided in Table 5. Sungroh Yoon et al. proposed a method lncRNAnet for classifying non-coding RNA. lncRNAnet identifies long non-coding RNA by means of deep learning and next generation sequencing[17]. Both recurrent neural networks (RNN) and convolutional neural network (CNN) are used in this method. RNN is used for RNA sequence modelling and CNN is used for spotting stop codons to find an open reading frame (ORF) indicator. lncRNAnet performs well for short length sequences. This method classifies lncRNA from protein-coding transcripts. The previous methods relied heavily on the features extracted from the known long non-coding RNA, their genomic profiles attained from database searches and multiple sequence alignments (MSA). lncRNAnet learns intrinsic features by RNN for RNA sequence modelling. The RNN uses backpropagation through time (BPTT) and one-hot coding scheme, the sequence is pre-processed for ORF indicator and transcript sequence. In one-hot coding scheme, each nucleotide, i.e., A, U, G, C is encoded as four-dimensional binary vectors. lncRNAnet uses datasets of GENCODE, ENSEMBL and Human and Vertebrate Analysis and Annotation (HAVANA) group databases. This technique provides robust performance irrespective of variation of sequence length and helps in identifying new lncRNA from the ample transcriptome data. Sungroh Yoon et al.. [49] proposed a method deep RNN for classification of non-coding RNA. This method uses secondary structure features for identifying non-coding RNA. It incorporates pairwise sequence alignment. The databases used in this method are NCBI, fRNAdb and NON-CODE.

Table 5: Studies related to ncRNA classification using RNN

Author, citation and year	Method	Features	Alignment	Database
Sungroh Yoon et al. [17], 2018	lncRNAnet [17]	Open reading frame (ORF) indicator	multiple sequence alignment	GENCODE, ENSEMBL and Human and Vertebrate Analysis and Annotation (HAVANA) group databases
Sungroh Yoon et al. [49], 2018	Deep RNN[49]	Secondary sequence features	pairwise sequence alignment	NCBI, fRNAdb, NON-CODE

3.2.3 Deep Sequencing

Deep sequencing refers to the concept of targeting for maximum unique reads of each section of a sequence [39]. This technique is also known as next generation sequencing. Deep sequencing is helping the researchers to detect rare microbes or cells that comprise of as small as 1% of the original sample. It is used in microbial genomics, oncology, cancer research and other researches that involve rare cell populations. The methods that implements deep sequencing for classification of non-coding RNA are provided in Table 6. Yasubumi Sakakibara et al. [15] proposed a method SHARAKU which implements deep sequencing for classification of non-coding RNA. SHARAKU incorporates a new algorithm that aligns two read mapping profiles of non-coding RNA's next generation sequencing data. This method implements a read mapping profile alignment program that uses decomposition for aligning and folding RNA sequences simultaneously (DAFS) program [51]. The read mapping profiles allows common processing patterns to be detected. Sequence and secondary structure information are taken simultaneously in this method. The sequences are read from BAM format that is a binary format for storing sequence data and also helps in compressing the data. This technique helps in finding non-coding RNAs articulated in the brain, more specifically, hippocampus of the left brain. This method uses NCBI Reference sequence database, ENSEMBLE database and next generation sequencing output. SHARAKU achieves higher accuracy than deepBlockAlign [40]. SHARAKU can only be implemented to labelled non-coding RNA sequences. Rosemarie Weikard et al. [53] proposed a method of deep next generation sequencing for classification of non-coding RNA. This method uses protein coding features to differentiate between coding and non-coding RNA. It incorporates pairwise sequence alignment. The databases used in this method are lncRNA, NCBI and NONCODE.

Table 6: Studies related to ncRNA classification using Deep Sequencing

Author, citation and year	Method	Features	Alignment	Database
Yasubumi Sakakibara et al. [15], 2016	SHARAKU [15]	Similarity score matrix	Pairwise sequence alignment	NCBI Reference sequence database, ENSEMBLE database and next generation sequencing output
Rosemarie Weikard et al.. [53], 2013	Deep next generation sequencing [53]	Protein coding features	Pairwise sequence alignment	lncRNA, NCBI, NONCODE

3.3 Reinforcement Learning

Reinforcement learning is the type of machine learning that takes action to maximize reward in a particular solution[54]. This learning is different from supervised learning. In supervised learning, there is a desired output for the corresponding inputs. While in reinforcement learning, there is no desired output for the corresponding input, but the reinforcement agent decides what to do to perform the task. It learns by experience as there is no training dataset. The decisions are made sequentially, i.e., current output depends upon the current input state and the next input state depends upon the output of the previous input. Most of the reinforcement algorithms use dynamic programming. They are used in information theory, genetic algorithms, simulation-based optimization, operations research, control theory, game theory, swarm intelligence and statistics[55]. They are implemented when exact mathematical model is not feasible.

4. Comparative Analysis between different Machine Learning Techniques used for Classification of Non-Coding RNA

In this section, comparative analysis between different machine learning techniques used for classification of non-coding RNA is provided. The comparison is done on the basis of learning algorithm, advantages, disadvantages and performance metrics. Different techniques perform well in particular situation and the type of dataset used. The link for the online resource tool is also provided wherever available. The input format is fasta in most of the methods that implement these machine learning techniques. The comparison is presented in Table 7.

Table 7: Comparative Analysis of Different Machine Learning Techniques used for Classification of Non-Coding RNA.

Technique used	Author, citation and year	Method	Learning Algorithm	Advantages	Shortcomings	Online resource tool	Input Format	Performance metrics
SVM	Peter F. Stadler et al.. [10], 2005	RNAz	Supervised	can detect variety of ncRNAs by using only thermodynamic stability and evolutionary conservation, suitable for large scale genomic annotation, predicts accurate model of consensus structure.	tmRNA and U70 snoRNA are difficult to detect, represents two diagnostic features that are not definite for a particular class of ncRNA, can be replaced by a direct statistical model, cannot identify new families of non-coding RNA from unannotated transcriptions	www.tbi.univie.ac.at/~wash/RNAz.	.maf and .aln	Thermodynamic stability measure and consensus secondary structure
	Jinfeng Liu et al.. [24], 2006	CONC	Supervised	accounts for the sequencing errors and can capture protein irrespective of the error's position and type of error	doesn't perform well on novel transcripts	-	.fasta	Area under the ROC curve

Random forest	Marasri Ruengjitchatchawalya et al.. [11], 2014	Hybrid random forest	Supervised	is robust due to random selection of features, provides high performance in identifying new ncRNAs, provides high accuracy for both known and unknown ncRNAs.	The framework is may not be completely accurate to incorporate new ncRNA families.	http://ncrna-pred.com/HLRF.htm	.fasta, sequence of variable length ranging from 75 to 200nt	SCORE based on -sequence, structure, structural robustness, modularity and coding potential
	Yanni Sun et al. [32], 2015	lncRNA-ID	Supervised	has high sensitivity, takes assistance of alignment-based features having good discriminative power, has shorter running time, easy to use, does not require a large number of training data	cannot work on unlabelled data, the performance and accuracy can further be improved.	https://github.com/zhangy72/LncRNA-ID	.fasta	Score cut-off of the dataset
Hierarchical Clustering	Yasubumi Sakakibara et al.[14], 2011	EnsembleClust	Unsupervised	utilizes all possible secondary structures and sequence alignments, performs clustering to identify new non-coding RNA from unannotated transcriptions, provides balance between quality of clustering and computational cost, accuracy is good even when sequence identity is below 60%, can decrease human labour costs of clustering.	performance can further be improved.	http://bpla-kernel.dna.bio.keio.ac.jp/clustering/	.mfa format-Multi fasta file	Structural alignment score

	Milad Miladi et al. [13], 2017	RNAscCLust	Unsupervised	has linear runtime that makes alignment of large clusters possible, do not need training data for clustering as it is unsupervised learning.	performance and accuracy can be further improved.	http://www.bioinf.uni-freiburg.de/Software/RNAscClust	.fasta	Average pairwise alignment score
CNN	Yasubumi Sakakibara et al.[16], 2018	CNNClust	Unsupervised	provides good performance and accuracy.	Better dataset can be used to achieve good results	http://www.dna.bio.keio.ac.jp/cnn/	.fa, .npy	Position weight matrix (PWM), read mapping profiles
	Antonino Fiannaca et al.[47], 2017	nRC	Unsupervised	provide good estimates with respect to both accuracy and speed, handles overfitting, has low standard deviation	execution time can further be improved.	https://github.com/ICarPA-TBlab/nrc/	.fasta	
RNN	Sungroh Yoon et al.[17], 2018	lncRNAnet	Unsupervised	provides robust performance irrespective of variation of sequence length, helps in identifying new long non-coding RNA, successfully detects shorter lncRNAs, help in identification of new ncRNAs.	meaning of the feature is hard to understand.	http://data.snu.ac.kr/pub/lncRNAnet	.fa, .h5	Open reading frame (ORF) indicator

	Sungroh Yoon et al.[49], 2018	Deep RNN	Unsupervised	Has good generalisation ability, handles overfitting effectively.	features have no explicit significance, learning secondary structure from input sequence is difficult as most ncRNA are not known.	https://github.com/eleventh83/deepMiRGene	.fasta	Secondary structure information
Deep sequencing	Yasubumi Sakakibara et al.[15], 2016	SHARAKU	Unsupervised	helps in detecting non-coding RNAs expressed in the brain, not only detects whole expression patterns of non-coding RNA sequence but also fragments due to splicing of the RNAs.	could be implemented only on labelled ncRNA sequences	http://www.dna.bio.keio.ac.jp/sharaku/	.pm, .sge, .fa	Computes minimum free energy structures, graph kernels, similarity score matrix
	Rosemarie Weikard et al.[53], 2013	Deep next generation sequencing	Unsupervised	Is capable of deciphering unlabelled transcriptional activity by detecting new transcripts.	focus is only on unlabelled transcripts and not on labelled transcripts, still difficult to identify between coding and non-coding RNA	-	.fastq, .gtf, .bam	Sequence similarity

The comparison of the performance of techniques is done on the basis of accuracy, specificity, sensitivity and area under curve (AUC). This comparison is presented in Table 8.

Table 8: Comparison of Performance of Different Machine Learning Techniques used for Classification of Non-Coding RNA

Technique used	Author, citation and year	Implementation	Accuracy	Specificity	Sensitivity	AUC
SVM	Peter F. Stadler et al.[10], 2005	RNAz	0.7527	0.9893	0.7527	-
	Jinfeng Liu et al. [24], 2006	CONC	-	0.9520	0.9380	-
Random forest	Marasri Ruengjitchatchawalya et al.[11], 2014	Hybrid random forest	0.9211	0.9350	0.9070	-
	Yanni Sun et al. [32], 2015	lncRNA-ID	0.9578	0.9528	0.9628	-
Hierarchical clustering	Yasubumi Sakakibara et al.[14], 2011	EnsembleClust	-	-	-	0.944
	Milad Miladi et al. [13], 2017	RNAscCLust	-	-	-	-
CNN	Yasubumi Sakakibara et al.[16], 2018	CNNClust	0.9800	-	-	-
	Antonino Fiannaca et al.[47], 2017	nRC	0.8181	0.9848	0.8181	-
RNN	Sungroh Yoon et al.[17], 2018	lncRNAnet	0.9179	0.8766	0.9591	-
	Sungroh Yoon et al.[49], 2018	Deep RNN	-	0.9920	0.8220	-
Deep sequencing	Yasubumi Sakakibara et al.[15], 2016	SHARAKU	-	-	-	0.985
	Rosemarie Weikard et al.[53], 2013	Deep next generation sequencing	-	-	-	-

4. CONCLUSION

In this paper, various machine learning techniques used for classification of non-coding RNA are discussed. Every technique has its own merits and demerits. These techniques can be used based on the particular situation to get the desired results. Convolutional neural networks are good tools for identification of non-coding RNA. Many other techniques have been used for clustering new non-coding RNA, but performance accuracy still has scope for improvement. These techniques can be combined to get good results. Better non-coding database can also be used for training of these machine learning techniques, to improve the accuracy further. If more non-coding RNA families are available in the training data, then more accurate clustering of unknown families is achieved. The current trend is more on convolutional neural networks, as it is giving better accuracy and performance. This review can help the researchers to select a particular technique according to their need.

CONFLICT OF INTEREST

Authors have no any conflict of interest.

REFERENCES

1. Mattick JS, Makunin IV. Non-coding RNA. Human molecular genetics. 2006 Apr 15;15(suppl_1):R17-29.
2. Taft RJ, Pang KC, Mercer TR, Dinger M, Mattick JS. Non-coding RNAs: regulators of disease. The Journal of pathology. 2010 Jan 1;220(2):126-39.
3. Palazzo AF, Lee ES. Non-coding RNA: what is functional and what is junk?. Frontiers in genetics. 2015 Jan 26;6:2.
4. Treangen TJ, Salzberg SL. Repetitive DNA and next-generation sequencing: computational challenges and solutions. Nature Reviews Genetics. 2012 Jan;13(1):36.
5. He L, Hannon GJ. MicroRNAs: small RNAs with a big role in gene regulation. Nature Reviews Genetics. 2004 Jul;5(7):522.
6. Esteller M. Non-coding RNAs in human disease. Nature Reviews Genetics. 2011 Dec;12(12):861.
7. Mattick JS. Non-coding RNAs: the architects of eukaryotic complexity. EMBO reports. 2001 Nov 1;2(11):986-91.
8. Samuel AL. Some studies in machine learning using the game of checkers. IBM Journal of research and development. 2000 Jan;44(1.2):206-26.
9. Bishop CM, Pattern Recognition and Machine Learning. 2006.
10. Washietl S, Hofacker IL, Stadler PF. Fast and reliable prediction of noncoding RNAs. Proceedings of the National Academy of Sciences. 2005 Feb 15;102(7):2454-9.
11. Lertampaiporn S, Thammarongtham C, Nukoolkit C, Kaewkamnerdpong B, Ruengjitchatchawalya M. Identification of non-coding RNAs with a new composite feature in the Hybrid Random Forest Ensemble algorithm. Nucleic acids research. 2014 Apr 25;42(11):e93-.

12. Cortes C, Vapnik V. Support-vector networks. Machine learning. 1995 Sep 1;20(3):273-97.
13. Miladi M, Junge A, Costa F, Seemann SE, Havgaard JH, Gorodkin J, Backofen R. RNAscClust: clustering RNA sequences using structure conservation and graph based motifs. Bioinformatics. 2017 Feb 27;33(14):2089-96.
14. Saito Y, Sato K, Sakakibara Y. Fast and accurate clustering of noncoding RNAs using ensembles of sequence alignments and secondary structures. BMC bioinformatics. 2011 Dec;12(1):S48.
15. Tsuchiya M, Amano K, Abe M, Seki M, Hase S, Sato K, Sakakibara Y. SHARAKU: an algorithm for aligning and clustering read mapping profiles of deep sequencing in non-coding RNA processing. Bioinformatics. 2016 Jun 11;32(12):i369-77.
16. Aoki G, Sakakibara Y. Convolutional neural networks for classification of alignments of non-coding RNA sequences. Bioinformatics. 2018 Jun 27;34(13):i237-44.
17. Baek J, Lee B, Kwon S, Yoon S. Lncrnanet: long non-coding rna identification using deep learning. Bioinformatics. 2018 May 29;34(22):3889-97.
18. Russell SJ, Norvig P. Artificial intelligence: a modern approach. Malaysia; Pearson Education Limited,; 2016.
19. Mohri M, Rostamizadeh A, Talwalkar A. Foundations of Machine Learning. Adaptive computation and machine learning. MIT Press. 2012;31:32.
20. Mitchell TM. Machine Learning, McGraw-Hill Higher Education. New York. 1997.
21. Griffiths-Jones S, Bateman A, Marshall M, Khanna A, Eddy SR. Rfam: an RNA family database. Nucleic acids research. 2003 Jan 1;31(1):439-41.
22. Washietl S, Hofacker IL. Consensus folding of aligned sequences as a new measure for the detection of functional RNAs by comparative genomics. Journal of molecular biology. 2004 Sep 3;342(1):19-30.
23. Hofacker IL, Fekete M, Stadler PF. Secondary structure prediction for aligned RNA sequences. Journal of molecular biology. 2002 Jun 21;319(5):1059-66.
24. Liu J, Gough J, Rost B. Distinguishing protein-coding from non-coding RNAs through support vector machines. PLoS genetics. 2006 Apr 28;2(4):e29.
25. Pang KC, Stephen S, Engström PG, Tajul-Arifin K, Chen W, Wahlestedt C, Lenhard B, Hayashizaki Y, Mattick JS. RNAdb—a comprehensive mammalian noncoding RNA database. Nucleic acids research. 2005 Jan 1;33(suppl_1):D125-30.
26. Liu C, Bai B, Skogerbø G, Cai L, Deng W, Zhang Y, Bu D, Zhao Y, Chen R. NONCODE: an integrated knowledge database of non-coding RNAs. Nucleic acids research. 2005 Jan 1;33(suppl_1):D112-5.
27. Carninci P, Kasukawa T, Katayama S, Gough J, Frith MC, Maeda N, Oyama R, Ravasi T, Lenhard B, Wells C, Kodzius R. The transcriptional landscape of the mammalian genome. Science. 2005 Sep 2;309(5740):1559-63.

28. Ho TK. Random decision forests. InProceedings of 3rd international conference on document analysis and recognition 1995 Aug 14 (Vol. 1, pp. 278-282). IEEE.
29. Pruitt KD, Tatusova T, Brown GR, Maglott DR. NCBI Reference Sequences (RefSeq): current status, new features and genome annotation policy. Nucleic acids research. 2011 Nov 24;40(D1):D130-5.
30. Amaral PP, Clark MB, Gascoigne DK, Dinger ME, Mattick JS. lncRNAdb: a reference database for long noncoding RNAs. Nucleic acids research. 2010 Nov 25;39(suppl_1):D146-51.
31. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. Journal of molecular biology. 1990 Oct 5;215(3):403-10.
32. Achawanantakun R, Chen J, Sun Y, Zhang Y. LncRNA-ID: Long non-coding RNA IDentification using balanced random forests. Bioinformatics. 2015 Aug 26;31(24):3897-905.
33. Chen G, Wang Z, Wang D, Qiu C, Liu M, Chen X, Zhang Q, Yan G, Cui Q. LncRNADisease: a database for long-non-coding RNA-associated diseases. Nucleic acids research. 2012 Nov 21;41(D1):D983-6.
34. Hinton GE, Sejnowski TJ, Poggio TA, editors. Unsupervised learning: foundations of neural computation. MIT press; 1999.
35. Tucker AB, Computer science handbook. 2004.
36. Maimon O, Rokach L, editors. Data mining and knowledge discovery handbook,2005.
37. Will S, Reiche K, Hofacker IL, Stadler PF, Backofen R. Inferring noncoding RNA families and classes by means of genome-scale structure-based clustering. PLoS computational biology. 2007 Apr 13;3(4):e65.
38. Havgaard JH, Torarinsson E, Gorodkin J. Fast pairwise structural RNA alignments by pruning of the dynamical programming matrix. PLOS computational biology. 2007 Oct 12;3(10):e193.
39. Smith TF, Waterman MS. Identification of common molecular subsequences. Journal of molecular biology. 1981 Mar 25;147(1):195-7.
40. McCaskill JS. The equilibrium partition function and base pair binding probabilities for RNA secondary structure. Biopolymers: Original Research on Biomolecules. 1990 May;29(6-7):1105-19.
41. Sankoff D. Simultaneous solution of the RNA folding, alignment and protosequence problems. SIAM journal on applied mathematics. 1985 Oct;45(5):810-25.
42. Sato K, Mituyama T, Asai K, Sakakibara Y. Directed acyclic graph kernels for structural RNA analysis. BMC bioinformatics. 2008 Dec;9(1):318.
43. Miladi M, Junge A, Costa F, Seemann SE, Havgaard JH, Gorodkin J, Backofen R. RNAscClust: clustering RNA sequences using structure conservation and graph based motifs. Bioinformatics. 2017 Feb 27;33(14):2089-96.
44. Costa F, Grave KD. Fast neighborhood subgraph pairwise distance kernel. InProceedings of the

27th International Conference on International Conference on Machine Learning 2010 Jun 21 (pp. 255-262). Omnipress.

45. van Gerven M, Bohte S, editors. Artificial neural networks as models of neural information processing. Frontiers Media SA; 2018 Feb 1.
46. Lee H, Grosse R, Ranganath R, Ng AY. Convolutional deep belief networks for scalable unsupervised learning of hierarchical representations. InProceedings of the 26th annual international conference on machine learning 2009 Jun 14 (pp. 609-616). ACM.
47. Fiannaca A, La Rosa M, La Paglia L, Rizzo R, Urso A. nRC: non-coding RNA Classifier based on structural features. BioData mining. 2017 Dec;10(1):27.
48. Graves A, Liwicki M, Fernández S, Bertolami R, Bunke H, Schmidhuber J. A novel connectionist system for unconstrained handwriting recognition. IEEE transactions on pattern analysis and machine intelligence. 2009 May;31(5):855-68.
49. Park S, Min S, Choi HS, Yoon S. Deep recurrent neural network-based identification of precursor micrornas. InAdvances in Neural Information Processing Systems 2017 (pp. 2891-2900).
50. Mardis ER. Next-generation DNA sequencing methods. Annu. Rev. Genomics Hum. Genet.. 2008 Sep 22;9:387-402.
51. Sato K, Kato Y, Akutsu T, Asai K, Sakakibara Y. DAFS: simultaneous aligning and folding of RNA sequences via dual decomposition. Bioinformatics. 2012 Oct 11;28(24):3218-24.
52. Langenberger D, Pundhir S, Ekstrøm CT, Stadler PF, Hoffmann S, Gorodkin J. deepBlockAlign: a tool for aligning RNA-seq profiles of read block patterns. Bioinformatics. 2011 Nov 3;28(1):17-24.
53. Weikard R, Hadlich F, Kuehn C. Identification of novel transcripts and noncoding RNAs in bovine skin by deep next generation sequencing. BMC genomics. 2013 Dec;14(1):789.
54. Bertsekas DP, Bertsekas DP, Bertsekas DP, Bertsekas DP. Dynamic programming and optimal control. Belmont, MA: Athena scientific; 1995 Jan.
55. Bertsekas DP, Tsitsiklis JN. Neuro-dynamic programming. Belmont, MA: Athena Scientific; 1996 Jan.