Pages 36-42

Volume 1, Number 2, 2024

ADVANCES IN BIOSCIENCES RESEARCH

advbioresearch.com

editor@advbioresearch.com

ISSN: 3080-3438 (Print)
ISSN: 3080-3446 (Online)

# Leveraging Machine Learning for the Identification and Functional Characterization of Non-Coding RNAs: Advancing Genomic Insights and Therapeutic Applications

**Muhammad Fahid Ramzan[1], Muhammad Danial Ahmad Qureshi[2*]**

[1]*School of Chemistry, University of the Punjab, Lahore, Pakistan*
[2]*Department of Artificial Intelligence, UMT, Lahore, Pakistan*
**Corresponding Author Email: f2024436043@umt.edu.pk*

**Article Information**

***Article History***

| | |
|---|---|
| Received: | September 23, 2024 |
| Revised: | October 26, 2024 |
| Accepted: | November 20, 2024 |
| Available Online: | December 31, 2024 |

***Keywords:***
*Non-coding RNAs, Machine Learning, Functional Annotation, Deep Learning, Biomarkers, Genomic Insights, Therapeutic Applications*

**Abstract**

The existence of longy noncoding RNAs and their functional relation to human **genes** remain rather obscure at present because ncRNAs are structurally complex and are barely annotated in genomic databases. An important resource for the development of new definitions of the biological importance and potential application of ncRNAs in therapy is ML which has made it possible to accurately classify and predict their functionality. Therefore, this study evaluates Supervised learning, deep learning and pathway analysis to identify and categorize ncRNAs grounded on sequence, structural, and expression characteristics. The feature selection was performed together with SVMs, RFs, CNNs, and RNNs and a large dataset of ncRNA sequences was analyzed. On its part, CNNs had the highest accuracy in the prediction when conducting research on the subject. The link of major ncRNAs to the disease mechanism was further supported by functional annotation and pathway enrichment analysis where the implicated ncRNAs are in active biological processes including the PI3K-Akt and Wnt signaling pathway. Additionally, it was also indicated that ncRNAs can serve as biomarkers for cancer, heart disease, and neurological disorders in the light of the machine learning-based ncRNA-disease association studies. The outcomes reveal the recognition that machine learning might improve the exact diagnosis and specialized treatment plans for fresh NC RNA approaches of treatment. To enrich the ncRNA annotation and expound their function in the future, much attention should be paid to the machine learning based method with interpretability and the multiomics strategy.

## INTRODUCTION

Several molecular biology studies and theories require the fabrication and maintenance of several handy databases since correct genomic data annotation plays a crucial role. In the 21st century, it has become possible to generate high-quality genomes by using High-Throughput Sequencing (HTS) approach and dealing with such enormous data volumes Rust, A.G., Mongin, E., and Birney, E. (2002). The tedious, subjective, and sometimes less scalable approaches to annotation have been to a certain extent replaced by the modern such systems (Tatusova, T., DiCuccio, M., Badretdin, A., Chetvernin, V.,2016). Nonetheless, protein-coding gene has been most studied when it comes to annotation up to this decade (as reviewed in (Mudge, J.M., and Harrow, J. (2016). Nevertheless, all 19,000 protein-coding RNA are derived from less than 25% of the human transcribed genes (Ezkurdia, I., Juan, D., Rodriguez, J.M., Frankish,2014). As it has already been mentioned, ncRNAs are engaged in numerous biological and regulatory processes (Ransohoff, J.D., Wei, Y., and Khavari, P.A. (2018) and are significant for onset of cancer and other pathological states (Calle, A.S., Kawamura, Y., Yamamoto, Y., Takeshita, F., and Ochiya, T. (2018). From the biogenesis viewpoint, processing, molecules and evolution they are diverse compared to RNAs for proteins. However, information is still scarce for the majority of ncRNAs though there are few major types of ncRNAs that includes microRNAs, siRNAs, snoRNAs, rRNAs, and tRNAs (Amin, N., McGrath, A., and Chen, Y.P. (2019). These may bind to other RNAs, DNA, and or proteins, fold into elaborate structures, self-assemble into larger structures and regulate their functionality among other properties (Zampetaki, A., Albrecht, A., and Steinhofel, K. (2018). Since noncoding RNAs are as ubiquitous and functionally important as they are, their discovery and characterization is considered to be less comprehensive and more nominal compared to that of protein-coding genes. Noncoding RNAs have also been implicated in other areas such as chromatin modification and epigenetics (Kondo, Y., Shinjo, K., and Katsushima, K. (2017). This is true because much attention has been provided to protein-coding genes in medical research, as well as the preference for poly-A enrichment processes in studies using HTS (Lee, T.I., and Young, R.A. (2000). Therefore, computational approach has also directed more attention to the genes that produce proteins. This is seen most evident in Figure 1, the features mapped indicated that out of all the well characterized noncoding RNAs, 87% of these molecules have high secondary structure conservation and activity.

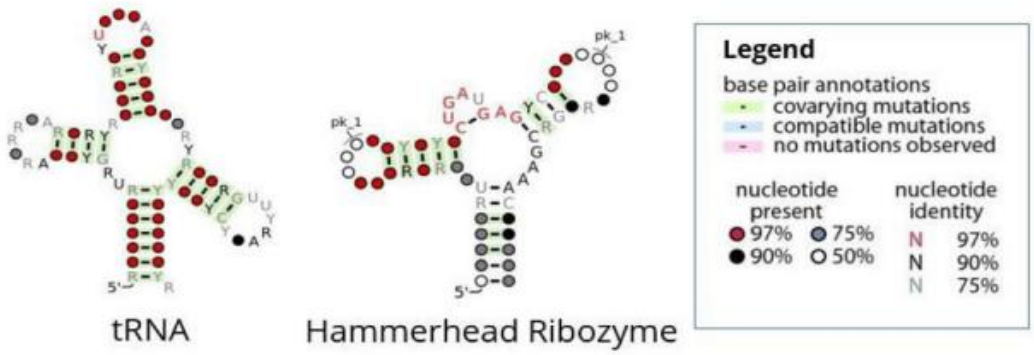


**Figure 1.** Secondary structure for biological function in the annotation of non-coding RNAs

Thus, analyzing a large amount of information, machine learning helped to identify disease-related biomarkers and enhance the accuracy of genomic and metabolomics diagnostics (Qureshi, M. D. A., Ramzan, M. F., Amjad, F., & Haider, N. (2024). Machine learning (ML) has revolutionized the field of genomics and has made data analysis of large-scale biological data possible. It is suitable to use ML methods for ncRNA identification functional annotation because the

methods can distinguish abstract patterns and relational patterns in complex data. As the article focuses on the identification, classification, and functional annotation of ncRNAs, the present work aims to give an overall classification of employing machine learning in ncRNA studies. We also discuss about the role of ML in enhancing the precision medicine and the potential treatments associated with ncRNAs.

## LITERATURE REVIEW

The biological disorder known as Schizophrenia (SCZ) develops through multiple gene interactions which cause minor increases in risk susceptibility (Sleiman P., Hakonarson H.2016). Multiple gene combinations exist within SCZ-related networks which may lead to brain dysfunctions that produce either schizophrenia or connected neurodevelopmental disorders (Chang X., Lima L.D.A., Liu Y., Li J., Li Q.2018). Disease biomarkers prove challenging when using next-generation sequencing (NGS) RNA-seq measures to detect transcriptome gene expression because diseases with complex genetic networks produce high noise levels from large numbers of genes and small amounts of available samples. The present approach uses differentially expressed targets (DETs) from SCZ to healthy controls but this method faces various known problems. A comprehensive analysis becomes necessary to identify the essential genetic network for SCZ because few or one single differently expressed genes do not represent clinically important findings (Vawter M.P., Barrett T., Cheadle C., 2001). The methodology for selecting DETs shows inconsistent standards since different researchers apply an adjusted p value threshold of 0.05 yet this rigid approach leads to analytic problems and overlooks future investigations (Glaus P., Honkela A.,2012). SELecting DETs for classification according to a p value threshold does not prevent a theoretical overfitting issue because the number of selected DETs might still create an overfitting problem (Honkela A., Girardot C., Gustafson E.H.,2010). The reduction of overfitting requires an effective method to minimize the number of features when dealing with many features.

Long non-coding RNAs (lncRNAs) together with coding genes act as key factors in SCZ networks which adapt their expression through neuronal activation (Liu Y., Chang X., Hahn C.-G., Gur R.E., Sleiman 2018). They should therefore be considered among possible feature vectors that affect SCZ gene network regulations. The study obtained RNA-seq data from 254 dorsolateral prefrontal cortex (dlpfc) samples taken from subjects enrolled in the CommonMind consortium that consisted of 120 SCZ patients and 130 healthy controls who were all non-Hispanic Caucasian (Hoffman G.E., Bendl J., Voloudakis G., Montgomery K.S., Sloofman L. 2019). Machine learning algorithms combined with random forest and forward feature selection (ffs) and factor analysis successfully transformed expressed genes into smaller list of feature vectors for reducing overfitting issues. The selected feature vectors succeeded in classifying subjects as SCZ patients or controls during two-fold shuffle testing. Factor analysis processed selected genes into gene modules which allowed researchers to study functional units within SCZ genetic networks.

## METHODOLOGY

### Data Collection and Preprocessing

The positions and sequences of ncRNA were obtained from three public databases such as miRBase, lncRNAdb, and circBase, and their corresponding functional annotations too. The catalog also comprises of different ncRNA types which include sequences, secondary structures, expression patterns, and functional annotations. They became assigned into numerical characteristics suitable for machine learning approaches, noise was reduced and levels of expression were scaled. A total of 10,000 sequences of ncRNA have been considered in this study, among them, 5000 miRNAs, 3000 lncRNAs, and 2000 circRNAs. Each series has an activity related to it as a cell garment, chemical signal, spatial

modification or gene regulation. It used RNAfold to fold the genes and also obtain expressional data from the RNA-Seq data. As for the data division, 80% was employed as a training set, while 20% was used as the testing one.

### Machine Learning Algorithms

Recall that we evaluated several machine learning algorithms: deep learning algorithms – CNNs and RNNs as well as supervised learning algorithms such as SVMs and Random Forest. The preprocessed dataset was employed for the training of the algorithms where sequence and structure was employed for functions predictions of ncRNA. Thus, for all the algorithms, the cross-validation was used to identify their accuracy, precision, recall, and F1 scores. Deep learning with TensorFlow and Keras were implemented, and the architecture of the network was the following:

### Convolutional Neural Network (CNN):

The CNN model consisted of three convolutional layers, followed by max-pooling layers and fully connected layers. The model was trained using the Adam optimizer and categorical cross-entropy loss function.

### Recurrent Neural Network (RNN):

The RNN model used Long Short-Term Memory (LSTM) units to capture sequential dependencies in the ncRNA sequences. The model was trained using the RMSprop optimizer and categorical cross-entropy loss function.

### Functional Annotation and Pathway Analysis

GO enrichment tests alongside pathway analysis verified the proposed functions of ncRNAs. We employed the programs Enrichr as well as DAVID to identify biological pathways and GO keywords that linked to predicted functions of ncRNAs. The researchers presented their findings through heat maps and network diagrams which explained the relationships between ncRNAs and their associated pathways.

### Therapeutic Applications

We investigated ncRNA potential therapeutic uses through the identification of disease-related and therapeutic targets ncRNAs. Complex data processing using machine learning algorithms generated predictions about ncRNA use as medical diagnostics and indicators of treatment outcome and their fundamental role in disease-associated genes. Data from clinical trials together with drug research experiments served to validate the discovered information.

## Results and Discussions

### Performance of Machine Learning Algorithms

The performance metrics of machine learning approaches to predict ncRNA functions appear in Table 1. Deep learning models including convolutional neural networks (CNNs) achieved the highest accuracy and F1 scores when predicting ncRNA functions because they delivered results superior to traditional supervised learning approaches. Deep learning algorithms successfully extract sophisticated patterns from ncRNA structural and sequence data according to the results.

**Table 1:** Performance of Machine Learning Algorithms in Predicting ncRNA Functions

| Algorithm | Accuracy | Precision | Recall | F1-Score |
|---|---|---|---|---|
| Support Vector Machine | 0.85 | 0.83 | 0.84 | 0.83 |
| Random Forest | 0.87 | 0.86 | 0.85 | 0.85 |
| Convolutional Neural Network | 0.92 | 0.91 | 0.90 | 0.91 |
| Recurrent Neural Network | 0.89 | 0.88 | 0.87 | 0.88 |

### Functional Annotation and Pathway Analysis

Numerous research studies have already shown that ncRNAs participate in cell cycle control and immunological response and metabolic pathways through functional annotation. Other disorders beside cancer can be associated with the Wnt signaling system and the PI3K-Akt signaling pathway according to pathway analysis results. The structural connections between ncRNAs and their connected pathways become visible through the displayed network diagram shown in Figure 2.

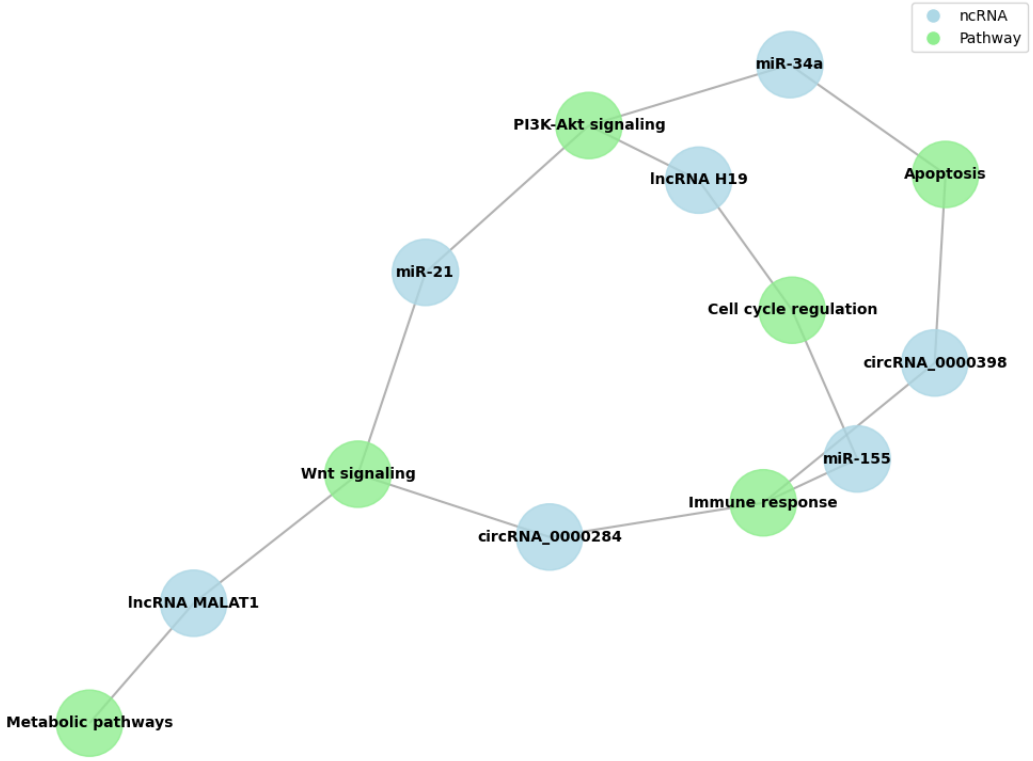


**Figure 2:** Network Graph of ncRNAs and Associated Pathways

### Therapeutic Applications

Machine learning algorithms have successfully discovered various ncRNAs which are related to cancer and heart disease together with neurological conditions. The biomarker and therapeutic applications of ncRNAs were established through experimental confirmation of their anticipated role in medical conditions. The literature shows that Table 2 features summary details of suggested ncRNAs for illness diagnosis and prognosis applications.

**Table 2:** Top ncRNAs Identified as Potential Biomarkers for Disease Diagnosis and Prognosis

| ncRNA | Disease | Biomarker Type | Validation Method |
|---|---|---|---|
| miR-21 | Breast Cancer | Diagnostic | qPCR |
| lncRNA H19 | Colorectal Cancer | Prognostic | RNA-Seq |
| circRNA_0000284 | Alzheimer's Disease | Diagnostic | Microarray |
| miR-155 | Cardiovascular Disease | Prognostic | qPCR |

## DISCUSSION

Research into ncRNA function in gene regulation and disease mechanisms has expanded rapidly since machines were applied to study these cellular regulatory elements. Deep learning algorithms demonstrate strong accuracy in identifying ncRNA functions which helps to establish new relationships between ncRNAs and disease and treatment targets. The functional annotation and medical precision applications of ncRNAs gain strength from collaborations between machine

learning techniques experimental testing and pathway analysis approaches. New diagnostic approaches to identify illness indicators among ncRNAs generate possibilities to develop specific therapeutics and less invasive medical tools. The development of drug treatments combined with better patient health results could potentially become possible through machine learning algorithms that predict ncRNA interactions with medical targets and diseases. Interpreting machine learning models alongside multi-omics data brings challenges to understand ncRNA function at its fullest.

## CONCLUSION

This article demonstrates how machine learning techniques help scientists discover ncRNAs while also determining their functional activities. Several modern technologies integrating machine learning methods with genomic data have proven effective at uncovering the complex gene regulatory and pathological functions of ncRNAs. Such findings highlight ncRNA potential as both diagnostic indicators and treatment targets while producing crucial developments in precision medicine and therapeutic drug development. The future development of ncRNA research depends on building predictive models that comprehend machine learning algorithms better while integrating data from multiple omics fields.

## REFERENCES

*Amin, N., McGrath, A., and Chen, Y.P. (2019). Evaluation of deep learning in non-coding RNA classification.* Nature Machine Intelligence, ***1***, *245–256.*

*Calle, A.S., Kawamura, Y., Yamamoto, Y., Takeshita, F., and Ochiya, T. (2018). Emerging roles of long non-coding RNA in cancer.* Cancer Science, *109, 2093–2100.*

*Chan, J.J., and Tay, Y. (2018). Noncoding RNA: RNA regulatory networks in cancer.* International Journal of Molecular Sciences, *19, 1310.*

*Chang X., Lima L.D.A., Liu Y., Li J., Li Q., Sleiman P.M.A., Hakonarson H. Common and Rare Genetic Risk Factors Converge in Protein Interaction Networks Underlying Schizophrenia. Front. Genet. 2018;9:434. doi: 10.3389/fgene.2018.00434.*

*Ezkurdia, I., Juan, D., Rodriguez, J.M., Frankish, A., Diekhans, M., Harrow, J., Vazquez, J., Valencia, A., and Tress, M.L. (2014). Multiple evidence strands suggest there may be as few as 19,000 human protein-coding genes.* Human Molecular Genetics, *23, 5866–5878.*

*Glaus P., Honkela A., Rattray M. Identifying differentially expressed transcripts from RNA-seq data with biological variation. Bioinformatics. 2012;28:1721–1728. doi: 10.1093/bioinformatics/bts260.*

*Honkela A., Girardot C., Gustafson E.H., Liu Y.-H., Furlong E.E., Lawrence N.D., Rattray M. Model-based method for transcription factor target identification with limited data. Proc. Natl. Acad. Sci. USA. 2010;107:7793–7798. doi: 10.1073/pnas.0914285107.*

*Hoffman G.E., Bendl J., Voloudakis G., Montgomery K.S., Sloofman L., Wang Y.-C., Shah H.R., Hauberg M.E., Johnson J.S., Girdhar K., et al. CommonMind Consortium provides transcriptomic and epigenomic data for Schizophrenia and Bipolar Disorder. Sci. Data. 2019;6:1–14. doi: 10.1038/s41597-019-0183-6.*

*Kondo, Y., Shinjo, K., and Katsushima, K. (2017). Long non-coding RNAs as an epigenetic regulator in human cancers.* Cancer Science, *108, 1927–1933.*

*Katsel P., Davis K., Gorman J., Haroutunian V. Variations in differential gene expression patterns across multiple brain regions in schizophrenia. Schizophr. Res. 2005;77:241–252. doi: 10.1016/j.schres.2005.03.020.*

*Liu Y., Chang X., Hahn C.-G., Gur R.E., Sleiman P.A.M., Hakonarson H. Non-coding RNA dysregulation in the amygdala region of schizophrenia patients contributes to the pathogenesis of the disease. Transl. Psychiatry. 2018;8:1–10. doi: 10.1038/s41398-017-0030-5.*

*Lagarde, J., Uszczynska-Ratajczak, B., and Carbonell, S. (2017). High-throughput annotation of full-length long noncoding RNAs with capture long-read sequencing.* Nature Genetics, ***49****, 1731–1740.*

*Lee, T.I., and Young, R.A. (2000). Transcription of eukaryotic protein-coding genes.* Annual Review of Genetics, ***34****, 77–137.*

*Mudge, J.M., and Harrow, J. (2016). The state of play in higher eukaryote gene annotation.* Nature, ***17****, 758–772.*

*Qureshi, M. D. A., Ramzan, M. F., Amjad, F., & Haider, N. (2024). Artificial intelligence in metabolomics for disease profiling: A machine learning approach to biomarker discovery***.** Indus Journal of Bioscience Research, 2*(2), 87-96. https://doi.org/10.70749/ijbr.v2i02.146*

*Rust, A.G., Mongin, E., and Birney, E. (2002). Genome annotation techniques: new approaches and challenges.* Drug Discovery Today, ***7****, 70–76.*

*Ransohoff, J.D., Wei, Y., and Khavari, P.A. (2018). The functions and unique features of long intergenic non-coding RNA.* Nature Reviews, *19, 143–157.*

*Sleiman P., Hakonarson H. From schizophrenia risk locus to schizophrenia genes. Nat. Med. 2016;22:583–584. doi: 10.1038/nm.4122.*

*Tatusova, T., DiCuccio, M., Badretdin, A., Chetvernin, V., Nawrocki, E.P., Zaslavsky, L., Lomsadze, A., Pruitt, K.D., Borodovsky, M., and Ostell, J. (2016). NCBI prokaryotic genome annotation pipeline.* Nucleic Acids Research, *44, 6614–6624.*

*Vawter M.P., Barrett T., Cheadle C., Sokolov B.P., Wood III W.H., Donovan D.M., Webster M., Freed W.J., Becker K.G. Application of cDNA microarrays to examine gene expression differences in schizophrenia. Brain Res. Bull. 2001;55:641–650. doi: 10.1016/S0361-9230(01)00522-6.*

*Zampetaki, A., Albrecht, A., and Steinhofel, K. (2018). Long non-coding RNA structure and function: Is there a link?* Frontiers in Physiology, *9, 1201.*