IJIRCCE

e-ISSN: 2320-9801 | p-ISSN: 2320-9798

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN COMPUTER & COMMUNICATION ENGINEERING

Volume 12, Issue 7, July 2024

INTERNATIONAL
STANDARD
SERIAL
NUMBER
INDIA

Impact Factor: 8.379

| DOI: 10.15680/IJIRCCE.2024.1207092|

Deploying Deep Learning Models to unlock the potential of Non-Coding Genomic Regions

Ravi Kumar S, Dr. Soumyasri SM

PG Student, Dept. of MCA, Visvesvaraya Technological University, The National Institute of Engineering, Mysuru India

Associate Professor, Dept. of MCA, Visvesvaraya Technological University, The National Institute of Engineering, Mysuru India

ABSTRACT: The analysis of non-coding genomic regions holds great promise for understanding complex biological processes and diseases. This study presents a system utilizing deep learning techniques to analyze these regions, providing insights into their functions and interactions. The system incorporates data acquisition, preprocessing, model training, and classification using various machine learning models. By developing advanced deep learning models, this research focuses on improving the accuracy of genomic analysis, particularly in identifying the functional roles of non-coding DNA. The ultimate goal is to facilitate more efficient genomic research, providing valuable insights for both clinical and scientific communities.

I. INTRODUCTION

Non-coding regions of the genome, once considered "junk DNA," are now recognized as crucial elements in gene regulation and other biological functions. Understanding these regions can lead to significant breakthroughs in genetics and medicine. This research aims to deploy deep learning models to classify and predict the functions of non-coding genomic sequences.

In the era of digital transformation, the role of advanced computational techniques in genomics has become indispensable. By integrating deep learning, this project seeks to unveil the functional roles of non-coding regions, offering a cohesive and intuitive analysis framework that bridges complex biological data and practical applications. By developing sophisticated deep learning models, we aim to enhance the precision of genomic analyses, enabling researchers to pinpoint the roles of non-coding DNA with greater accuracy and efficiency. This initiative holds the promise of revolutionizing genomic research, offering profound implications for both clinical applications and scientific inquiry.

II. OBJECTIVE

The project aims to leverage machine learning techniques to analyze non-coding regions of the genome. By developing advanced deep learning models, the research focuses on improving the accuracy of genomic analysis, particularly in identifying the functional roles of non-coding DNA. The ultimate goal is to facilitate more efficient genomic research, providing valuable insights for both clinical and scientific communities.

III. LITERATURE SURVEY

The Report Examine the body of the Knowledge regarding Deploying Deep Learning Models to Unlock the Potential of Non-Coding Genomic Regions, Relevant Studies on the Following and Analysed.

[1] “Deep Learning In Non Coding Variant ”

[2] “NCNet:Deep Learning Network Models For Predicting Function Of Non-Coding DNA”

[3] “Deep Learning Approaches for lncRNA-Mediated Mechanisms: A Comprehensive Review of Recent Developments”

[4] “Genomic Benchmarks: A Collection Of Datasets For Genomic Sequence Classifcation”

[5] “Computational Approaches for Identifying Non-Coding RNAs”

[6] “Classification of DNA Sequences Using Machine Learning”

[7] “De Novo Approach to Classify Protein-Coding and Non-Coding Transcripts”

[8] “Chromatin signature reveals over a thousand highly conserved large non-coding RNAs in mammals”

IV. METHODOLOGY

The "RandomForestClassifier" is a widely-used ensemble learning method designed for classification tasks. This classifier enhances model performance and robustness by combining the outputs of multiple decision trees. The core idea behind this method is to train several decision trees independently and aggregate their predictions to produce a more accurate and stable final output. This ensemble approach leverages the strengths of individual trees while mitigating their weaknesses, particularly overfitting.

During the training phase, the RandomForestClassifier employs the bagging (Bootstrap Aggregating) technique. This involves generating multiple bootstrap samples from the original training dataset. Each bootstrap sample is created by randomly selecting samples from the training data with replacement, meaning that some samples may appear multiple times in a sample while others may not appear at all. This step ensures that each decision tree in the forest is trained on a unique subset of the data, promoting diversity among the trees.

As each decision tree is constructed, the classifier introduces additional randomness through feature selection. At each node of a tree, a random subset of the available features is selected. The best feature and threshold for splitting the node are then determined from this subset, using criteria such as Gini impurity or entropy. This random feature selection reduces the correlation between trees and enhances the overall diversity of the forest, which is crucial for the ensemble's effectiveness.

The process of creating bootstrap samples, constructing decision trees, and selecting random features is repeated for a predefined number of trees, known as n_estimators. Each tree is trained independently on its respective bootstrap sample, ensuring that the forest is composed of a diverse set of models. This diversity is a key factor in the RandomForestClassifier's ability to generalize well to new, unseen data.

In the prediction phase, the trained RandomForestClassifier uses the ensemble of decision trees to make predictions on new input data. Each tree independently predicts the class label for the input features. The final prediction is determined by aggregating these individual predictions through a process known as majority voting. In this method, the class label that receives the most votes from the trees is selected as the final output. This aggregation method helps to smooth out the predictions, leading to improved accuracy and robustness compared to any single decision tree.

The RandomForestClassifier's combination of bootstrap sampling, random feature selection, and majority voting makes it a powerful tool for classification tasks. It achieves high accuracy by leveraging the collective wisdom of multiple decision trees, robustness by reducing overfitting, and versatility by being applicable to a wide range of classification problems. This ensemble approach ensures that the final model is both reliable and capable of handling complex datasets with varying characteristics.

V. SYSTEM ARCHITECTURE

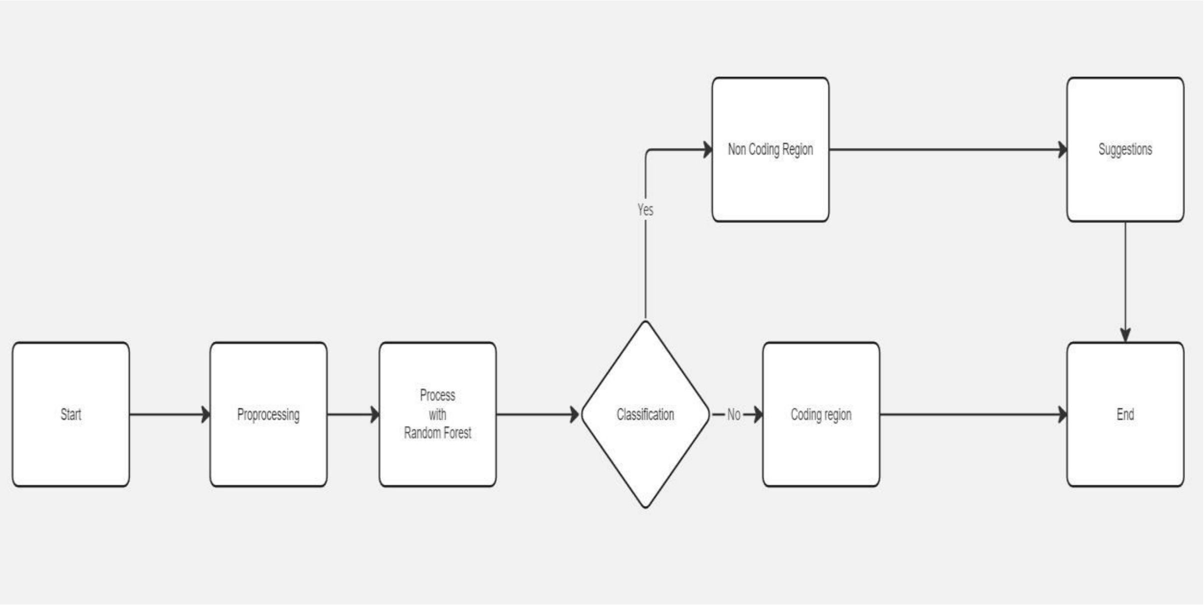

FIG. 1 Software Architecture

VI. CONCLUSION

Deploying deep learning models to analyze non-coding genomic regions has the potential to unlock significant insights into genome functionality. This project successfully implemented a Random Forest Classifier, leveraging data sourced from Kaggle and undergoing rigorous preprocessing, model development, and integration. The system accurately classifies genomic sequences into coding and non-coding regions, demonstrating the efficacy of machine learning in genomics. The developed web application ensures user-friendly interaction, while robust deployment and monitoring strategies guarantee reliable performance. This work paves the way for further research and practical applications in understanding the vast non-coding regions of the genome.

REFERENCES

1. Lee Kuan Xin, Afnizanfaizal Abdullah,"Deep Learning In NonCoding Variant",doi:dx.doi.org/10.11591/ije
2. Hanyu Zhang, Che-Lun Hung, Meiyuan Liu , Xiaoye Hu And Yi-Yang Lin,"Deep Learning Network Models For Predicting Function Of Non-Coding DNA",doi.org/10.3389/fgene.2019.00432
3. Yoojoong Kim and Minhyeok Lee,"Deep Learning Approaches for lncRNA-Mediated Mechanisms: A Comprehensive Review of Recent Developments",doi.org/10.3390/ijms241210299
4. Katarína Grešová, Vlastimil Martine k, David Čechák, Petr Šimeček And Panagiotis Alexiou,"Genomic Benchmarks: A Collection Of Datasets For Genomic Sequence Classifcation ",doi.org/10.1186/s12863-023-01123-8
5. Alipanahi, B., Delong, A., Weirauch, M. T., and Frey, B. J.. Predicting the sequence specificities of DNA- and RNA-binding proteins by deep learning. Nat. Biotechnol. 33,831– 838. doi: 10.1038/nbt.3300
6. Kelley, D. R., Snoek, J., and Rinn, J. L. (2016). Basset: learning the regulatory code of the accessible genome with deep convolutional neural networks. Genome Res.26, 990–999. doi: 10.1101/gr.200535.115
7. Xuan, P.; Wang, S.; Cui, H.; Zhao, Y.; Zhang, T.; Wu, P. Learning global dependencies and multi-semantics within heterogeneous graph for predicting disease-related lncRNAs.
8. Xie, F.; Yang, Z.; Song, J.; Dai, Q.; Duan, X. DHNLDA: A Novel Deep Hierarchical Network Based Method for Predicting lncRNA-Disease Associations.
9. Li, A., Zhang, J., & Zhou, Z., "PLEK : a tool for predicting long non-coding RNAs and messenger RNAs based on an improved k- mer scheme,"
10. Zhao, J., Song, X., & Wang, K., "lncScore: alignment-free identification of long noncoding RNA from assembled novel transcripts. "

INTERNATIONAL JOURNAL
OF INNOVATIVE RESEARCH
IN COMPUTER & COMMUNICATION ENGINEERING