[BIOINFO 2023]

DNA Sequence-Based Virus Classification Using Deep Learning and Machine Learning

Kavya Dasaramoole Prakash[1], and Kyungsup Kim[1,*]

[1] *Department of Computer Engineering, Chungnam National University*
**Corresponding author:sclkim@cnu.ac.kr*

Deoxyribonucleic acid (DNA) is a complex biomolecule, often extending to thousands of nucleotides, that contains the fundamental genetic information as a foundation for all living organisms. DNA sequences play a pivotal role in genomics, offering insights into genetic information and molecular processes. Deciphering and categorizing these sequences are essential for understanding the genetic basis of life and for applications in fields such as disease diagnosis, evolutionary studies, and drug development. However, it can indeed be quite challenging to classify DNA sequences. This study employs two distinct methodologies, Tokenization, and One-Hot Encoding, to reformat the DNA sequences. For classification, we utilize a Convolutional Neural Network (CNN) that integrates both tokenized and one-hot encoded data. Additionally, traditional Machine Learning models, including Support Vector Machine (SVM), XGBoost, and Random Forest, are integrated forcomparative analysis. The core objective is to evaluate and contrast these methodologies in DNA sequence classification, illuminating their respective strengths and characteristics. Based on the experimental findings, Random Forest, XGBoost, and CNN with One-Hot Encoding achieve higher accuracy rates of 98.5%, 98.1%, and 96.16%, respectively.