Vol.64: e21210075, 2021
https://doi.org/10.1590/1678-4324-2021210075
ISSN 1678-4324 Online Edition

Article – Biological and Applied Sciences

A Novel Frequency Based Feature Extraction Technique for Classification of Corona Virus Genome and Discovery of COVID-19 Repeat Pattern

Muthulakshmi Murugaiah[1*]
https://orcid.org/0000-0003-1833-5311P

Murugeswari Ganesan[2]
https:/orcid.org/0000-0002-7356-518X

[1,2]ManonmaniamSundaranar University, Department of Computer Science and Engineering, Tirunelveli, Tamil Nadu, India.

Editor-in-Chief: Alexandre Rasi Aoki
Associate Editor: Raja Soosaimarian Peter Raj

Received: 2021.02.12; Accepted: 2021.07.21.

*Correspondence: muthulakshmi311292@gmail.com;Tel.: +91-9952627205 (M.M.).

HIGHLIGHTS

- A Feature extraction technique is proposed which extracts 120 features from Corona virus genome sequences.
- A classification accuracy of 97.96% is achieved for a dataset with 1000 samples of seven strains of Corona virus genome sequences.
- The most repeat pattern “TTGTT” is identified in COVID-19 genome sequence.

Abstract: Genome sequence regulates the life of all living organisms on earth. Genetic diseases cause genomic disorders and therefore early prediction of severe genetic diseases is quite possible by Genome sequence analysis. Genomic disorders refer to the mutation that is rearrangement of bases in the Genome of an organism. Genome sequence analysis and mutation identification can help to classify the diseased genome which can be accomplished using Machine Learning techniques. Feature Extraction plays a crucial role in classification as it is used to convert the Genome sequences into a set of quantitative values. In this article, we propose a novel feature extraction technique called Frequency based Feature Extraction Technique which extracts 120 features from genome sequences for classification. In the current scenario, COVID-19 is the pandemic disease and Corona virus is the source of this disease. So, in this research work, we tested the proposed feature extraction technique with 1000 samples of Genome sequences of Corona virus affected patients across the world. The extracted features were classified using both Machine Learning and Deep Learning techniques. From the results, it is evident that the proposed feature extraction technique performs well with Convolutional Neural Network classifier giving an accuracy of 97.96%. The proposed technique also helps to find the most repeat patterns in the genome sequences. It is discovered that the pattern “TTGTT” is the most repeat pattern in COVID-19 genome.

Keywords: Genome Sequences; Feature Extraction; Classification; Corona virus; COVID-19; Machine Learning.

INTRODUCTION

Bioinformatics is an interdisciplinary field which applies computer algorithms for biological data analysis. Biological data refers to the package of instructions that makes up and regulates the organism. The instructions are expressed using the genetic materials such as Deoxyribo Nucleic Acid (DNA) or Ribo Nucleic Acid (RNA). DNA is made up of four alphabets A,C,G and T which are called nucleotide bases where A refers to Adenine, C refers to Cytosine, G refers to Guanine and T refers to Thymine. In parallel with DNA, RNA also consists of four nucleotide bases but instead of Thymine it contains Uracil. So, the bases in RNA are A,C,G and U where U refers to Uracil. The complete set of DNA or RNA is called Genome. Genome is made up of DNA in almost all organisms except some viruses. This Genome needs to be sequenced as it provides the most significant information related to the survival of all living organisms. The orderly arrangement of the four nucleotide bases is called Genome Sequences. Since the Genome is made up of DNA bases, it may also be called as DNA sequence. The sample Genome sequence is AGCGTTGATCGTTGACGAGA. Genome Sequence analysis is the most significant area of Bioinformatics as it deals with the life of an organism. Genome sequence analysis not only used to predict genetic diseases but almost all severe diseases such as cancer, cardio vascular disease etc. This is possible because these diseases lead to the rearrangement of nucleotide bases (or) Mutation in Genome sequences. Mutation refers to the changes such as addition, deletion or substitution of nucleotide bases in the Genome sequences. Feature extraction from Genome sequences is the essential step as it is used for finding mutations and also for classification to predict diseases.

Motivation and Justification of Research work

In the state of play, COVID-19 is the pandemic disease which is spreading rapidly, and becoming a threat to humans' lives worldwide. It is found that there are 7 human infecting Corona virus strains are available. They are HCoV-OC43, HCoV-HKU1, HCoV-229E, HCoV-NL63, MERS-CoV, SARS-CoV and SARS2-CoV2 [1]. Among the 7 strains of Corona virus, SARS2-CoV2 is the novel human infecting Corona virus which was first identified in Wuhan, China in December 2019 and it is named as COVID-19.The number of affected patients and the mortality rate is inordinate and it is still increasing gradually. A confirmed number of 167,958,998cases including 3,492,673 deaths, as of 27th May2021, have been reported in World Health Organization (WHO) [2]. By analyzing the daily record of Corona affected cases are reported in WHO, it is obvious that the number of COVID-19 cases is steadily increasing and there is a need to analyze and predict the disease at an early stage. But the fact is that, identification of SARS2-CoV2 can give wrong results because of their genetic similarity with other strains of Corona virus. The effective computer aided diagnosis and classification system for diseases is of great significant in Medical Informatics[3].Motivated by all these facts, it is required to analyze and classify the Genome of Corona virus despite the laboratory investigations. Therefore, in this research, a novel feature extraction techniqueis proposed to classify the Genome sequences of Corona virus for accurate prediction of COVID-19.

Contribution to Research

The main contributions in the proposed Feature extraction technique are as follows:

1. We extracted a total of 120 features from the genome sequencesbased on storage, Base(s) frequency, arrangement of patterns and composition of amino acids.
2. We evaluated the proposed feature extraction technique using both Machine Learning and Deep Learning classifiers.
3. We compared the proposed feature extraction technique with the existing techniques.
4. Using the proposed feature extraction technique, we identified the most repeat pattern of COVID-19 genome.

Related Works

A detailed study of feature extraction methods for Genome sequences was done by Robson ParmezanBonidiaR and coauthors [4]. The features from DNA sequenceswere also extracted by using multifractal analysis [5] to identify the biological functionality using the information present in the DNA sequences. PyFeatis a python-based feature extraction tool [6] which can effectively be applied to the biological sequences such as DNA, RNA and protein sequences.A tool named "FastFeatGen" for parallel feature extraction from Genome sequences was developed by Rahman M. [7]. A comprehensive web-based tool named “Seq2Feature” was developed by NikamRand coauthors [8] which computes features including physicochemical, energetic and conformational properties of DNA.

Liu Z and coauthors [9] developed a feature extraction algorithm for Genome sequences. The author presented a novel feature called Base-Base Correlation using joint probabilities of nucleotide bases in the sequence. The author also extracted features such as Word Frequency and Dinucleotide Relative Abundance in addition to the proposed feature and used them to classify the functional regions of the human chromosome 22. You W and coauthors [10] developed a feature extraction method called dinucleotide composition to classify the DNA sequences. This method transforms DNA sequences into 16-dimension vector which are the features. Then the features are fed into the Artificial Neural Network to classify the DNA sequences of several bacteria. Zhou Qand coauthors [11] developed a feature extraction methodology for the classification of DNA sequences. To extract features, the authors introduced the windows method whichcalculates the frequency distribution of DNA double nucleic acids. The extracted features are given as input to Support Vector Machine (SVM) classification algorithm and Artificial Neural Network (ANN) algorithm. The proposed method was implemented on 2 bacterial dataset and it is able to classify bacteria. Mathew A and coauthors [12] used aminoacid index features (AAIndex features) for the classification of mutated driver gene. For obtaining AAIndex feature, the author assigned a complex number to each amino acid. Then, the author used Daubechies wavelet function to extract features and wavelet transforms to obtain coefficient matrix. Back propagation Neural Network is used to classify the normal and mutated genome. Wavelet packet analysis is also used to extract features from Genome sequences for classification purpose [13]. Vijayanandcoauthors [14] extracted features using Chaos Game Representation (CGR) of DNA sequences. The author plotted CGR using C-GRex tool. The obtained CGR is divided into 64 cells where black dots in CGR represent the frequency of trinucleotides. The author divided the trinucleotide frequency by the total length of the DNA sequence and framed a 64-element vector which is further used as features to classify the eight categories of Eukaryotic organisms using ANN.Sathish Kumar S and Duraipandian N [15] introduced a classification technique for DNA sequences by integrating Data mining and Artificial Neural Networks which is based on Pattern mining. The authors generated dinucleotide and trinucleotide as patterns and calculated the frequency for both set of patterns. The frequency values are used as features for the classification using ANN. Kristensenand coauthors [16] classified the DNA sequences using Multi Layer Perceptron (MLP) and SVM network. The authors used sliding window technique to extract the subsequence frequencies and then normalized them using a normalization condition. The authors set the window size as 2,3 and 4 and calculated the frequencies of dinucleotide, trinucleotide and tetranucleotide patterns. The extracted frequency values are used as features and given into MLP and SVM to classify the Prokaryotic and Eukaryotic DNA. Researchers also used the advanced Machine Learning concept called Deep Learning for classification and prediction of Genome sequences for various applications. KassimN and coauthors [17] implemented Convolutional Neural Network to identify and classify the genetic marker for liver cancer caused by Hepatitis B virus DNA sequences. TampuuA and coauthors [18] developed ViraMiner in which both used Convolutional Neural Network with One-hot encoding for classification and identification of SARS-2COV2 genome. Bhonde S and coauthors [19] developed a predictive modelling for COVID-19 genome sequences using Deep Learning. The authors had done feature extraction and feature selection using principal component analysis and Random forest. Subsequently, the genome sequences are classified using Convolutional Neural Network (CNN) and Bidirectional-Long Short-Term Memory (BLSTM) to predict the risk of COVID-19. Habib P and coauthors [20] developed a tool named “COVIDier” for the classification of various virus strains of Corona viruses’ genome and virulence proteins classification which works on the basis of Deep Learning approach.For feature extraction, the keyword extraction methods can also be considered which was used for text classification [21]. Probabilistic (or) statistical classifiers such as Naive Bayes (NB) classifier, Decision Tree (DT) classifier, Support Vector Machines (SVM) and instance-based classifiers such as the k Nearest Neighbour(kNN) algorithm, deep learning models such as Neural Networks (NN) are widely used for text classification [22]. In case classifiers, NB and kNN are usedfor text sentiment classification [23].Deep learning algorithms such as Recurrent Neural Network (RNN) are also used for text mining and text classification [24]. Bidirectional Long Short Term Memory (LSTM) is a type of RNN whichfinds its application in sarcasm classification in text documents [25].

MATERIALS AND METHODS

The Genome sequences of Corona virus are stored and managed in databases such as National Centre for Biotechnology and Information (NCBI), Virus Pathogen Database and Analysis Resource (ViPR) and hCoV-19 etc. More than hundreds of Genome of COVID-19 patients are sequenced and stored in the databases every day. In this research work, Corona virus genome sequences are taken from the NCBI database and passed as an input to the proposed feature extraction technique to extract features. Thereupon, the extracted features are fed in to Machine Learning and Deep Learning classifiers to classify the 7 strains of Corona virus genome sequences precisely, which in turn help to spot the COVID-19 genome accurately. The outline of the work is shown in Figure.1.

Proposed feature extraction technique

Features describe the complete set of data (Two-dimensional data, sequential data and text data)by means of set of quantitative values efficiently. The set of values is collectively called as feature vectorand individual quantitative value is called as a feature. In case of Genome sequence analysis andclassification, it finds too difficult to analyze the raw genome sequence data as it is whopping in size.Thecomplete human genome consists of around 6 billion bases which are strenuous to read and analyze.Besides, it takes more computation time.Feature extraction addresses this issue by extracting the significant features from the Genome sequences which makes the analysis much easier in a short period of computation. Most of the existing feature extraction techniques extract very few features from genome.In this research work, we consider a maximum number of possible features from Genome sequences which help to predict and classify the diseased genome.

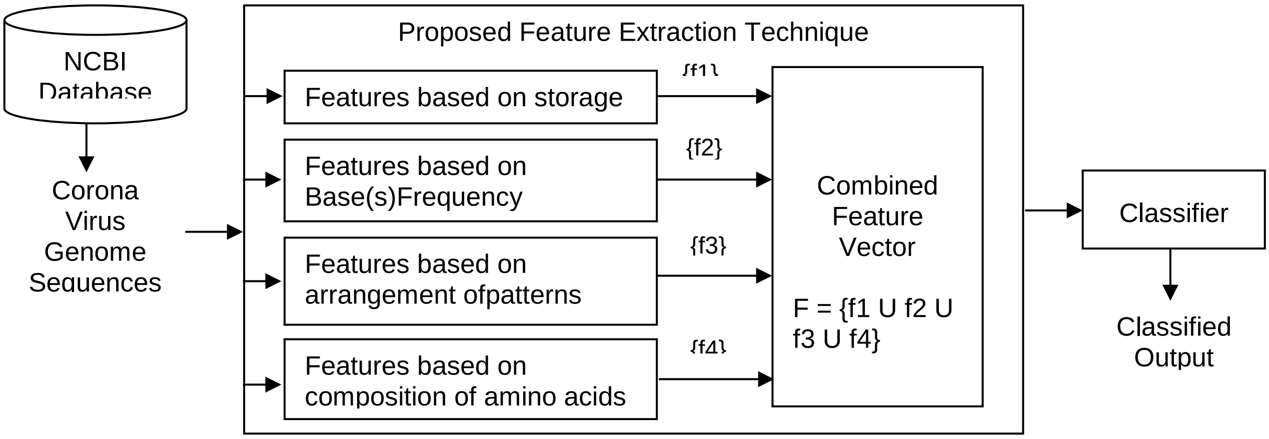
NCBI Database
Corona Virus Genome Sequences
Proposed Feature Extraction Technique
Features based on storage
Features based on Base(s)Frequency
Features based on arrangement ofpatterns
Features based on composition of amino acids
{f1}
{f2}
{f3}
{f4}
Combined Feature Vector
F = {f1 U f2 U f3 U f4}
Classifier
Classified Output

Figure 1. Outline of the work

In the proposed Frequency based Feature Extraction Technique (FFET), based on the characteristics of the extraction methods, the features are categorized as follows:Features based on storage, Features based on Base(s) Frequency, Features based on arrangement of patterns and Features based on composition of Amino acids.

Features based on storage

The length and size of the Genome varies from one organism to another which will be helpful for the classification process. This category includes two features namely length and size of the Genome. Both the features are interrelated with one another.

Length

Let S be the Genome sequence and B be the Base in the Genome sequence.Length refers to the length of the input Genome sequence. Length is computed by counting the total number of bases in the sequence as given in Equation (1). For instance, the Length of the Genome sequence ACGTGATCAGTACG is 15.

$$Length(S) = Total\ Number\ of\ bases(B) \tag{1}$$

Size

Genome sequence consists of sequence of bases. Each base takes 1 byte for storage thereby the length of the sequence is equal to its size when expressed in terms of bytes. In the proposed feature extraction technique, Size of the sequence is expressed in terms of KB (Kilo Bytes) using Equation (2).

$$Size(S) = \frac{Length(S)}{1024} \tag{2}$$

Such that 1 Kilo Byte =1024 Bytes.The features based on storage are given as f1 = {Length(S), Size(S)}.

Features based on Base(s) Frequency

This category includes features such as N Count, Base Count, Dimer Count and Codon Count. A total of 85 features are extracted under this category. Identification of mutation in the genome is the preliminary step in the prediction of disease. A mutation refers to the changes in the bases of the genome sequences. The feature (N Count) does not contribute to the mutation prediction and classification of diseased genome. But it plays a role in the classification of organisms which is called as Taxonomic classification. The other base(s) frequency features such as Base Count, Dimer Count and Codon Count help to find the mutation inthe genome sequences. They also help in the classification and comparison of genome. The descriptions of the features are as follows:

N Count

Genome sequences consist of a Non template base N in addition to the four nucleotide bases (A,C,G and T). The feature "N Count" (NC) refers to the number of occurrences of Non-template base (N) in the Genome sequence as given in Equation (3).

$$NC(S) = Number\ of occurences\ of\ Base\ N \tag{3}$$

Base Count

Base Count refers to the number of occurrences of individual nucleotide bases (A,C,G and T) in the Genome sequence. Then the Base Count BC(S) is as follows:
Bases = {A,C,G,T}
BC(S) = {Number of occurrences of individual Bases}

Dimer Count

Two individual bases join together to form a Dimer. The feature "Dimer Count (DC)" refers to the number of occurrences of all possible combinations of Dinucleotides in the Genome sequences. There are total of 16 Dimers in a Genome sequence.
Dinucleotides = {AA, AC, AG, AT, CA, CC, CG, CT, GA, GC, GG, GT, TA, TC, TG, TT}
DC(S) = {Number of occurrences of individual Dinucleotides}

Codon Count

Three individual bases group together to form a Codon. The feature "Codon Count (CC)" refers to the number of occurrences of all possible combinations of trinucleotides in the Genome sequence. There are total of 64 codons in a Genome sequence.
Codons = {Number of all possible combinations of trinucleotides}
Codons = {AAA, AAC, AAG, AAT, CCA, CCC, CCG, CCT...TTG, TTT}
CC(S) = {Number of occurrences of individual Codons}

The features based on Base(s) frequency are given as f2 = {NC(S), BC(S), DC(S), CC(S)}

Features based on arrangement of patterns

The 6 features in this category are Most Repeat Pattern (MRP) Count, Palindrome Count, Palindrome Threshold, Entropy, Open Reading Frame (ORF) Count and ORF Threshold. All these features are mainly based on the arrangement of patterns in the genome which contributes much in the classification and comparison of genome. ORF Count helps in the process of Gene prediction and comparison.

Most Repeat Pattern Count (MRP)

Most Repeat Pattern Count refers to the number of occurrences of MRP in the Genome Sequence. In this work, the Pattern length is set to 4 and so the MRP count gives the number of occurrences of most repeat tetranucleotide pattern in the sequence.
Tetranucleotides = {AAAA, AAAC, AAAG, AAAT, CCCC, CCCA, CCCG, ..., TTTT}
MRP(S) = Number of occurrences of most repeat tetranucleotide pattern

Palindrome Count (PC)

There are a number of Palindrome patterns in the Genome sequences. Hence, the Palindrome Count refers to the total number of palindrome patterns in the sequence. An example palindrome pattern is ACTGCCGTCA.

PC(S) = Total number of Palindrome patterns

Palindrome Threshold (PT)

Palindrome Threshold returns the length of the palindrome which is maximum.

PT(S) = max {length (palindrome)}

Entropy

Entropy refers to the calculation of randomness in the sequence. Let a Genome sequence be S such that S=Y1,Y2,...Yn, which occurs with the probability P(Y1), P(Y2),...P(Yn), then the Entropy of S that is E(S) is calculated using Equation (4).

$$E(S) = -\sum_{i=1}^{n}(\mathrm{P}(\mathrm{Y_i}) \log_2 \mathrm{P}(\mathrm{Y_i})) \tag{4}$$

Open Reading Frame Count (ORF)

Open Reading Frame is a part of the reading frame in the genome which has the ability to be translated into Protein sequence. In general, an ORF begins with the start codon (ATG) and ends with any one of the stop codons (TAA, TAG, and TGA) [26]. ORF count refers to the total number of ORFs present in the sequence.

ORF Count(S) = Total number of ORFs

Open Reading Frame Threshold

ORF Threshold represents the length of the ORF which contains maximum number of bases.

ORF Threshold (S) = max {length (ORF)}

The features based on arrangement of patterns are given as f3 = {MRP(S), PC(S), PT(S), E(S), ORF Count(S), ORF Threshold(S)}

Features based on Composition of Amino acids

There are total of 27 features identified under this category. The features such as Amino acid count, Stop Codon (SC) count, Atomic Composition and Molecular Weight are calculated by considering the composition of amino acids in the genome sequences. The features such as Atomic Composition and Molecular Weight vary from genome to genome, hence it plays a major role in the classification process.

Amino acid Count

Amino acids are encoded by group of codons (three bases). Some Amino acids are encoded by one codon whereas; some are encoded by more than one codon. There are total of 20 different amino acids in human body that functions as building blocks of proteins which are shown in Table 1. For calculating Amino acid count, the Genome sequence is converted into Amino acid sequence by replacing the codons with the respective amino acid codes as given in Table 1. Amino acid count refers to the number of occurrences of each amino acid in the Genome sequence.

AA(S) = {Number of occurrences of individual amino acids}

Stop Codon Count

The codon ATG which codes for Methionine is called as Start codon. Similarly, the three codons namely TAA, TAG, TGA are called as termination codons for which the amino acid codes do not exist. Stop Codon (SC) count refers to the sum of the number of occurrences of the three termination codons in a Genome sequence.

SC(S) = {Number of occurrences of Stop Codons}

Atomic Composition

Atomic composition is calculated for the amino acid sequence. From the Molecular formula of amino acids in Table 2, it is clear that amino acids are made up of 5 different atoms namely Carbon (C), Hydrogen (H), Oxygen (O), Nitrogen (N) and Sulphur (S). The feature "Atomic composition" of an amino acid sequence refers to the number of occurrences of the 5 different atoms (C,H,O,N and S). As the amino acid sequence produces protein, there is a terminal H and OH bond in all atoms of amino acids. Hydrogen and Oxygen are the internal elements and so we have to subtract the H and OH bond from Hydrogen and Oxygen atoms while calculating the Atomic composition. The number of H and O in terminal bond is 2 and 1respectively.Let i be the amino acid, then the Atomic composition is calculated as shown in Table 2.

Table 1. Amino acid Description

Sl. No.	Codons	Amino acid	Amino acid Code	Molecular Formula	Molar Mass
1	GCT, GCC, GCA, GCG	Alanine	A	$C_3H_7NO_2$	89.09
2	CGT, CGC, CGA, CGG, AGA, AGG	Arginine	R	$C_6H_{14}N_4O_2$	174.20
3	AAT, AAC	Asparagine	N	$C_4H_8N_2O_3$	132.12
4	GAT, GAC	Aspartic acid	D	$C_4H_7NO_4$	133.10
5	TGT, TGC	Cysteine	C	$C_3H_7NO_2S$	121.15
6	CAA, CAG	Glutamine	Q	$C_5H_{10}N_2O_3$	146.15
7	GAA, GAG	Glutamic acid	E	$C_5H_9NO_4$	147.13
8	GGT, GGC, GGA, GGG	Glycine	G	$C_2H_5NO_2$	75.07
9	CAT, CAC	Histidine	H	$C_6H_9N_3O_2$	155.16
10	ATT, ATC, ATA	Isoleucine	I	$C_6H_{13}NO_2$	131.17
11	TTA, TTG, CTT, CTC, CTA, CTG	Leucine	L	$C_6H_{13}NO_2$	131.17
12	AAA, AAG	Lysine	K	$C_6H_{14}N_2O_2$	146.19
13	ATG	Methionine	M	$C_5H_{11}NO_2S$	149.21
14	TTT, TTC	Phenylalanine	F	$C_9H_{11}NO_2$	165.19
15	CCT, CCC, CCA, CCG	Proline	P	$C_5H_9NO_2$	115.13
16	TCT, TCC, TCA, TCG, AGT, AGC	Serine	S	$C_3H_7NO_3$	105.09
17	ACT, ACC, ACA, ACG	Threonine	T	$C_4H_9NO_3$	119.12
18	TGG	Tryptophan	W	$C_{11}H_{12}N_2O_2$	204.23
19	TAT, TAC	Tyrosine	Y	$C_9H_{11}NO_3$	181.19
20	GTT, GTC, GTA, GTG	Valine	V	$C_5H_{11}NO_2$	117.15

Table 2. Atomic Composition features

Sl.No.	Features in Atomic Composition	Feature Description	Formula
1	C	Total Number of Carbon atoms in the Amino acid sequence	$C = \sum_{i=1}^{n} \left(\mathrm{AC(i)} * \mathrm{NC(i)}\right)$
2	H	Total Number of Hydrogen atoms in the Amino acid sequence	$H = \sum_{i=1}^{n} (\mathrm{AC(i)} * \mathrm{NH(i)}) - 2$
3	N	Total Number of Nitrogen atoms in the Amino acid sequence	$N = \sum_{i=1}^{n} (\mathrm{AC(i)} * \mathrm{NN(i)})$
4	O	Total Number of Oxygen atoms in the Amino acid sequence	$O = \sum_{i=1}^{n} (\mathrm{AC(i)} * \mathrm{NO(i)}) - 1$
5	S	Total Number of Sulphur atoms in the Amino acid sequence	$S = \sum_{i=1}^{n} (\mathrm{AC(i)} * \mathrm{NS(i)})$

Where n refers to number of amino acids in the sequence; in general there are 20 amino acids.AC(i) refers to the Amino acid Count (number of occurrences of particular amino acid i in the sequence); NC(i) is the Number of Carbon atoms in the respective amino acid i; NH(i) is the Number of Hydrogen atoms in the respective amino acid i; NN(i) is the Number of Nitrogen atoms in the respective amino acid i; NO(i) is the Number of Oxygen atoms in the respective amino acid i; NS(i) is the Number of Sulphur atoms in the respective amino acid i.

Molecular Weight

Molecular Weight refers to the Weight of the molecules in the amino acid sequence which is calculated using Molar mass which was shown in Table 1. The feature Molecular Weight is labelled as f44 which is calculated using Equation (5).

$$\mathrm{MW} = \sum_{i=1}^{n}(\mathrm{AC(i)} * \mathrm{MM(i)}) - \mathrm{WW} \tag{5}$$

Where MW is the Molecular Weight of the amino acid sequence; n refers to number of amino acids in the sequence;in general there are 20 amino acids.AC(i) is the number of occurrences of respective amino acid i in the amino acid sequence; MM(i) is the Molar Mass of respective amino acid i; WW is the Weight of Water Molecule. The Weight of Water molecule is 18.015.

All the features that are extracted under the four category are combined together to form a feature vector which consists of 120 features totally. The combined feature vector (F) is given as input to the classification process.

Feature Vector = F = {f1 U f2 U f3 U f4}

Classification

Classification is a kind of supervised Machine Learning concept which predicts the class to which the data belongs and this can be accomplished by extracting features from the input data. The process of classification in this research work is shown in Figure 2.After extracting features from the Genome sequences, Classification algorithm is applied to the dataset which contains the extracted features for classification of seven strains of Corona virus. Machine learning algorithms are extensively used for automatic detection of illnesses [27].Deep learning algorithms are widely used in text mining and text classification[28].Deep Learning models are generally used for raw input data. But we can also use Deep Learning models for manually extracted feature data set. In most cases, both feature based and raw data based Deep Learning models perform correspondingly in terms of accuracy [29].For this research work, we employed both Machine Learning and Deep Learning classification algorithms. In this research, we used the Machine Learning algorithms such as k Nearest Neighbour (kNN), Support Vector Machine (SVM) and Naive Bayes (NB).Deep learning algorithms such as Artificial Neural Network(ANN), Convolutional Neural Network (CNN)and Recurrent Neural Network (RNN) are employed. In case of RNN, Long Short-Term Memory (LSTM)is used for Classification. Before applying the classifier, the dataset is split into training and testing dataset using Cross Validation. The classifier is trained with the training dataset and then the trained classifier is applied to the testing dataset in order to classify the strains of Corona virus.

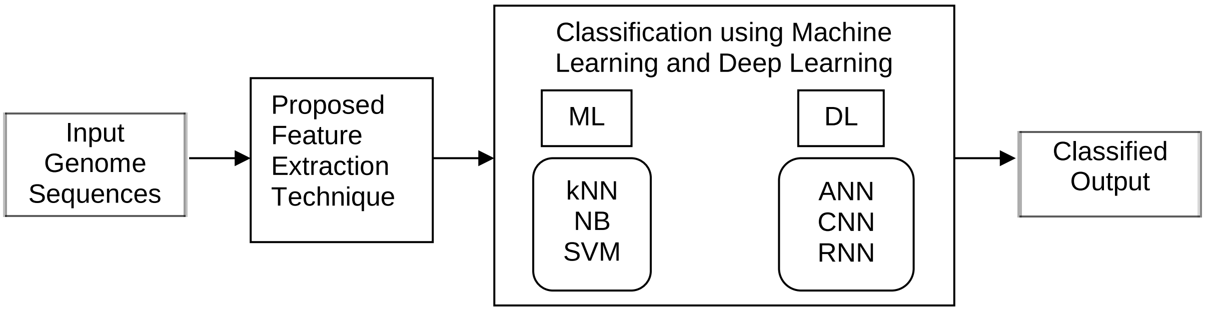
Input Genome Sequences
Proposed Feature Extraction Technique
Classification using Machine Learning and Deep Learning
ML
DL
kNN
NB
SVM
ANN
CNN
RNN
Classified Output

Figure 2. Flow of Classification process

RESULTS AND DISCUSSIONS

The dataset and its description, experimental results and discussions are as follows

Dataset

The Genome sequences of Corona virus is downloaded in FASTA format from NCBI [30] which is a freely available database and it contains the Genome sequences of COVID-19 patients across the world.

Description of Corona virus Genome Dataset

A virus has DNA (or) RNA as its genetic material and is called as DNA virus (or) RNA virus. Corona virus is a RNA virus which causes respiratory infections that ranges from mild to severe. RNA of Corona virus is sequenced and stored in the database in the form of complementary DNA (cDNA) sequences. For the purpose of this research, we downloaded 1000 samples of RNA sequences of 7 human strains of Corona virus from the NCBI database. These collected 1000 samples include the RNA sequences of patients from various countries such as Belgium, China, Denmark, France, India, Peru, Taiwan, United States of America (USA), Vietnam etc. We grouped the seven strains of virus into 5 classes based on the severity of diseases which is shown in Table 3.

Table 3. Dataset Description

Sl.No.	Strains	No. of Samples	Symptoms /Disease	Class Name
1	HCoV-OC43 and HCoV-229E	156	Common Cold	CC
2	HCoV-HKU1 and HCoV-NL63	114	Mild Lower Respiratory Infection	MLRI
3	MERS-CoV	160	Middle East Respiratory Syndrome	MERS
4	SARS-CoV	220	Severe Acute Respiratory Syndrome	SARS
5	SARS2-CoV2	350	COVID-19	COVID-19

Experimental Result Analysis of the Proposed Feature Extraction technique

The proposed feature extraction technique is applied to the Corona virus Genome sequences and we generated a dataset containing a feature vector which consists of 120 features. We extracted the features from one sample of all seven strains of Corona virus and they are listed in Table (4-8).

Table 4. Features based on storage

Sl.No.	Features	HCoV-OC43	HCoV-229E	HCoV-HKU1	HCoV-NL63	MERS-CoV	SARS-CoV	SARS2-CoV2
1	Length	30737	27055	29604	27553	30119	29711	29797
2	Size (in KB)	30	26	29	27	29	29	29

Table 4 shows the features based on storage such as Length and Size of the genome. Table 5 shows the Base(s) frequency features such as Base and Dimer count.

Table 5. Base and Dimer Count features

Sl.No.	Features	HCoV-OC43	HCoV-229E	HCoV-HKU1	HCoV-NL63	MERS-CoV	SARS-CoV	SARS2-CoV2
				Base Count				
1	N	0	31	2727	0	0	0	636
2	A	8481	7342	7475	7259	7900	8452	8720
3	C	4654	4488	3526	3985	6116	5938	5355
4	G	6624	5801	5096	5509	6304	6185	5743
5	T	10978	9393	10780	10800	9799	9136	9343
				Dimer Count				
6	AA	2453	2304	2264	2223	2195	2490	2786
7	AC	1329	1540	1063	1372	1766	1978	1973
8	AG	1827	1390	1381	1281	1654	1750	1706
9	AT	2872	2108	2766	2383	2284	2234	2254
10	CA	1607	1629	1057	1358	1931	2205	2030
11	CC	823	682	544	593	1165	974	868
12	CG	476	475	302	325	711	567	429
13	CT	1747	1702	1623	1708	2309	2191	2028
14	GA	1620	1372	1271	1193	1472	1680	1583
15	GC	1309	1135	771	876	1490	1432	1142
16	GG	1279	1063	862	1010	1234	1203	1076
17	GT	2416	2230	2192	2430	2108	1870	1942
18	TA	2801	2037	2883	2485	2302	2077	2321
19	TC	1193	1131	1147	1143	1695	1554	1372
20	TG	3041	2872	2551	2893	2704	2665	2532
21	TT	3943	3353	4199	4279	3098	2840	3118

The set of features which are based on the arrangement of patterns in the Genome sequences are shown in Table 6.

Table 6. Features based on arrangement of patterns in the genome

Sl. No.	Features	HCoV-OC43	HCoV-229E	HCoV-HKU1	HCoV-NL63	MERS-CoV	SARS-CoV	SARS2-CoV2
1	MRP Count	505	504	628	654	333	317	324
2	Palindrome Count	1222	1000	1317	1277	1032	1036	1089
3	Palindrome Threshold	17	16	21	20	17	17	20
4	Entropy	1.9326	1.9473	1.8814	1.9044	1.9728	1.9747	1.9576
5	ORF Count	709	623	610	621	667	724	699
6	ORF Threshold	13149	10833	7260	12180	13173	13146	6672

Table 7 shows the set of features which are based on the composition of amino acids in the Genome sequence. Table 8 shows the frequency of codons in the genome sequences. Codon count is the feature which comes under the category of Base(s) Frequency.

Table 7. Features based on composition of amino acids

Sl.No	Features	HCoV-OC43	HCoV-229E	HCoV-HKU1	HCoV-NL63	MERS-CoV	SARS-CoV	SARS2-CoV2
				Amino acid Count				
1	A	537	334	354	293	418	343	274
2	R	368	355	300	318	520	551	525
3	N	438	335	484	347	285	334	436
4	D	406	147	440	238	121	158	165
5	C	387	455	511	696	506	565	560
6	Q	355	302	193	184	337	402	383
7	E	280	142	261	174	157	198	210
8	G	456	256	410	330	293	274	328
9	H	212	290	142	221	338	379	339
10	I	711	502	526	475	565	471	491
11	L	1306	1565	908	1058	1711	1425	1213
12	K	522	402	419	339	345	410	412
13	M	330	375	163	194	299	261	198
14	F	554	491	670	749	453	430	492
15	P	323	215	225	200	384	317	265
16	S	724	553	719	639	852	835	768
17	T	506	502	427	494	564	643	656
18	W	166	182	147	279	203	221	217
19	Y	420	272	559	504	348	346	383
20	V	949	824	669	650	651	613	566
21	SC Count	295	509	432	802	689	727	839
				Atomic Composition				
22	C	51747	45317	44979	45682	49250	48103	46742
23	H	81159	71983	67606	67909	78078	75756	72700
24	N	12959	11365	10954	10927	12756	12954	12582
25	O	13766	11042	12312	11375	12293	12449	12258
26	S	717	830	674	890	805	826	758
27	Molecular Weight	1128073	979309	980397	980702	1071385	1061213	1031340

Experimental setup

In case of Machine Learning, kNN, NB and SVM are used for classification. For kNN, the value of k is set to 3 and we use Euclidean distance as a distance measure. In SVM, we use multiclass SVM as there are 5 classes for classification.In case of Deep Learning, ANN, CNN and RNN are used for classification. For ANN, the number of hidden layers is set to 100. In CNN, the Softmax function is used in the fully connected layer for the purpose of classification. The number of hidden layers in CNN and RNN is also set to 100. The number of epochs for all the three Deep Learning algorithms is set to 100.

Table 8. Frequency of Codons

Sl.No.	Features	HCoV-OC43	HCoV-229E	HCoV-HKU1	HCoV-NL63	MERS-CoV	SARS-CoV	SARS2-CoV2
1	AAA	272	198	294	221	164	231	256
2	AAC	96	175	93	135	152	175	219
3	AAG	250	204	125	118	181	179	156
4	AAT	342	160	391	212	133	159	217
5	ACA	174	206	123	167	199	276	256
6	ACC	79	94	49	94	114	140	131
7	ACG	37	71	20	46	98	61	63
8	ACT	216	131	235	187	153	166	206
9	AGA	132	130	111	121	167	200	231
10	AGC	104	58	59	83	116	150	121
11	AGG	78	115	58	89	159	187	133
12	AGT	225	160	211	172	172	155	158
13	ATA	260	181	147	134	193	147	148
14	ATC	103	94	68	72	122	105	100
15	ATG	330	375	163	194	299	261	198
16	ATT	348	227	311	269	250	219	243
17	CAA	167	183	134	121	178	238	236
18	CAC	73	128	30	82	164	190	169
19	CAG	188	119	59	63	159	164	147
20	CAT	139	162	112	139	174	189	170
21	CCA	117	109	61	85	128	120	98
22	CCC	47	23	23	20	71	54	45
23	CCG	22	30	13	17	56	40	25
24	CCT	137	53	128	78	129	103	97
25	CGA	27	18	17	16	40	48	47
26	CGC	31	24	23	19	44	36	28
27	CGG	22	17	13	12	44	30	40
28	CGT	78	51	78	61	66	50	46
29	CTA	154	239	73	112	287	222	187
30	CTC	68	84	47	60	169	127	76
31	CTG	160	254	60	109	305	234	164
32	CTT	207	183	160	193	237	245	233
33	GAA	157	83	171	117	87	98	124
34	GAC	81	73	87	66	63	85	79
35	GAG	123	59	90	57	70	100	86
36	GAT	325	74	353	172	58	73	86
37	GCA	171	124	98	98	158	114	87
38	GCC	67	46	44	38	75	70	56
39	GCG	43	34	17	29	78	54	28
40	GCT	256	130	195	128	107	105	103
41	GGA	89	58	49	58	77	88	96
42	GGC	56	55	64	54	80	65	74
43	GGG	45	41	27	36	53	54	57
44	GGT	266	102	270	182	83	67	101
45	GTA	216	180	139	128	167	152	146
46	GTC	76	92	53	77	85	84	79
47	GTG	238	294	59	100	247	235	166
48	GTT	419	258	418	345	152	142	175
49	TAA	117	184	215	347	243	266	346
50	TAC	104	140	129	171	170	195	209
51	TAG	87	129	95	148	192	158	175
52	TAT	316	132	430	333	178	151	174
53	TCA	116	129	120	134	184	228	169
54	TCC	56	39	55	60	118	79	83
55	TCG	30	49	23	35	77	63	54
56	TCT	193	118	251	155	185	160	183
57	TGA	91	196	122	307	254	303	318
58	TGC	112	157	99	207	207	263	233
59	TGG	166	182	147	279	203	221	217
60	TGT	275	298	412	489	299	302	327
61	TTA	337	342	335	277	334	258	289
62	TTC	102	143	117	188	185	196	161
63	TTG	380	463	233	307	379	339	264
64	TTT	452	348	553	561	268	234	331

Performance Evaluation of the Proposed Feature extraction technique

After applying the proposed FFET, the dataset is generated which consists of 1000 samples and 120 features and is splitted into training dataset and testing dataset using K fold Cross Validation, where the value of K is set to 10. The training dataset is fed into the classifier with the assigned labels or classes. Subsequently, the testing dataset is given into the classifier to predict the class accurately.

Performance Metrics

The results of the feature extraction together with the classifier are evaluated using some performance metrics which was shown in Table 9.

Table 9. Performance Metrics

Sl.No.	Performance Metrics	Description	Formula
1	Accuracy	Proportion of number of correct predictions from overall number of samples	$\frac{TP+TN}{TP+TN+FP+FN}$
2	Specificity	Proportion of number of negative predictions out of the total number of samples which were actually negative	$\frac{TN}{TN+FP}$
3	Precision	Proportion of number of correct positive predictions out of the total number of positive predictions	$\frac{TP}{TP+FP}$
4	Recall	Proportion of number of correct positive predictions out of the total number of samples which were actually positive	$\frac{TP}{TP+FN}$
5	F1 Score	Weighted average of precision and recall.	$2*\frac{Precision*Recall}{Precision+Recall}$
6	Error Rate	Proportion of number of incorrect predictions from overall number of samples	$\frac{FP+FN}{TP+TN+FN+FP}$

Performance Analysis of the proposed Feature Extraction technique with classifiers

The performances of the proposed feature extraction technique together with various classifiers are shown in Table 10.

Table 10. Performance of the proposed feature extraction technique with classifiers

Sl.No.	Classifier	Accuracy (%)	Specificity (%)	Precision (%)	Recall (%)	F1 Score (%)	Error Rate (%)
1	kNN	92.80	92.00	75.00	68.00	79.07	7.20
2	NB	88.00	86.05	53.85	100.00	70.00	12.00
3	SVM	91.84	92.86	66.67	85.71	75.00	8.16
4	ANN	93.60	93.00	77.42	72.00	81.82	6.40
5	**CNN**	**97.96**	**97.62**	**87.50**	**100.00**	**93.33**	**2.04**
6	RNN	89.80	92.86	62.50	71.43	66.67	10.2

From Table 10, it is evident that the proposed feature extraction technique with Convolutional Neural Network gives better accuracy than any other classifiers. Table 10 infers that CNN performs better than other classification algorithms and so statistical test was carried out to evaluate whether the observed difference in performance is statistically significant or not. For this purpose, 5×2 CV Paired t test [31] was employed using accuracy as a metric and Naive Bayes classifier as a Base Classifier because it performs worse than other classifiers. In this test, 5 repetitions of 2 fold cross validation was performed with 5 different datasets and the mean accuracy is shown in Table 11. In 5×2 CV Paired t test, the default significant p value is 0.05. In this test, each of the classifier is tested with the base classifier and the t-statistic value and p value was calculated. The performance of the algorithms are classified as Better, Same and Worse based on the p value as follows. The results of the statistical test are shown in Table 11.

(i) Better – Calculated p value is lower than the significant value with better accuracy than base classifier.
(ii) Worse - Calculated p value is lower than the significant value with least accuracy than base classifier.
(iii) Same - Calculated p value is higher than the significant value.

Table 11. Results of 5X2 CV Paired t test for classifiers

Test Metric	NB	kNN	SVM	ANN	CNN	RNN
Mean Accuracy	0.85	0.90	0.87	0.91	**0.94**	0.80
Performance (No.of times)	(Better/Same/Worse)	(2/3/0)	(1/4/0)	(2/3/0)	**(3/2/0)**	(1/2/2)

From Table 11, it is clear that the result of CNN is statistically significant and it also performs better than rest of the algorithms.

Result Analysis of existing and proposed Feature extraction techniques

The results and the performance of the existing and proposed feature extraction algorithms with CNN classifier are analyzed for Corona virus genome and are shown in Table 12.

Table 12. Result Analysis of existing feature extraction techniques and proposed FFET with CNN

Sl.No	Existing Techniques	Features	Accuracy (CNN) (%)
1	Zhou Q et al. [11]	Probability distribution of double nucleotide bases	96.40
2	You W et al. [10]	Dinucleotide composition	95.92
3	Mathew A et al.[12]	Amino acid index features	89.80
4	Sathish Kumar S&Duraipandian N [15]	Pattern occurrence of double and triple nucleotide bases	93.88
5	Kristensen Tet al.[16]	Frequency distribution of double nucleotide bases	95.92
6	VijayanKet al.[14]	Probability of occurrence of trinucleotides	91.18
7	Rasheed Z &RangwalaH.[32]	GC content and Oligonucleotide frequencies	87.20
8	Proposed FFET	Base(s) Frequency features, Storage features, Features based on arrangement of patterns, Compositional features.	**97.96**

From Table 12, it is obvious that the proposed FFET performs well for classification than other existing feature extraction techniques. It is observed from Table 12 that most of the research works extract dinucleotide base composition features. It is clear from the analysis that dinucleotide base composition is good enough to perform classification by giving an accuracy of more than 90%. For the applications like taxonomic classification and genome comparison, dinucleotide base composition alone does not provide satisfactory results. For medical applications like disease prediction, more accurate classification is required. By considering all these facts, we extract 120 features, so that an improved performance is obtained. From the analysis, it is also found that the features such as Entropy, ORF count, Most Repeat Pattern count and Palindrome count help to improve the classification accuracy whereas other existing techniques failed to consider those features. The proposed FFET provides an accuracy of 97.96% with CNN for Corona virus genome classification.

DISCUSSIONS

We demonstrated a feature extraction technique to classify the genome sequences of Corona virus. We compared the proposed feature extraction technique with the prior methods which are shown in Table 12. The table apparently shows that, our proposed technique extracts maximal number of features from the genome sequences which in turn helps to upgrade the performance of classification. By extracting features using the proposed feature extraction technique; we performed a detailed analysis of the Corona virus genome. From the extracted features, one can understand the characteristics and mutating ability of the genome. From the experimental results, we identified a tetranucleotide repeat and a pentanucleotide repeat of COVID-19 genome. We observed that “TGTT” is the most repeat pattern in the COVID -19 genome when the pattern length is set to 4. On the other hand, “TTGTT” is the most repeat pattern when the pattern length is set to 5. By performing detailed experimental analysis, we also found the most repeat patterns of all strains of Corona virus genome which is shown in Table 13.The information in the Table 13 and the proposed FFET will definitely be beneficial in the area of drug designing for COVID-19 which is necessary in the medical field nowadays. The main application of the proposed feature extraction technique is that it can be used to predict the genetic diseases and also for the prediction of other diseases such as cardiovascular disease and cancer etc., as it has the ability to classify the diseased genome whatever the disease may be. It can also be used for taxonomic classification of organisms to understand the evolutionary history of the organisms. It can be used for DNA Paternity testing, Criminal identification in Forensics, Personalized medicine and Drug Designing. The limitation of the proposed FFET is that it takes more computation time especially for the features such as Atomic Composition and Molecular Weight.

Table 13. Repeat Patterns of strains of Corona virus

Sl.No.	Strains of Corona virus Genome	Most Repeat Pattern	
		Pattern Length = 4	Pattern Length = 5
1	HCoV-OC43	TGTT	TTTTA
2	HCoV-HKU1	TTTT	TTTTA
3	HCoV-229E	TGTT	TGTTG
4	HCoV-NL63	TGTT	TTGTT
5	MERS-CoV	TGTT	TGTTG
6	SARS-CoV	TGCT	TGCTG
7	SARS2-CoV2	TGTT	TTGTT

CONCLUSION

Biological sequence analysis places a vital role in predicting severe diseases. In this work, a novel feature extraction technique namely FFET is proposed which extracts 120 features from genome sequences. These features are used to classify the Corona virus genome sequences which were taken from NCBI data repository. We tested the proposed FFET with 1000 samples using 3 Machine Learning algorithms such as kNN, NB and SVM and 3 Deep Learning algorithms such as ANN, CNN and RNN. From the experimental results, it is proved that the proposed FFET achieves better accuracy of 97.96% with CNN classifier. It is also observed that the performance of Deep Learning algorithms dominates the performance of Machine Learning algorithms. One of the features namely “MRP Count” helps to find the most repeat patterns in the genome sequences. It is discovered that the pattern “TTGTT” is the most repeat pattern in the COVID-19 genome. In future, it is planned to develop a feature selection method which will help to achieve better accuracy with minimum number of features which in turn minimize the computation time.

REFERENCES

1. Human Coronavirus Types | CDC [Internet]. Cdc.gov. 2021 [cited 7 August 2021]. Available from: https://www.cdc.gov/coronavirus/types.html
2. WHO Coronavirus (COVID-19) Dashboard [Internet]. Covid19.who.int. 2021 [cited 7 August 2021]. Available from: https://covid19.who.int/
3. Onan A. A fuzzy-rough nearest neighbor classifier combined with consistency-based subset evaluation and instance selection for automated diagnosis of breast cancer. Expert Syst. Appl. 2015;42(20):6844-52.
4. Bonidia R, Sampaio L, Domingues D, Paschoal A, Lopes F, de Leon Ferreira de Carvalho A et al. Feature Extraction Approaches for Biological Sequences: A Comparative Study of Mathematical Models. 2020.
5. Zhang H, Kinsner W. Feature extraction from DNA sequences by multifractal analysis: Engineering in Medicine and Biology Society. 23rd Annual International Conference of the IEEE. Istanbul, Turkey: IEEE; 2001. p. 1567-1572.
6. Muhammod R, Ahmed S, MdFarid D, Shatabda S, Sharma A, Dehzangi A. PyFeat: a Python-based effective feature generation tool for DNA, RNA and protein sequences. Bioinformatics. 2019;35(19):3831-3.
7. Rahman M. FastFeatGen: Faster Parallel Feature Extraction from Genome Sequences and Efficient Prediction of DNA N^6-Methyladenine Sites. Computational Advances in Bio and Medical Sciences. 2020;52-64.
8. Nikam R, Gromiha M. Seq2Feature: a comprehensive web-based feature extraction tool. Bioinformatics. 2019;35(22):4797-9.
9. Liu Z, Jiao D, Sun X. Classifying Genomic Sequences by Sequence Feature Analysis. Genomics, Proteomics & Bioinformatics. 2005;3(4):201-5.
10. You W, Wang K, Li H, Jia Y, Wu X, Du Y. Classification of DNA Sequences Basing on the Dinucleotide Compositions. 2009 Second International Symposium on Computational Intelligence and Design. 2009;2:390-4.

11. Zhou Q, Jiang Q, Wei D. A new method for classification in DNA sequence. 2011 6th International Conference on Computer Science & Education (ICCSE). 2011;218-221.
12. Mathew A, John J, Thomas T. Classification of Mutated Cancer Genome Using Machine Learning Approaches. Int. J. Adv. Res. Comput. Commun. Eng. 2014;3(8).
13. Zhao J, Yang X, Li J, Tang Y. DNA Sequences Classification Based on Wavelet Packet Analysis. Wavelet Analysis and Its Applications. 2001;424-429.
14. Vijayan K, Nair V, Gopinath D. Classification of Organisms using Frequency-Chaos Game Representation of Genomic Sequences and ANN. 2009.
15. Sathish Kumar S, Duraipandian N. An Effective Identification of Species from DNA Sequence: A Classification Technique by Integrating DM and ANN. Int J Adv Comput Sci Appl. 2012;3(8):104-14.
16. Kristensen T, Guillaume F. Classification of DNA Sequences by a MLP and SVM Network. International Conference on Bioinformatics & Computational Biology (BIOCOMP). The Steering Committee of the World Congress in Computer Science, Computer Engineering and Applied Computing (WorldComp); 2013.
17. Kassim N, Abdullah A. Classification of DNA Sequences Using Convolutional Neural Network Approach. Innovations in Computing Technology and Applications [Internet]. UTM Computing Proceedings; 2017.
18. Tampuu A, Bzhalava Z, Dillner J, Vicente R. ViraMiner: Deep learning on raw DNA sequences for identifying viral genomes in human samples. PLOS ONE. 2019;14(9):e0222271.
19. Bhonde S, Bhati M, Prasad J. Predictive Analytics to Combat with COVID-19 using Genome Sequencing. SSRN Electron. J. 2020.
20. Habib P, Alsamman A, Saber-Ayad M, Hassanein S, Hamwieh A. COVIDier: A Deep-learning Tool For Coronaviruses Genome And Virulence Proteins Classification. 2020.
21. Onan A, Korukoğlu S, Bulut H. Ensemble of keyword extraction methods and classifiers in text classification. Expert Syst. Appl. 2016; 57:232-47.
22. Onan A. Classifier and feature set ensembles for web page classification. J. Inf. Sci. 2015;42(2):150-65.
23. Onan A, Korukoğlu S. A feature selection model based on genetic rank aggregation for text sentiment classification. J. Inf. Sci. 2016;43(1):25-38.
24. Onan A. Mining opinions from instructor evaluation reviews: A deep learning approach. Comput. Appl. Eng. Educ. 2019;28(1):117-38.
25. Onan A, Tocoglu M. A Term Weighted Neural Language Model and Stacked Bidirectional LSTM Based Framework for Sarcasm Identification. IEEE Access. 2021;9:7701-7722.
26. Sieber P, Platzer M, Schuster S. The Definition of Open Reading Frame Revisited. Trends Genet. 2018;34(3):167-70.
27. Onan A. On the Performance of Ensemble Learning for Automated Diagnosis of Breast Cancer. Adv. Intell.Syst.Comput. 2015;119-29.
28. Thomas J, Thomas S, Sael L. Feature versus raw sequence: Deep learning comparative study on predicting pre-mirna. arXiv preprint arXiv:1710.06798. 2017 Oct 17.
29. ONAN A. Sentiment analysis on massive open online course evaluations: A text mining and deep learning approach. Comput. Appl. Eng. Educ. 2020;29(3):572-589.
30. National Center for Biotechnology Information [Internet]. Ncbi.nlm.nih.gov. 2021 [cited 7 August 2021]. Available from: https://www.ncbi.nlm.nih.gov/
31. Raschka, Sebastian, STAT 451: Introduction to Machine Learning Lecture Notes. 2020.
32. Rasheed Z, Rangwala H. TAC-ELM: Metagenomic Taxonomic Classification with Extreme Learning Machines. InBICoB 2011 (pp. 92-97).

© 2021 by the authors. Submitted for possible open access publication under the terms and conditions of the Creative Commons Attribution (CC BY NC) license (https://creativecommons.org/licenses/by-nc/4.0/).