mathematics

MDPI

Article

COVID-19 Genome Sequence Analysis for New Variant Prediction and Generation

Amin Ullah [1], Khalid Mahmood Malik [2], Abdul Khader Jilani Saudagar [3,*], Muhammad Badruddin Khan [3], Mozaherul Hoque Abul Hasanat [3], Abdullah AlTameem [3], Mohammed AlKhathami [3] and Muhammad Sajjad [4,5,*]

1 CORIS Institute, Oregon State University, Corvallis, OR 97331, USA
2 Department of Computer Science and Engineering, Oakland University, Rochester, MI 48309, USA
3 Information Systems Department, College of Computer and Information Sciences, Imam Mohammad Ibn Saud Islamic University (IMSIU), Riyadh 11564, Saudi Arabia
4 Color and Visual Computing Lab, Department of Computer Science, Norwegian University of Science and Technology (NTNU), 2815 Gjøvik, Norway
5 Digital Image Processing Laboratory, Department of Computer Science, Islamia College Peshawar, Peshawar 25000, Pakistan
* Correspondence: aksaudagar@imamu.edu.sa (A.K.J.S.); muhammad.sajjad@icp.edu.pk (M.S.)

Abstract: The new COVID-19 variants of concern are causing more infections and spreading much faster than their predecessors. Recent cases show that even vaccinated people are highly affected by these new variants. The proactive nucleotide sequence prediction of possible new variants of COVID-19 and developing better healthcare plans to address their spread require a unified framework for variant classification and early prediction. This paper attempts to answer the following research questions: can a convolutional neural network with self-attention by extracting discriminative features from nucleotide sequences be used to classify COVID-19 variants? Second, is it possible to employ uncertainty calculation in the predicted probability distribution to predict new variants? Finally, can synthetic approaches such as variational autoencoder-decoder networks be employed to generate a synthetic new variant from random noise? Experimental results show that the generated sequence is significantly similar to the original coronavirus and its variants, proving that our neural network can learn the mutation patterns from the old variants. Moreover, to our knowledge, we are the first to collect data for all COVID-19 variants for computational analysis. The proposed framework is extensively evaluated for classification, new variant prediction, and new variant generation tasks and achieves better performance for all tasks. Our code, data, and trained models are available on GitHub (https://github.com/Aminullah6264/COVID19, accessed on 16 September 2022).

Keywords: artificial intelligence; deep learning; RNA analysis; SARS-CoV-2; self-attention; uncertainty analysis; variational autoencoders

MSC: 68T07

Citation: Ullah, A.; Malik, K.M.; Saudagar, A.K.J.; Khan, M.B.; Hasanat, M.H.A.; AlTameem, A.; AlKhathami, M.; Sajjad, M. COVID-19 Genome Sequence Analysis for New Variant Prediction and Generation. *Mathematics* **2022**, *10*, 4267. https://doi.org/10.3390/math10224267

Academic Editor: Victor Leiva

Received: 17 September 2022
Accepted: 2 November 2022
Published: 15 November 2022

1. Introduction

The first human SARS-coronavirus-2 case was reported in Wuhan, China, in December 2019, and this disease later became known as COVID-19 [1]. In a short interval of time, it spread all over the world, with statistical studies showing that the number of cases increased exponentially. As of December 2021, more than 267 million confirmed cases and 5.27 million deaths had been recorded worldwide [2]. In its early spread, the World Health Organization (WHO) set the COVID-19 risk assessment at the regional and global levels to "Very High". Due to the absence of effective COVID-19 treatments and its self-evolution, while spreading in different regions and races, the early detection of this life-threatening infectious disease is crucial. Since the start of this pandemic, several approaches have become available to diagnose COVID-19, including nucleic acid-based techniques called

polymerase chain reaction (PCR) [3–5], where patients need to undergo a clinical test, a computed tomography (CT) scan [6], and a chest X-ray [7]. PCR-based methods with a backbone of deoxyribonucleic acid (DNA) or ribonucleic acid (RNA) sequences for analysis utilize basic local alignment search tools (BLASTs) and fast-all (FASTA)-like alignment methods to detect the virus in the sequence. These methods aim to find common patterns in two given sequences, which are very effective in finding similarities. When a patient becomes infected with the virus, the samples and genomes taken from the patient are sequenced. The sequenced genomes were compared in GenBank (NCBI) to identify the virus. The four building blocks of the genome sequences are adenine (A), cytosine (C), guanine (G), and thymine (T). GenBank maintained a BLAST server to check genomic sequence similarity and find the patterns in the nucleotides. GenBank has 106 billion nucleoid bases of 108 million distinct sequences, with 11 million new ones added in the past year [8]. The BLAST- and FASTA-based methods rely on the alignment score between the query of the reference sequences in the dataset. The main disadvantages of alignment-based methods are that the classification prediction entirely depends on selecting one of the few initial alignments and hyperparameters. Furthermore, these techniques are not effective and efficient for matching complete genome sequences in millions of DNA databases, such as NCBI GenBank.

Machine learning and deep learning techniques were initially applied to analyze DNA sequences and overcome alignment-based methods' challenges. For instance, the researchers in [9,10] utilized CpG island features and machine learning methods to classify different families of SARS, including AlphaCoV, BetaCoV-1, MERS-CoV, HKU1-CoV, NL63-CoV, and 229E-CoV. However, the current works are limited to the prediction of only COVID-19. To the best of our knowledge, no study has been presented which predicts the evolving deadly variants of COVID-19 that are currently causing breakthrough connections and threatening the economy. Likewise, there is also no approach to predicting the possible sequences of non-existing variants. To address these research challenges, the main contributions of this paper are as follows:

- We proposed a convolutional neural network (CNN) to classify the COVID-19 variants. The proposed model utilizes 1D convolutions, batch normalization, and self-attention layers to extract discriminative features from the nucleotide sequences. Self-attention is employed to cope with the relationship and mutation of adenine (A), cytosine (C), guanine (G), and thymine (T) in the sequence.
- The detection of new variants of COVID-19 is an important task for the scientific community to develop better healthcare and economic plans to deal with its spread. To the best of our knowledge, our proposed framework is the first attempt to detect new variants. Our framework employs uncertainty calculation via entropy followed by an optimum threshold to detect unknown variants of COVID-19.
- We introduce a novel COVID-19 new variant generation technique, which is used to evaluate the proposed variant detection technique. Our variational autoencoder-decoder (VAE) network is able to predict new sequences with significant similarity with the original coronavirus and its variants using the BLAST method. We also believe that the proactive nucleotide sequence prediction of possible new variants of COVID-19 could assist vaccine providers in increasing the efficacy of vaccines.

The remainder of this paper is organized as follows: Section 2 discusses the state-of-the-art literature related to COVID-19 diagnosis using artificial intelligence techniques, Section 3 describes the proposed framework, Section 4 explains the data collection, experimental setup, and results, and Section 5 concludes the paper with a summary of the proposed work and future directions in COVID-19 research.

2. Related Works

The precise detection and diagnosis of the rapidly spreading coronavirus have become challenging due to excessive mutations in its structure [11]. For this purpose, researchers have proposed many techniques using machine learning and deep learning technology for

different data modalities, such as chest X-ray images, genomic sequences, and cough audio signals [12], to detect COVID-19-infected persons. For instance, Asraf et al. [13] presented a short survey in which they explored the importance of deep learning approaches by applying different modalities of data to control the novel COVID-19 pandemic. Using genomic data, Arsalan [9] used a support vector machine, naïve Bayes, k-nearest neighbor (KNN), and random forest for COVID-19 classification, where his method achieved 93% accuracy using a decision tree. In a similar approach [10], CpG island features and a KNN-based technique were proposed for human genome classification using a dataset recorded in 2019 for novel coronavirus and other SARS virus families. In another method, He et al. [14] proposed a nucleic transformer for COVID-19 classification, showcased the extraction of promoter motifs from learned attention and discussed how direct visualization of self-attention maps assists in informed decision-making using deep learning models. Dasari et al. [15] proposed an explainable deep learning model using CNN, and long short-term memory (LSTM)-based features called EDeep VPP and EDeep VPP-hybrid. Their framework performs model interpretability, which allows them to automatically extract important features that have a primary role in predicting COVID-19 from genome sequences. They obtained 0.992 mean AUC-ROC and 0.990 AUC-PR on 19 human metagenomic contig experimental datasets using 10-fold cross-validation. Currently, a very limited number of methods have been proposed using genomic data, and the state-of-the-art method has very limited accuracy and a high false prediction rate.

The majority of the work on COVID-19 using machine learning and deep learning is based on chest X-ray data. For instance, Wang et al. [16] collected COVID-19 patients' CT images and used a transfer learning mechanism to train the inception CNN model to detect COVID-19. They achieved 82.9% accuracy, and the model was tested on external samples and achieved 73% precision. Similarly, Barstugan et al. [17] classified CT images of COVID-19 into three major classes: COVID-19, influenza, and viral pneumonia cases. The dataset was obtained from a hospital in Zhejiang city in China, which contained a total of 618 pictures, 214 from 110 infected people with COVID-19, 224 photos with viral pneumonia and influenza A, and 175 photos of normal people. They built a 3-dimensional significant learning model and achieved 87.6% classification accuracy. In another method, Gozes et al. [18] developed robust 2D-3D deep learning models and combined them with clinical understanding for the classification of COVID-19 CT images, where they achieved 96.4% sensitivity, 98% specificity, and 0.996 AUC. Similarly, Ozkaya et al. [19] proposed deep feature fusion and ranking techniques for COVID-19 analysis. They achieved a display matrix with 95.60% accuracy, 93.3% specificity, 95.60% sensitivity, 97.87% precision, 91.29% MCC and 97.77% F1-score matric execution gained. Mucahid et al. [17] gathered 150 CT pictures of COVID-19. They used five feature extraction techniques, which were improved by applying them on small patches to obtain many features, resulting in high precision. Muhammad et al. [20] proposed a coordinated model for the acknowledgment of COVID-19 in Mexico and used five classification algorithms, DT, LR, NB, SVM, and ANN. Among these estimations, the decision tree classifier achieved an incredible accuracy of 94.99%. Comparable to this, Muhammad et al. [21] proposed four predictive data mining models for novel COVID-19-infected patients' recovery; among these algorithms, decision trees achieved the highest accuracy of 99.85%. In another study, Ali et al. [21] presented a comparative study of five deep convolution neural network models, including InceptionV3, Inception-ResNetV2, ResNet50, ResNet101, and ResNet152, for the detection of COVID-19 disease using X-ray images. Song et al. [22] assembled chest CT scans of 88 patients diagnosed with COVID-19 and 86 healthy individuals and 101 patients infected with bacterial pneumonia for comparison from different clinical centers in China, which was then used to diagnose novel coronavirus.

State-of-the-art COVID-19 machine learning classification and detection methods have performed significantly well on different data modalities. However, they are limited to binary classes such as COVID-19 and non-COVID-19 or very few classes such as COVID-19, viral pneumonia, and influenza data analysis. In contrast, the proposed framework classi-

fies and detects new variants of COVID-19, which are crucial for the proactive management of new waves of the pandemic.

3. COVID-19 Nucleotide Sequences Analysis Framework

This section discusses the proposed framework for COVID-19 nucleotide sequence classification and the prediction of new variants. The proposed framework consists of three phases, as shown in Figure 1 and mathematically explained in Algorithm 1. First, the classification network determines the prediction of a COVID-19 variant for a given nucleotide sequence. Second, an uncertainty check is performed to check whether this prediction is from the existing or new variants. The third phase of this framework aims at generating possible sequences of new variants based on the evolutions of existing variants by using random noise with VAE. The details of the proposed framework are explained in subsequent sections.

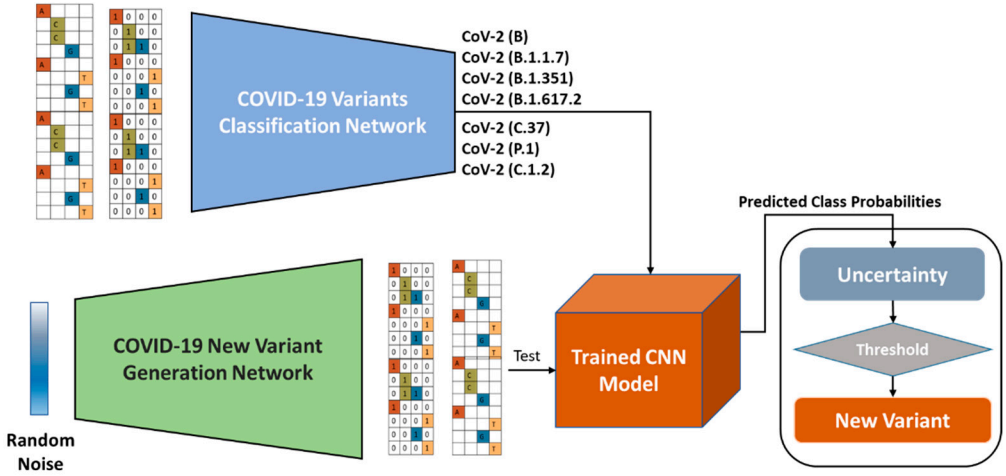


Figure 1. The proposed framework for COVID-19 variant classification and new variant prediction and generation for the evaluation of the proposed method.

3.1. Nucleotides Data Preprocessing

DNA analysis for any application requires text data processing, utilizing either nucleotides or protein sequences. Machine learning and deep learning techniques have been widely applied in various test data analysis tasks, such as spam detection, text sentiment analysis, and topic categorization. In the proposed framework, we employ CNN for feature extraction and classification, which takes numerical tensors as the input, i.e., 2D image pixel data. On the other hand, nucleotide sequences are consecutive alphabetical letters in the form of 1D text data. For that reason, the alphabetical letters need to be transformed into numerical representations so that they can be processed via CNN.

In the text analysis literature, researchers have used different techniques to transform alphabetical sequences into numerical representations, such as lookup tables, which find each word's corresponding vector in the given vocabulary, also known as word vectors. The vocabulary length is fixed in the lookup table-based representation. Word2vec is its updated form, where the vocabulary is learned via a neural network to produce word embeddings based on its position in the sentence. These two representations are compelling in data analysis, have a large-scale vocabulary, and each word has a different meaning in different appearances. However, in the nucleotide sequence, we have only four characters. Therefore, we employed a one-hot encoding [23] scheme for the transformation of nucleotide sequences to numerical values. In one-hot encoding, binary variables are introduced for the given words. For instance, in nucleotide sequences formed from four characters, we will have four binary variables in the one-hot sequence, where one variable is 1 and the other is 0, representing one character. For the sequence "AGCT", one-hot encoding can be calculated as given in Equation (1).

$$\text{“AGCT”} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \quad (1)$$

3.2. COVID-19 Variants Classification Network

Convolutional neural networks (CNNs) are fundamentally designed for 2D image data analytics, including object classification, detection, localization, etc. [24]. The success of CNNs is due to learned filters designed during the training process of the network to extract discriminative features for a given task. Due to this property, researchers have utilized CNNs in the form of 1D-CNNs for various modalities in data analytics, including natural language processing, electricity energy consumption prediction, and RNA sequence analysis. Inspired by these techniques, we also utilized 1D-CNN for COVID-19 variant classification tasks. CNN is one of the best deep learning techniques to extract features from raw data because it can select features automatically by tuning the applied filters in the stacked layers. In the genome nucleotide dataset, the length of sequences is very long, and capturing long-term dependencies between the sequences is very challenging. Therefore, CNN is the best choice to classify this type of data.

In 1D-CNNs, the filters are convolved over the one-dimensional input in a sliding window manner. Let us consider the COVID-19 preprocessed one-hot sequence as input x, where we have four channels in one-dimensional form, and each channel represents adenine (A), cytosine (C), guanine (G), and thymine (T).

3.3. Self-Attention

CNN layers are very powerful for extracting features from given ordered data. However, to achieve better performance, multiple stacked layers are needed to globally accumulate locally captured filter activations. Furthermore, filters are applied to identify likely relationships and cannot relate long distances among sequences. Learning long-range dependencies is an important aspect in natural language processing (NLP). To capture such dependencies, transformer blocks have been introduced in NLP. Self-attention is a typical form of transformer that utilizes pairwise entity interactions with a content-based addressing mechanism to represent a rich hierarchy of associative features across long sequences. Self-attention is computationally primitive compared to multiple stacked convolutional layers in the model. The self-attention takes X feature maps as the input and repeats them as key, query, and value vectors [25]. Different convolutional filters convolve these vectors to have different projections of inputs and weights, where the features from the key and query vectors are aggregated by pairwise dot product and Softmax function. Finally, self-attention is achieved by the dot product of the features from the value vector and Softmax output. This setup helps the model capture the internal long-term sequential dependencies in the given X feature maps.

The proposed classification network consists of four 1D convolutional, two pooling, batch normalization, self-attention, dropout, fully connected, and Softmax layers, as visualized in Figure 2. The kernel sizes in the convolutional layers are kept at four due to the number of characters—‘A’, ‘C’, ‘G’, and ‘T’—which assists the proposed network in learning the patterns between them. The convolutional layers consist of 128, 64, 32, and 16 learnable kernels. The max pooling layer with a kernel size of four is added to reduce the dimensions of the feature for efficient processing. The batch normalization layer is added to avoid overfitting and achieve better performance. These normalized features are then passed to the self-attention layer. Furthermore, a dropout layer with value = 0.5 is added after the self-attention layer to regularize the network and generalize it for unseen data. Then, a fully connected layer yields logits for the number of output classes propagated forward to the Softmax layer to obtain the final predicted probability values.

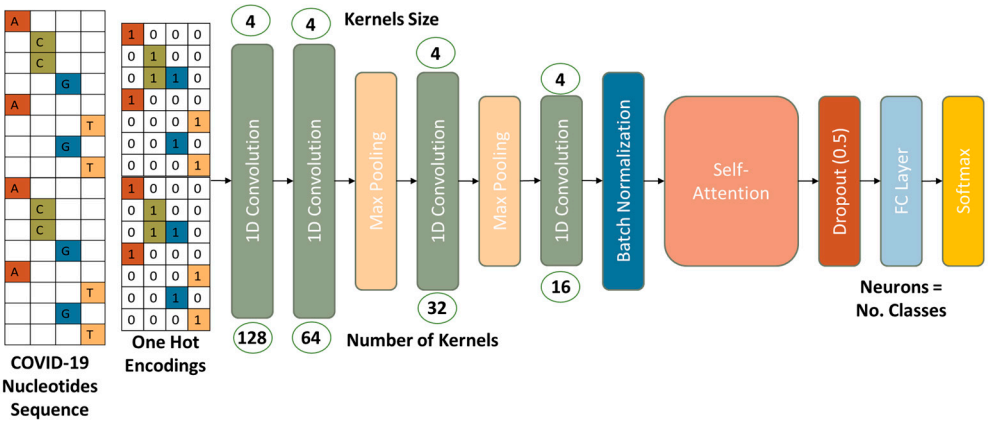


Figure 2. The proposed self-attention 1D-CNN network for COVID-19 variant classification.

3.4. New Variant Detection

For decades, entropy has been considered an effective method to predict uncertainty and disorder in a given probability distribution [26]. In the proposed framework, we used it to predict uncertainty in the classification of COVID-19 variants. For instance, if the predicted probabilities of a classification network are equally distributed among (more or less) all variants, it means that the given input is not from any of the trained classes in the network. Furthermore, it is obvious that the difference between these variants is very small; however, with the generalization ability for unseen data of the proposed classification network, we can achieve a confidence score greater than 90% for the correctly predicted class. This helps us to differentiate between seen and unseen variants of COVID-19. The class uncertainty U is calculated using Equation (2), where c is the class predicted probability and i is the class index.

$$U(C) = -\sum_{i=1}^{n} P\left(c^i\right) logP\left(c^i\right) \tag{2}$$

The decision to predict the new variant is predicted based on a predefined threshold value; if $U(C)$ is greater than that threshold, we consider it a new variant of COVID-19.

3.5. COVID-19 Variant Generation

Synthetic data generation has been an active area of research in the past decade, for which researchers have proposed various techniques, including encoder-decoder networks, generative adversarial networks (GANs), and VAEs. These techniques have been very successful in computer vision and natural language processing domains, such as human face generation, cartoon characters, text-to-image translation, and human pose generation. However, they have not been explored well in sensitive data such as DNA sequence generations. For instance, Nathan et al. [27] proposed a GAN network for synthetic DNA sequence generation that can be tuned to have the desired properties for protein binding microarrays. This method achieved outstanding results for this task; however, the length of the synthetic sequence was concise, containing only fifty nucleotides, and the creation was straightforward for any generative model's family to outperform in the results. On the other hand, such methods are ineffective for sequence generation, where the long-term dependencies are important to cope with the desired patterns for analysis, such as COVID-19 nucleotide sequences. This is observed in experiments using different lengths of sequences for COVID-19 new variant generation using GANs.

The newly discovered COVID-19 variants are highly infectious and cause vaccine breakthrough cases. Therefore, their detection is important for the diagnosis and proper treatment of affected humans [28]. This encouraged us to develop a method for the early detection of new COVID-19 variants. The proposed detection method is very effective

in discriminating between existing and new variants. However, the challenge is how to evaluate the proposed method and where to obtain new variants for the evaluation of the proposed method. For this, we followed two approaches: (1) we took a few variants from the existing ones for the testing of the proposed method, and (2) we proposed a new variant generation technique that provided us with a synthetic COVID-19 variant for the testing of the proposed method.

To generate a synthetic COVID-19 sequence, we first employed a GAN network with convolutional layers for feature representation in both generator and discriminator networks. However, despite the varied tuning of hyperparameters in both networks' layers, we were not able to converge the model to generate an equivalent sequence. We observed that GAN did not work because of the long nucleotide sequence length of at least three thousand characters. Next, we investigated a variant of GAN known as VAE for this task and proposed an effective network to generate a new COVID-19 variant, which is discussed in detail in the upcoming section.

3.6. Variational Autoencoder-Decoder

The VAE is from the family of generative models, and predicts the probability density function of the input data, where the latent space is a probability distribution, unlike traditional autoencoders, where latent space is a high-level representation of the input data, as given in Figure 3. The VAE makes use of Bayesian inference and the robustness of deep neural networks to acquire a nonlinear low-dimensional latent space [29]. First, the input data x are fed into an encoder, which learns the parameters of distributions $Q(z \mid x)$. Second, Bayesian inference is obtained by sampling the latent representation z with a prior distribution $p(z)$ (standard Gaussian N (0, I)), where I is the identity matrix. In VAE, the encoder generates mean μ and standard deviation σ vectors, where z can be calculated as given in Equation (3).

$$z_i = \mu_i + \sigma_i \cdot \varepsilon \tag{3}$$

where i is the index of each component in the μ and σ vectors, and ε is a Gaussian distribution ($\varepsilon \sim$ N (0, 1)). Conceptually, the decoder reconstructs the input from the given latent space, i.e., $P(x \mid z)$. However, in the proposed method, we generate COVID-19 variants xv from the input x. Therefore, the reconstruction loss is replaced by variant construction loss and can be represented as given in Equation (4).

$$L_{variant\_cons} = -E\, Q(x)[logP(z)] \tag{4}$$

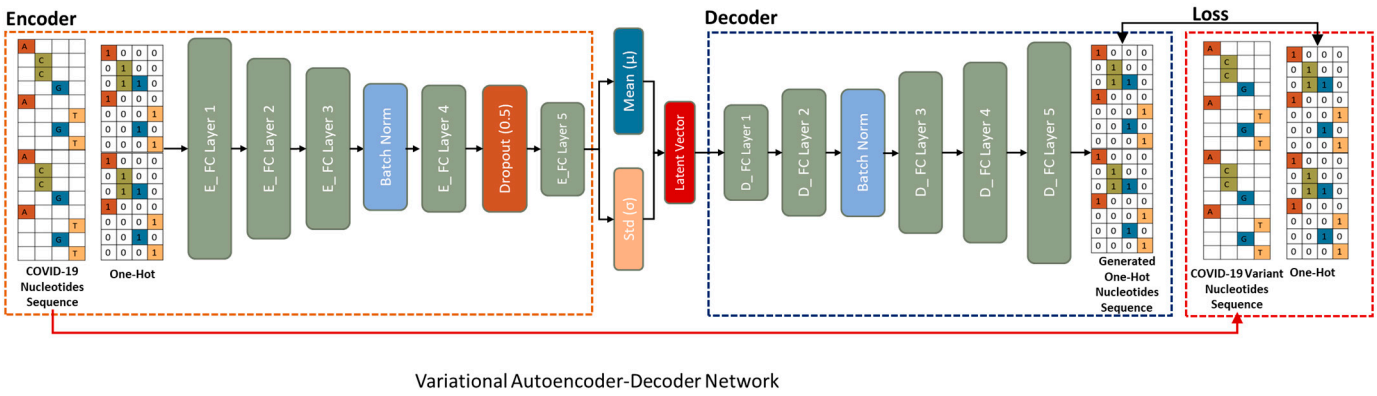


Figure 3. The proposed VAE network for COVID-19 variant generation. The network takes the original COVID-19 nucleotide sequence as the input and outputs a predicted COVID-19 variant. The reconstruction loss is calculated between the variant of COVID-19 not with the original sequence because we want to converge our model to generate different variants.

Furthermore, the VAE also minimizes the loss between the generated latent space and the specified normal distribution. To do this, Kullback–Leibler (Dkl) divergence is used to measure the proximity between two densities $Q(z \mid x)$ and Gaussian distribution $p(z)$, as given in Equation (5).

$$L_{KLdiv} = D_{Kl}\left(Q(x) \mid\mid P(z)\right) \tag{5}$$

The overall loss to converge to the proposed COVID-19 variant generation network can be represented as given in Equation (6).

$$L = L_{variant\_cons} + L_{KLdiv} \tag{6}$$

Algorithm 1: COVID-19 variant prediction

Steps:

```
M ← Load Attention-CNN model
seq ← Load nucleotide sequences
th ← 0.15        # threshold for uncertainty prediction
one_seq ← ONE_HOT_Conversion (seq)
preds ← M (one_seq)
score ← ENTROPY (preds)
if (score > th):
    print ('New variant')
else:
    varinat ← MAX (preds)
    print (varinat)
```

4. Experimental Results and Discussion

In this section, the proposed COVID-19 variant classification and new variant prediction and generation framework are experimentally evaluated for COVID-19 nucleotide and protein sequence data. The overall accuracy metric is used for the classification network, and receiver operating characteristic (ROC) curves, area under the curve (AUC) values, dot plot similarity between sequences, and sequence alignment are utilized to evaluate the performance of the new variant generation network in ablation studies and comparisons with state-of-the-art methods. The experimental environment consists of a Windows 10 operating system installed over a Corei7-1050 processor, 16 GB RAM, and aided by a dedicated 8 GB RTX-2070 GPU. The code is implemented in Python-3.7 and the deep learning framework 'Kears = 2.4.0'.

4.1. Dataset

COVID-19 data analysis is a recently emerging topic; therefore, data are not available as a benchmark for analysis. The existing methods utilized the limited COVID-19 data along with other known virus data from the NCBI website for the analysis. However, we collected data for all diagnosed COVID-19 variants from NCBI for the evaluation of the proposed framework. The data were collected by searching the scientific name of the virus variant on the NCBI website, which retrieves the recently available data for that variant. The Centers of Disease Control and Prevention (CDC) in the United States of America divided these variants into different categories based on their severity level, including Variant Being Monitored (VBM), Variant of Interest (VOI), Variant of Concern (VOC), and Variant of High Consequence (VOHC). In this research, we analyzed collected data from all of the categories for the effective and efficient classification and prediction of new variants based on these variant data. The overall statistics of the collected data are given in Table 1.

Table 1. Data collected from NCBI for the classification and new variant prediction tasks.

No.	Variants Scientific Codes	Known Names	No. of Nucleotide Sequences Length = 30,000	No. of Protein Sequences Length = 3000	Classification Network Training	New Variant Prediction Test
1	B	SARS-CoV-2	1500	1109	✓	-
2	B.1.1.7	CoV-Alpha	1500	1053	✓	-
3	B.1.351	CoV-Beta	1500	1058	✓	-
4	B.1.617.2	CoV-Dalta	1015	1078	✓	-
5	C.37	CoV-Lambda	661	1056	✓	-
6	P.1	CoV-Gamma	1500	1073	✓	-
7	B.1.525	CoV-Eta	1500	1720	✓	-
8	C.1.2	-	119	14	-	✓
9	P.2	CoV-Zeta	1047	1059	-	✓
10	B.1.427	CoV- Epsilon	1500	1071	-	✓

4.2. Evaluation of the Classification Network

A detailed discussion of the network architecture is given in Section 3.2. The proposed COVID-19 variant classification network is compared with different DNA/RNA analysis methods, and the statistics are shown in Table 2. The training and validation loss and accuracy recorded during training epochs are visualized in Figure 4. To our knowledge, no existing method has been proposed for COVID-19 variant classification. Therefore, we tried our best to run the codes of the given methods for the proposed task. We compared the results of nucleotide sequence data with handcrafted feature-based and deep learning-based approaches. The dinucleotide composition (DNC), trinucleotide composition (TNC), tetranucleotide composition, and composite nucleotide features were used in [30] for splicing site prediction. The DNC, TNC, TetraNC, and composite NC encode different nucleotides to generate feature vectors. For instance, DNC encodes two nucleotides and generates a $4 \times 4 = 16$-dimensional feature vector; each nucleotide is converted to a 4-dimensional one-hot vector.

Table 2. Comparison with state-of-the-art nucleotide sequence analysis methods for COVID-19 variant classification on the test set of the dataset.

Method	Overall Accuracy
DNC [30]	64.7
TNC [30]	63.1
TetraNC [30]	64.5
Composite NC [30]	66.9
DeepBind [31]	85.6
DeeperBind [31]	89.2
MPPIF-Net [32]	85.7
Our baseline (Protein)	**88.6**
Our Attention-CNN (Protein)	**90.6**
Our baseline (Nucleotides)	**93.4**
Our Attention-CNN (Nucleotides)	**96.1**

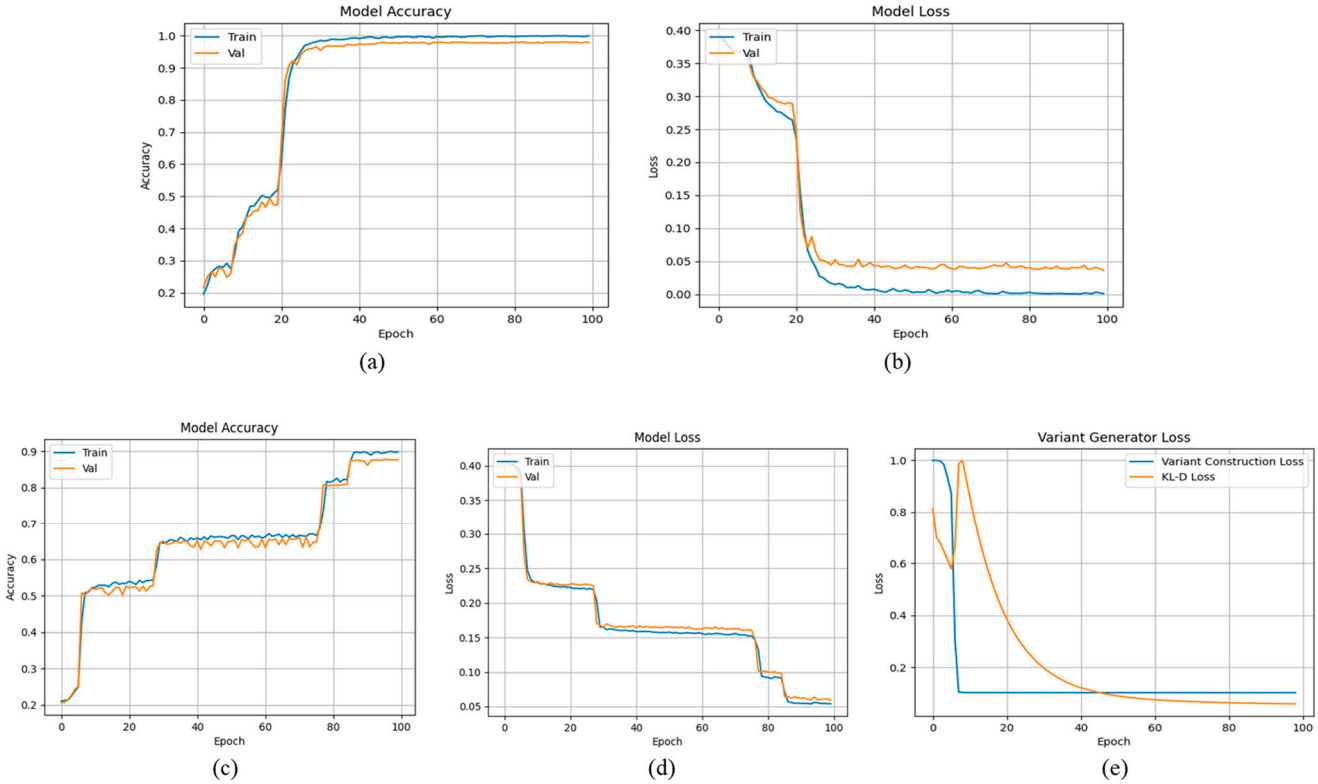
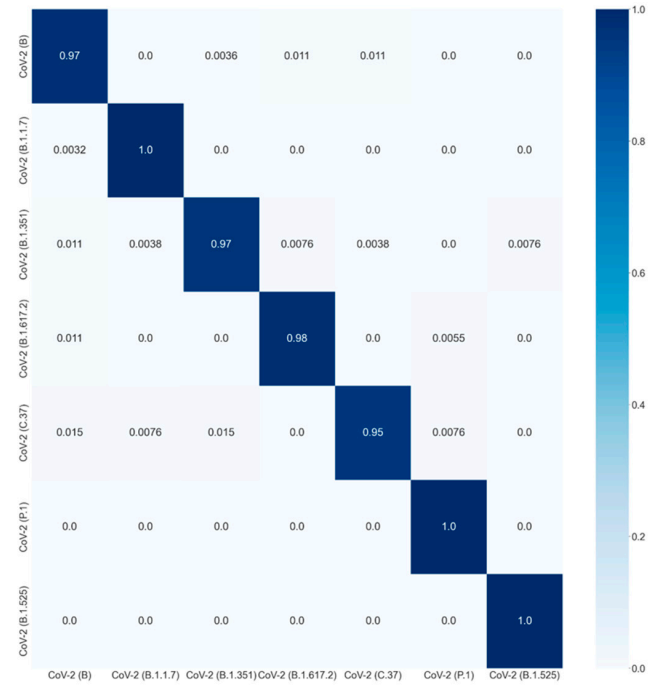


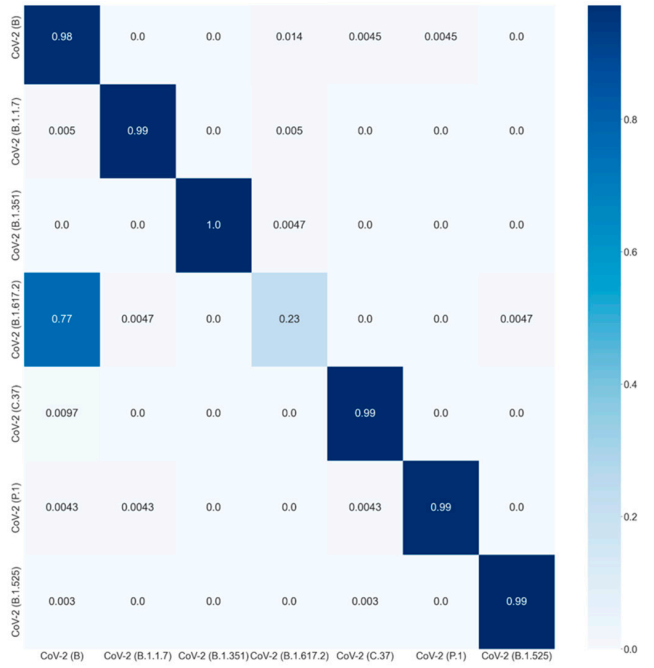
Figure 4. The training and validation performance recoded during training epochs: (**a**) accuracy for nucleotides, (**b**) loss for nucleotides, (**c**) accuracy for protein, (**d**) loss for protein sequence data, and (**e**) shows the new variant construction loss.

Similarly, TNC encodes three relative nucleotides, TetraNC encodes four nucleotides, and composite NC is the fusion of all of these features. We tuned that work for COVID-19 variant classification, where we achieved overall accuracies of 64.7%, 63.1%, 64.5%, and 66.9% for DNC, TNC, TetraNC, and composite NC, respectively. The DeepBind and DeeperBind [31] networks were used to predict the sequence specificities of DNA-binding proteins. DeepBind only used CNN features, while DeeperBind utilized CNN and LSTM for sequence representation. For the underlying task, these techniques achieved 85.6% and 89.2% accuracy. MPPIF-Net [32] utilized CNN with multilayer bidirectional LSTM for the identification of plasmodium falciparum parasite mitochondrial proteins, where it achieved state-of-the-art accuracy; however, for the proposed problem, it attained only 85.7% accuracy.

The proposed baseline network consists of only 1D convolutional layers that achieved 88.6% and 93.4% accuracy for protein and nucleotide sequences, respectively. Our attention-CNN for protein and nucleotide sequences achieved 90.6% and 96.1% overall accuracy, respectively, on the test dataset. The confusion matrix created for nucleotide data is given in Figure 5a, while for protein sequences, the confusion matrix is shown in Figure 5b. Overall, our network performs better on nucleotide sequences because it has only four characters, "AGCT", and the network is able to learn the relationship between the characters effectively using attention mechanism. On the other hand, the protein sequences consist of 26 characters that are very hard to encode, and the sequence length is only 3000 characters, due to which the network lacks the ability to discriminate between similar patterns. Therefore, COV-B.1.67.2 is confused with the original COVID-19 samples.



(a)



(b)

Figure 5. The confusion matrix attained using the proposed attention-CNN for the test set of (**a**) nucleotide sequences and (**b**) protein sequences.

4.3. Evaluation of the New Variant Generation Network

The proposed new variant generation technique is evaluated using dot plot similarity [33] and Needleman–Wunsch [34] sequence alignment methods. The dot plot finds similarity between two sequences by arranging one sequence on the horizontal axis and the second sequence on the vertical axis of the plot. When the residues of both sequences match at the same location on the plot, a dot is drawn at the corresponding position. Note that the sequences can be written backwards or forwards; however, the sequences on both axes must be written in the same direction. The closeness of the sequences in similarity will

determine how close the diagonal line of the sequences is. The dot plots for COVID-19 and its different variants are given in Figure 6a–e, while the dot plot between COVID-19 and the generated sequence from random noise using the proposed method is given in Figure 6f. The reason for drawing dot plots between other COVID-19 variants is to show the matching patterns between the variants and the generated syntactic variant. As mentioned earlier, the diagonal line indicates that there is probably a good alignment between sequences, and the other positions show different kinds of mutations in the original sequence.

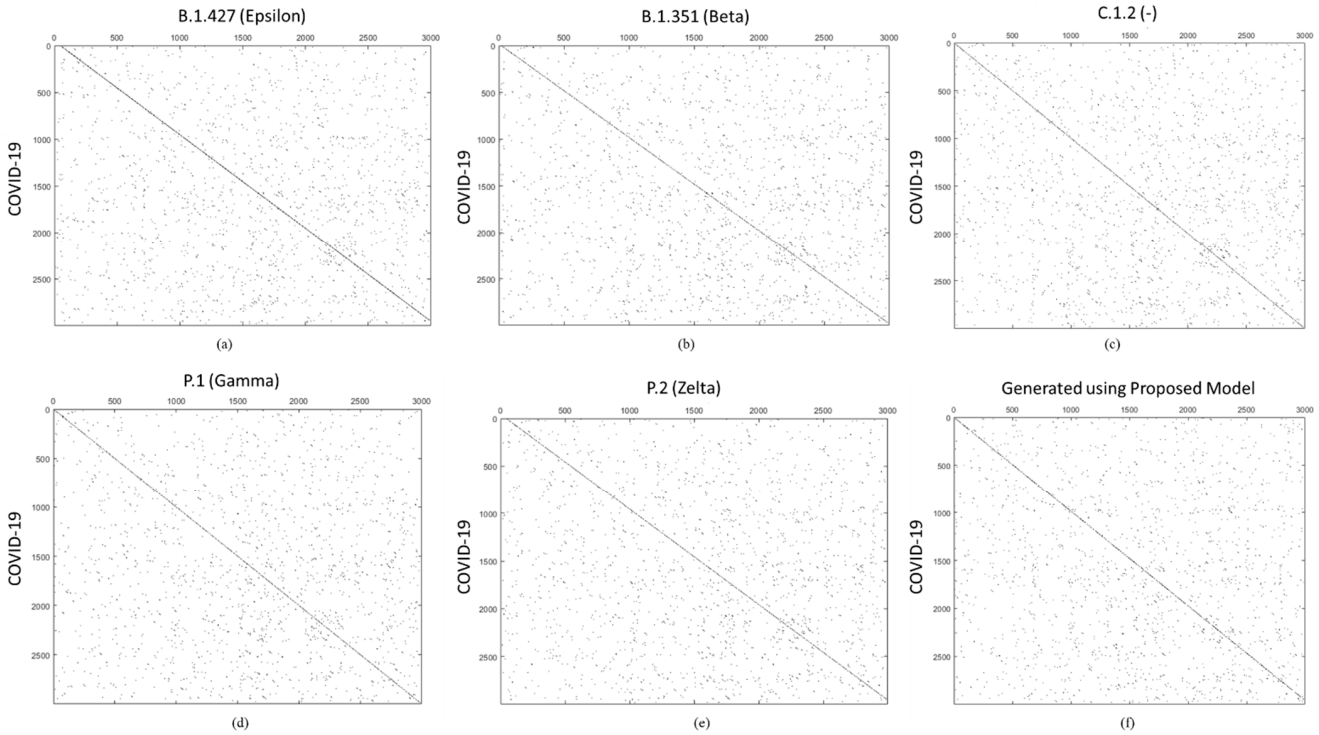


Figure 6. Dot plot similarities between nucleotide sequences of COVID-19 and its variants: (**a**) CoV-Epsilon, (**b**) CoV-Beta, (**c**) CoV-C.1.2, (**d**) CoV-Gamma, (**e**) CoV-Zeta, and (**f**) the new variant generated using the proposed technique. The number on the axis represents sequence length.

Therefore, we claim that our model can learn the COVID-19 sequential patterns and mutations from its variants to generate any possible new variants based on the trained data.

The dot plot similarity proved that our generated sequence is very similar to the original COVID-19 sequence from its diagonal outputs in the plot. Now, we can look at the global alignment of the generated sequence with the original COVID-19 sequence to check for mutation. The sequence alignment between the original COVID-19 and the generated nucleotide sequence is given in Figure 7. The length of the generated sequence is 3000, and we selected a few portions of the sequence from different positions for visualization in the figure. In Figure 7, 'Sequence 1' is the original COVID-19 for reference, and 'Sequence 2' is the query-generated sequence. The consensus sequence is the estimated order of most frequent residues in the nucleotides found at each position in a sequence alignment. Figure 7 shows that the generated sequence mostly agrees with the consensus sequence of the COVID-19 variants, and few places have mutations. Based on these evaluations, we claim that our network can learn to generate new COVID-19 variants from random noise, which can be very beneficial for the detection and prediction of new variants. Finally, this study does not claim that the predicted variants generated using the proposed method exist; rather, it predicts a potential new variant that shares some characteristics with existing variants in terms of mutations from the original COVID-19. We used the dot plot and sequence alignment approaches to demonstrate comparable trends in our investigations. We believe that COVID-19 researchers can use the generated sequences to advance their investigation and develop a diagnosis and vaccine.

Figure 7. Snapshots of different positions from sequence alignment between the original COVID-19 and generated nucleotide sequence. Sequence 1 is the original COVID-19 sequence, and sequence 2 is the generated COVID-19 variant using the proposed technique. Consensus sequence is defined in Section 3.3.

4.4. New Variant Prediction Evaluation

One of the contributions of our paper is the detection of new variants of COVID-19. The question is, how can we evaluate the new variant prediction mechanism? For this, we followed a strategy where we used seven variants for training the classification network and three variants for evaluating new variant prediction, including C.1.2, P.2, and B.1.427. The train and test splits are given in Table 1. The ability of the new variant to predict the proposed technique was assessed using the ROC curve and AUC values. The ROC curve calculates the contrast between the true positive rate (TPR) and the false positive rate (FPR) at different threshold values for classification decisions. In this problem, the TPR reflects the new variant samples detected as new variants, while the FPR reflects the old variant samples detected as new variants. Figure 8 shows that the proposed technique achieved a 0.72 AUC value. The current achieved accuracy is very reasonable to begin a detailed analysis of such sequences from genomic experts for the early treatment of new variants.

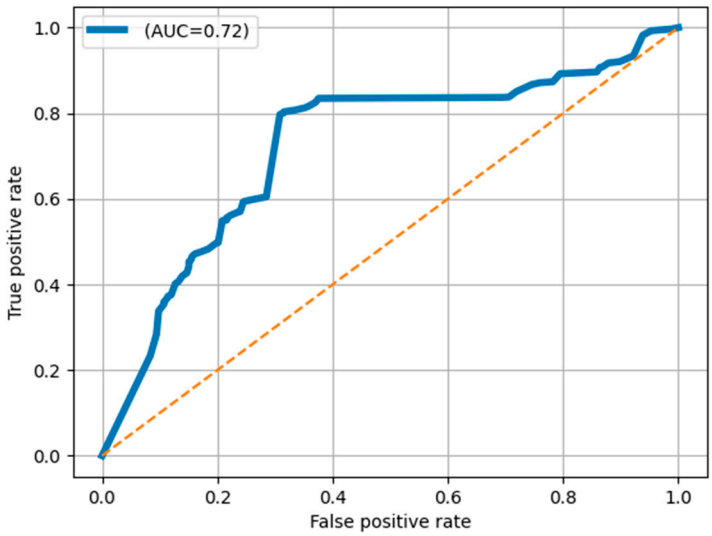


Figure 8. ROC curve and AUC value achieved for new variant prediction using the proposed technique.

5. Conclusions

To date, the classification of COVID-19 variants and the generation of synthetic COVID-19 sequences are unexplored research directions. In this paper, we presented a unified framework for the classification of COVID-19 variants using the CNN and self-attention model, a new variant prediction using uncertainty calculations, and a new variant generation network using VAE. Furthermore, we collected the most recent data for all the COVID-19 variants and ran different baseline techniques for its classification, where our proposed classification network achieved the best performance. Furthermore, the new variant prediction and generation techniques are also extensively evaluated, which proves to be very effective for the underlying tasks.

There are some shortcomings in this study that can be fixed in follow-up investigations. We construct the entropy score based on a single model's prediction, and the uncertainty AUC is quite low. This can be improved by introducing efficient ensemble or Bayesian methods for robust uncertainty prediction. Furthermore, we do not know which patterns of the nucleotide sequences are part of the decision. In the future, we aim to introduce explainable artificial intelligence (XAI) for these COVID-19 analysis tasks so that we can know how the neural networks learn the mutation and how they classify a given sequence.

Author Contributions: Conceptualization, A.U. and M.S.; methodology. A.U., M.S. and K.M.M.; software, M.B.K. and M.H.A.H.; validation, K.M.M., M.B.K., A.A. and M.H.A.H.; formal analysis, M.S., A.U. and M.A.; investigation, A.A., M.A., A.K.J.S.; resources, M.A. and A.K.J.S.; data curation, A.U. and M.H.A.H.; writing—original draft preparation, M.S. and A.U.; writing—review and edit-

ing, K.M.M., M.B.K., A.A., M.H.A.H., M.A. and A.K.J.S.; visualization, A.U.; supervision, A.K.J.S.; project administration, M.B.K., A.A., M.H.A.H., M.A. and A.K.J.S.; funding acquisition, M.B.K., A.A., M.H.A.H., M.A. and A.K.J.S. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the Deputyship for Research & Innovation, project number 959, Ministry of Education, Saudi Arabia.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Acknowledgments: The authors extend their appreciation to the Deputyship for Research Innovation, Ministry of Education in Saudi Arabia for funding this research work through project number 959. Furthermore, any thoughts, conclusions and assumptions or recommendations expressed in this paper are those of the authors and do not necessarily reflect the views of any official organization.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Lv, M.; Luo, X.; Estill, J.; Liu, Y.; Ren, M.; Wang, J.; Wang, Q.; Zhao, S.; Wang, X.; Yang, S.J.E. Coronavirus disease (COVID-19): A scoping review. *Eurosurveillance* **2020**, *25*, 2000125. [CrossRef] [PubMed]
2. World Health Organization. *COVID-19 Weekly Epidemiological Update*, 54th ed.; WHO: Geneva, Switzerland, 2021.
3. Abdulkareem, K.H.; Mohammed, M.A.; Salim, A.; Arif, M.; Geman, O.; Gupta, D.; Khanna, A. Realizing an effective COVID-19 diagnosis system based on machine learning and IOT in smart hospital environment. *IEEE Internet Things J.* **2021**, *8*, 15919–15928. [CrossRef]
4. Esbin, M.N.; Whitney, O.N.; Chong, S.; Maurer, A.; Darzacq, X.; Tjian, R. Overcoming the bottleneck to widespread testing: A rapid review of nucleic acid testing approaches for COVID-19 detection. *RNA* **2020**, *26*, 771–783. [CrossRef] [PubMed]
5. Delgado, E.J.; Cabezas, X.; Martin-Barreiro, C.; Leiva, V.; Rojas, F. An Equity-Based Optimization Model to Solve the Location Problem for Healthcare Centers Applied to Hospital Beds and COVID-19 Vaccination. *Mathematics* **2022**, *10*, 1825. [CrossRef]
6. Akram, T.; Attique, M.; Gul, S.; Shahzad, A.; Altaf, M.; Naqvi, S.S.R.; Damaševičius, R.; Maskeliūnas, R. A novel framework for rapid diagnosis of COVID-19 on computed tomography scans. *Pattern Anal. Appl.* **2021**, *24*, 951–964. [CrossRef] [PubMed]
7. Sahlol, A.T.; Yousri, D.; Ewees, A.A.; Al-Qaness, M.A.; Damasevicius, R.; Abd Elaziz, M. COVID-19 image classification using deep features and fractional-order marine predators algorithm. *Sci. Rep.* **2020**, *10*, 15364. [CrossRef] [PubMed]
8. Benson, D.A.; Cavanaugh, M.; Clark, K.; Karsch-Mizrachi, I.; Lipman, D.J.; Ostell, J.; Sayers, E.W. GenBank. *Nucleic Acids Res.* **2012**, *41*, D36–D42. [CrossRef] [PubMed]
9. Arslan, H. Machine Learning Methods for COVID-19 Prediction Using Human Genomic Data. *Proceedings* **2021**, *74*, 20. [CrossRef]
10. Arslan, H.; Arslan, H. A new COVID-19 detection method from human genome sequences using CpG island features and KNN classifier. *Eng. Sci. Technol. Int. J.* **2021**, *24*, 839–847. [CrossRef]
11. Cortés-Carvajal, P.D.; Cubilla-Montilla, M.; González-Cortés, D.R. Estimation of the instantaneous reproduction number and its confidence interval for modeling the COVID-19 pandemic. *Mathematics* **2022**, *10*, 287. [CrossRef]
12. Sharma, N.; Krishnan, P.; Kumar, R.; Ramoji, S.; Chetupalli, S.R.; Ghosh, P.K.; Ganapathy, S. Coswara—A Database of Breathing, Cough, and Voice Sounds for COVID-19 Diagnosis. *arXiv* **2020**, arXiv:2005.10548.
13. Asraf, A.; Islam, M.Z.; Haque, M.R.; Islam, M.M. Deep learning applications to combat novel coronavirus (COVID-19) pandemic. *SN Comput. Sci.* **2020**, *1*, 363. [CrossRef] [PubMed]
14. He, S.; Gao, B.; Sabnis, R.; Sun, Q. Nucleic Transformer: Deep Learning on Nucleic Acids with Self-Attention and Convolutions. *bioRxiv* **2021**. [CrossRef]
15. Dasari, C.M.; Bhukya, R. Explainable deep neural networks for novel viral genome prediction. *Appl. Intell.* **2022**, *52*, 3002–3017. [CrossRef] [PubMed]
16. Wang, S.; Kang, B.; Ma, J.; Zeng, X.; Xiao, M.; Guo, J.; Cai, M.; Yang, J.; Li, Y.; Meng, X.; et al. A deep learning algorithm using CT images to screen for Corona Virus Disease (COVID-19). *Eur. Radiol.* **2021**, *31*, 6096–6104. [CrossRef]
17. Barstugan, M.; Ozkaya, U.; Ozturk, S. Coronavirus (COVID-19) classification using ct images by machine learning methods. *arXiv* **2020**, arXiv:2003.09424.
18. Gozes, O.; Frid-Adar, M.; Greenspan, H.; Browning, P.D.; Zhang, H.; Ji, W.; Bernheim, A.; Siegel, E. Rapid ai development cycle for the coronavirus (COVID-19) pandemic: Initial results for automated detection & patient monitoring using deep learning ct image analysis. *arXiv* **2020**, arXiv:2003.05037.
19. Özkaya, U.; Öztürk, Ş.; Barstugan, M. Coronavirus (COVID-19) classification using deep features fusion and ranking technique. In *Big Data Analytics and Artificial Intelligence against COVID-19: Innovation Vision and Approach*; Springer: Berlin/Heidelberg, Germany, 2020; pp. 281–295.

20. Muhammad, L.; Algehyne, E.A.; Usman, S.S.; Ahmad, A.; Chakraborty, C.; Mohammed, I.A. Supervised machine learning models for prediction of COVID-19 infection using epidemiology dataset. *SN Comput. Sci.* **2021**, *2*, 11. [CrossRef]
21. Narin, A.; Kaya, C.; Pamuk, Z. Automatic detection of coronavirus disease (COVID-19) using X-ray images and deep convolutional neural networks. *Pattern Anal. Appl.* **2021**, *24*, 1207–1220. [CrossRef]
22. Song, Y.; Zheng, S.; Li, L.; Zhang, X.; Zhang, X.; Huang, Z.; Chen, J.; Wang, R.; Zhao, H.; Zha, Y.; et al. Deep learning enables accurate diagnosis of novel coronavirus (COVID-19) with CT images. *IEEE/ACM Trans. Comput. Biol. Bioinform.* **2021**, *18*, 2775–2780. [CrossRef]
23. Pan, X.; Rijnbeek, P.; Yan, J.; Shen, H.-B. Prediction of RNA-protein sequence and structure binding preferences using deep convolutional and recurrent neural networks. *BMC Genom.* **2018**, *19*, 511. [CrossRef]
24. Ullah, A.; Muhammad, K.; Hussain, T.; Baik, S.W.; De Albuquerque, V.H.C. Event-oriented 3D convolutional features selection and hash codes generation using PCA for video retrieval. *IEEE Access* **2020**, *8*, 196529–196540. [CrossRef]
25. Muhammad, K.; Ullah, A.; Imran, A.S.; Sajjad, M.; Kiran, M.S.; Sannino, G.; de Albuquerque, V.H.C. Human action recognition using attention based LSTM network with dilated CNN features. *Future Gener. Comput. Syst.* **2021**, *125*, 820–830. [CrossRef]
26. Song, Y.; Fu, Q.; Wang, Y.-F.; Wang, X. Divergence-based cross entropy and uncertainty measures of Atanassov's intuitionistic fuzzy sets with their application in decision making. *Appl. Soft Comput.* **2019**, *84*, 105703. [CrossRef]
27. Killoran, N.; Lee, L.J.; Delong, A.; Duvenaud, D.; Frey, B.J. Generating and designing DNA with deep generative models. *arXiv* **2017**, arXiv:1712.06148.
28. Rangasamy, M.; Chesneau, C.; Martin-Barreiro, C.; Leiva, V. On a Novel Dynamics of SEIR Epidemic Models with a Potential Application to COVID-19. *Symmetry* **2022**, *14*, 1436. [CrossRef]
29. Pu, Y.; Gan, Z.; Henao, R.; Yuan, X.; Li, C.; Stevens, A.; Carin, L. Variational autoencoder for deep learning of images, labels and captions. *Adv. Neural Inf. Process. Syst.* **2016**, *29*, 2352–2360.
30. Ullah, W.; Muhammad, K.; Ul Haq, I.; Ullah, A.; Ullah Khattak, S.; Sajjad, M. Splicing sites prediction of human genome using machine learning techniques. *Multimed. Tools Appl.* **2021**, *80*, 30439–30460. [CrossRef]
31. Hassanzadeh, H.R.; Wang, M.D. DeeperBind: Enhancing prediction of sequence specificities of DNA binding proteins. In Proceedings of the 2016 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), Shenzhen, China, 15–18 December 2016; pp. 178–183.
32. Khan, S.U.; Baik, R. MPPIF-Net: Identification of Plasmodium Falciparum Parasite Mitochondrial Proteins Using Deep Features with Multilayer Bi-directional LSTM. *Processes* **2020**, *8*, 725. [CrossRef]
33. Cabanettes, F.; Klopp, C. D-GENIES: Dot plot large genomes in an interactive, efficient and simple way. *PeerJ* **2018**, *6*, e4958. [CrossRef]
34. Likic, V. *The Needleman-Wunsch Algorithm for Sequence Alignment*; Lecture given at the 7th Melbourne Bioinformatics Course; Bi021 Molecular Science and Biotechnology Institute, University of Melbourne: Melbourne, Australia, 2008; pp. 1–46.