A Machine Learning Approach for Analyzing Match Percentage of COVID-19 Nucleotide Sequence Alignment with Related Diseases

Arpita Gupta, Anirudh Krishnan, Archana Ganesh, Sriram Venkatesan, Ramadoss Balakrishnan

Abstract— In the current scenario, there is a pandemic around the world caused by a type of novel coronavirus, also known as COVID-19. Because of this reason, there is a need to understand the nucleotide sequence of the COVID-19. In our study, we have experimented on machine learning based models trained with transfer learning. The aim is to calculate the similarity of sequences of COVID-19 concerning eight diseases (Ebola, H5N1, H1N1, HIV, Alphacorona, MERS, Picornaviral, and Pneumonia). These diseases are either of the same family as COVID-19 or the diseases of which the medicines are being used for the treatment of patients. Nucleotide Sequence matching percentage is the process where a nucleotide sequence is compared with other sequences to find the similarity percentage in between these sequences. We have used transfer learning to overcome the issue of lack of COVID- 19 data. The model has highest achieved accuracy of 98.45. The top three diseases which are found to be most common are MERS, Alphacorona, and Ebola, respectively. The results have proved that the nucleotide sequence matching percentage could be utilized with great accuracy using machine learning. At the same time, it would help in better analyzing the COVID- 19 genomic data.

Index Terms— COVID- 19, Machine Learning, Transfer Learning, Genome

1 INTRODUCTION

In current time coronaviruses has caused a large-scale pandemic all over the world. Coronavirus has been creating pandemic from the last two decades in the form of SARS and the Middle East Respiratory System (MERS) [1,2,3]. A coronavirus outbreak happened in the city of Wuhan in China in December 2019. It was identified to be a novel coronavirus on 9th January 2020, now known as COVID-19 [4]. It is a form of coronavirus which is more dangerous than any virus outbreak that happened in the 21st century. The outbreak of this disease started from the seafood market in China. The symptoms associated with COVID-19 are fever, dry cough, headache, breathing difficulties, and Pneumonia [5]. The main cause of death because of COVID-19 is due to failure of respiratory systems leading to alveolar damage [5]. The cases of COVID-19 have either came in contact with the infected in Chinese seafood market or came in contact who have traveled back from China with infection leading to transmission all over the world. There is an exponential rise in the cases reported of coronavirus, and the current count has reached around 1.4 million cases all over the world. The country with maximum cases is the USA, followed by Italy and Spain. WHO has announced this as a pandemic all over the world as it is transmitting faster because of its person-to-person spread and no immunity to COVID-19 in humans.

The pathogens of COVID-19 are the most dangerous ones because of their rate of reproduction (2-3) and serial interval of (5-7.5) [4]. COVID-19's RNA virus family of single strands and which is commonly found in animals [6]. COVID-19 can incite SARS leading to respiratory failure. COVID- 19 is more severe than the SARS because it could be spread more easily [4]. COVID-19 has currently caused 5462 cases as of 8th April 2020 in India as it could go to the worst situation in India because of its dense population and weak health care facilities. Most of the countries affected by COVID- 19 are of lockdown, leading to severe loss in the economy.

Machine learning tools could be used in analyzing the outbreak of COVID- 19 and its characteristics in genomics

- *A. Gupta is with the Department of Computer Applications, National Institute of Technology, Tiruchirappalli, India, 620015 Email : arpitagupta2993@gmail.com*
- *A. Krishnan is with the Department of Chemical Engineering, National Institute of Technology, Tiruchirappalli, India, 620015 Email : aniroodh1998@gmail.com*
- *A. Ganesh is with the Department of Instrumentation and Control Engineering, National Institute of Technology, Tiruchirappalli, India, 620015 Email : archana.2098@gmail.com*
- *S. Venkatesan is with the Department of Instrumentation and Control Engineering, National Institute of Technology, Tiruchirappalli, India, 620015 Email : sriramvenkat98@gmail.com*
- *R. Balakrishnan is with the Department of Computer Applications, National Institute of Technology, Tiruchirappalli, India, 620015 Email : brama@nitt.edu*

and other categories. Machine learning techniques could help in better understanding the genomics of the COVID-19 and how it is affecting the humans machine leaning using the feature-based techniques to analyze the data better. The basic need for machine learning techniques is large training data, which is not possible in the case of COVID- 19 because of the pandemic it has created and lack of understanding of this disease [7]. Machine learning techniques need proper training for all the possible cases while this deficiency could be overcome by using the technique of transfer learning. Transfer learning is the technique in which the models could be pretrained on any other large dataset available for training and then tested on the available dataset [8]. There are different types of transfer learning. Since the models of machine learning need a well-annotated dataset for training here, transfer learning comes into picture. Machine learning also requires a large set of data for training, this problem could be overcome using transfer learning [7]. The basic feature of machine learning is feature extraction, that plays a vital role in analyzing the data.

Here we have used the transfer learning feature of machine learning to train the model. In transfer learning the learning from one domain could be used in another domain. The domains of the training dataset for pre-training and the target dataset are important as more relation leads to better transferring of knowledge [8]. Transfer learning is the key to solving the problem of lack of dataset and in solving one problem with the knowledge of the other. Here we have analyzed the nucleotide dataset available of the disease COVID- 19 from NCBI website, containing data collected from different cities. We have compared the coronavirus nucleotide data [9] with eight other diseases nucleotide data, they are Ebola, H5N1, H1N1, HIV, Alphacorona, MERS, Picornaviral, and Pneumonia [10,11,12,13,14,15,16,17], using Convolution Neural Network(CNN), Multilayer Perceptron(MLP) and Long Short Term Memory (LSTM) networks. We have compared the disease of the same family virus or the disease whose medicines are being used for the treatment of the patients.

1.1 Motivation

This work is to better understand the nucleotide sequence of COVID- 19 and to compare it and find out the disease it is similar to(Ebola, H5N1, H1N1, HIV, Alphacorona, MERS, Picornaviral, and Pneumonia), helping in future for vaccine preparation or to find the cure. The comparative study has shown very promising results, and it has proved that machine learning techniques could help in analyzing the nucleotide data and give a better understanding of the current scenario.

1.2 Problem Statement

To design a model to find the matching patterns in the nucleotide dataset of the eight diseases (Ebola, H5N1, H1N1, HIV, Alphacorona, MERS, Picornaviral, and Pneumonia) and to understand the sequence of the COVID- 19 datasets. To provide nucleotide sequence matching and to provide the matching percentage. Understanding the different classification model's performance and applying to find useful results.

1.3 Organization

The following paper has been organized as follows: in section 2 explain the related work, existing applications of machine learning in genomic sequence. Section 3 describes the proposed model based on machine learning and transfer learning. Section 4 explains the results achieved and the datasets used, followed by section 5 conclusion.

2 Related Work

Machine learning has certain applications in the field of genomics. Machine learning has proved to be very helpful in most of the fields. Machine learning techniques like CNN extract the features of the dataset and analyses the local and the global features of the dataset [18]. There are many techniques in machine learning which could be applied in the field of genomics. We have used CNN, LSTM, and MLP as these models' characteristics of feature extraction helps in better nucleotide sequence matching. These models have achieved great accuracy. This section will explain the basic concepts of transfer learning and machine learning techniques on how it could be used and has been used in the field of genomics and other fields.

2.1 Machine Learning

Recently there are many machine learning techniques being used in the analysis of genomic, also known as nucleotide sequence. RNN is the networks that are great at analyzing the sequential data so they could be used for genomic analysis [18]. CNN is great at image analysis so that it could be used in genomic image analysis. CNN could be used to extract the features better and analyze these features for a better understanding of mutation and genomic applications [19]. CNN's are also used in the sequence specificity of protein binding. For denoising, the data autoencoders could be used, which could be of great help. In the field of genomics, researchers could use the advantages of different models to solve the problems in the medical field. An existing model has used LSTM based neural machine translation for a language translation problem by understanding protein sequence [20]. Another model used LSTM with CNN for the prediction of protein subcellular localization from protein sequence [21]. Some works have used the hybrid models, they have fed the CNN output to LSTM for better classification, or they have grouped CNN and RNN for better protein sequence analysis.

2.2 Transfer Learning

Transfer learning techniques are adapted from the way humans learn from their experience and solve a problem that they have not faced before with the help of knowledge acquired from other problems. Transfer learning helps the machine learning model to acquire

knowledge from other problems and use them in a more relevant problem. It has proved to be useful in other fields where there is a lack of data for training like computer vision [22], natural language processing [23]. There are models like SVM, CNN used in the analysis of genomic expression. There is a need in the field of genomics to apply advanced deep learning techniques and to understand the pattern better.

In this paper, we have used the transfer learning technique for training the models, and these models are CNN, MLP, and LSTM. The main aim of this study is to find the sequence matching percentage in the sequences of the COVID- 19 concerning the eight diseases dataset on which we have trained the models. The pre-training is done on the large datasets of the eight diseases of the networks (CNN, LSTM, and MLP), using the knowledge acquired in the form of features from the pre-training of the model. We have compared the results of different networks and have found that CNN works the best among the three. This methodology has proved to be effective in finding the sequence matching percentage in the nucleotide sequence and has given the state-of-the-art results.

There are many works, but there is no other work that has proposed sequence matching of COVID-19 genomic sequence with other diseases. This research could be used to understand better and find a solution to the current pandemic situation.

3 Proposed Work

The network architecture we have experimented and compared are CNN, MLP, and LSTM, which can find the match percentage of the nucleotide sequence of the COVID-19 dataset in comparison to the eight diseases(Ebola, H5N1, H1N1, HIV, Alphacorona, MERS, Picornaviral, and Pneumonia). The network aims to find the match percentage of the nucleotide and to attain better accuracy. Here we have experimented, trained, and tested three networks. We have found that CNN performs the best in all the terms. The networks take the dataset of different diseases for training and then use the acquired knowledge to find the matching pattern percentage. We have chosen these networks because they have performed well on other genomic applications.

3.1 Pre-processing

The nucleotide sequence of different diseases is of different length, which demanded the preprocessing of the data. So, we have split the original sequence in 300 length subsequences [24]. These different subsequences of length 300 are given their corresponding parent sequence label. The nucleotide sequence preprocessing has helped us in achieving better results and in better training of the models. The input features for the model are obtained after one-hot encoding of these sequences. We have tried different sequence preprocessing, but this technique has given us the best results in training accuracy. This preprocessing technique of the nucleotide sequence is done in all the diseases dataset.

3.2 MLP

MLP model consists of three layers: - the input layer, hidden layer, and output layer. Our model has a hidden layer consisting of 16 nodes and a fully connected SoftMax layer. The SoftMax layer consists of 8 nodes. The learning rate used is 0.005, which we have found to give the best accuracy. The model uses the ReLu activation function as it has outperformed all the other activation functions like sigmoid, tanh. The loss function we have used is binary cross-entropy, and the optimizer is Adam. We have used batch normalization in our models as it leads to better accuracy. As shown in figure 1.

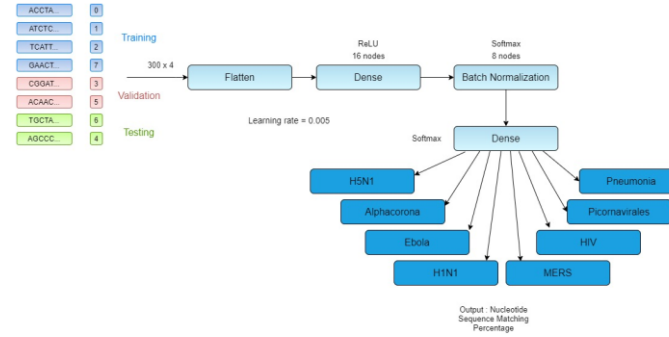

Fig.1 MLP Architecture

3.3 CNN

In our proposed CNN model, we have used two convolution 1D layers consisting of 32 filters, and the size of the kernel is 5. We have used l1 regularizer for the kernel with parameter values of 10^-5, we have found to be the optimum value by iterating through an array of parameter values. This is followed by a flattening layer, which is

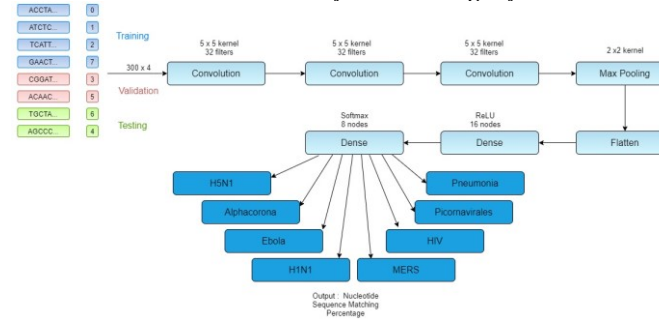

Fig.2 CNN Architecture

followed by a hidden layer consisting of 16 nodes and a fully connected SoftMax layer consisting of 8 nodes. Adding more hidden layers would have resulted in overfitting. The optimum learning rate we have found is 0.01. In this model, also, we have used the ReLu function as this was the optimum activation function for this model to perform the sequence match percentage calculation. As shown in figure 2 We have used the SGD optimizer and cross-entropy loss function. This model has outperformed all the other models and has achieved state-of-the-art ac-

curacy.

3.4 LSTM

In this model we have used bidirectional LSTM layer consisting of 15 nodes. The next layer is a dense layer with 16

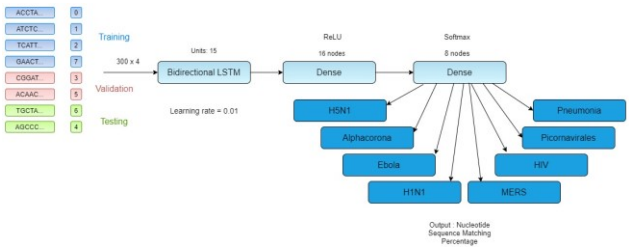

Fig.3 LSTM Architecture

nodes. We have used the ReLu activation function in this model also as it is the optimum one. As shown in figure 3. This model consists of the SoftMax layer consisting of 8 nodes, and the optimum learning rate we have found is 0.01. We have used regularization for better fitting and binary cross-entropy loss.

4 Results and Discussion

In this paper we have experimented three models for nucleotide sequence matching

Table 1. Details of the similarity percentage of the nucleotide sequence in comparison to other diseases

SIMILARITY PERCENTAGE WITH COVID-19			
DISEASE	**CNN**	**MLP**	**LSTM**
H5N1	0.435	0.069	0.081
ALPHACORONA	24.439	11.865	29.514
EBOLA	15.297	18.673	9.022
H1N1	1.811	0.691	0.928
MERS	**41.407**	**33.772**	**39.756**
HIV	0.147	1.672	1.404
PICORNAVIRALES	7.560	19.482	8.084
PNEUMONIA	8.900	13.771	11.156

and to find the matching percentage concerning the eight diseases, of the same family virus or the disease whose medicines are being used for the treatment of the patients. In this work, we have conducted experiments on the nucleotide data available at NCBI website for COVID- 19, and we have used the dataset available for the eight diseases (Ebola, H5N1, H1N1, HIV, Alphacorona, MERS, Picornaviral, and Pneumonia) on the NCBI website. These datasets are collected and made available from different countries. The COVID-19 has proved to be a pandemic and finding as much as about it on genomics level will help us in better understanding of the disease.

We have used Ebola, H5N1, H1N1, HIV, Alphacorona, MERS, Picornaviral, and Pneumonia nucleotide sequence dataset for training, and validation. COVID-19 dataset was tested using this model and the sequence match percentages of it with the 8 diseases were obtained. The following subsections describe the results achieved by these methods. As table 1 shows that the similarity percentage of the nucleotide is maximum in all the cases.

Table 2. Details of performance evaluation of the models

PARAMETERS	CNN	MLP	LSTM
ACCURACY	**98.45**	95.71	96.73
F1_SCORE	**93.73**	82.00	86.37
PRECISION	**94.71**	85.92	89.70
RECALL	**92.80**	78.54	83.38

4.1 MLP

This model of MLP is firstly trained on the eight-disease dataset, the pre-training. We have tested on the COVID-19 dataset and found the similarity percentage of the nucleotide sequence. The model is trained on 30 epochs. We have used a checkpoint feature and early stopping, monitoring the validation loss. The batch size used is 32. This model has achieved an accuracy of 95.71%, f1 score of 82.00%, a precision of 85.92%, and a recall of 78.54%.

The below figure 4 (a)shows graph of variation in precision value and 4 (b)shows variation in recall of the model in the MLP model while training and validation on the eight diseases dataset and COVID-19 dataset, respectively.

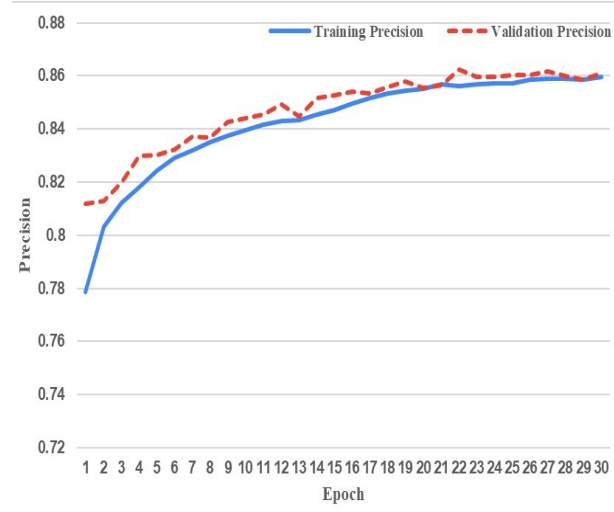

Fig. 4 (a) Graph showing variation in precision of the MLP model

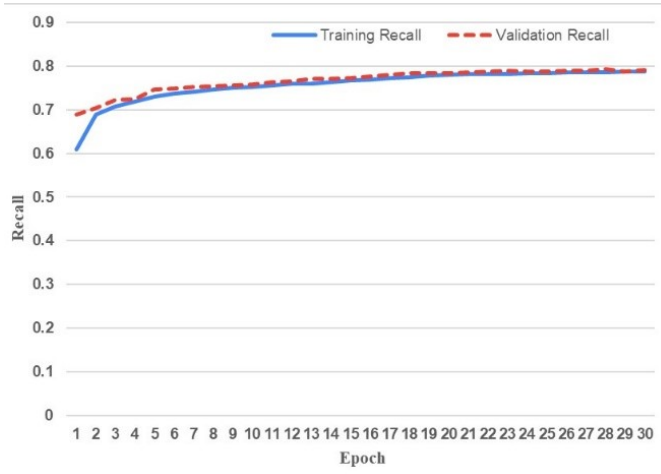

Fig. 4 (b) Graph showing variation in recall of the MLP model

The below figure 5 (a) shows the graph of the variation in the f1 score and 5 (b) shows variation in accuracy during training and validation of the model. The accuracy achieved in this model is 95.71, which is plausible. The model has shown that with non-deep models can also perform well in the field of genomics.

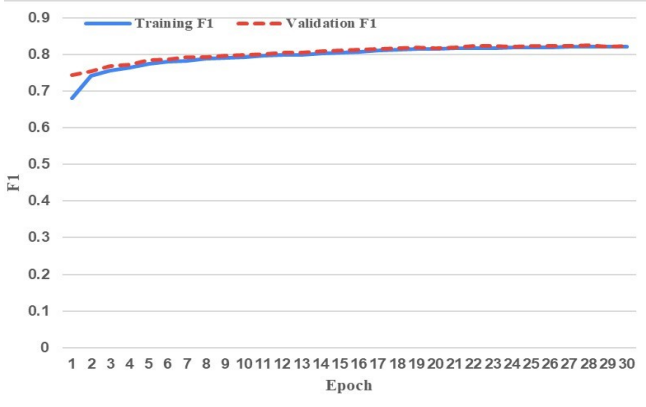

Fig. 5 (a) Graph showing variation in f1 score of the MLP model

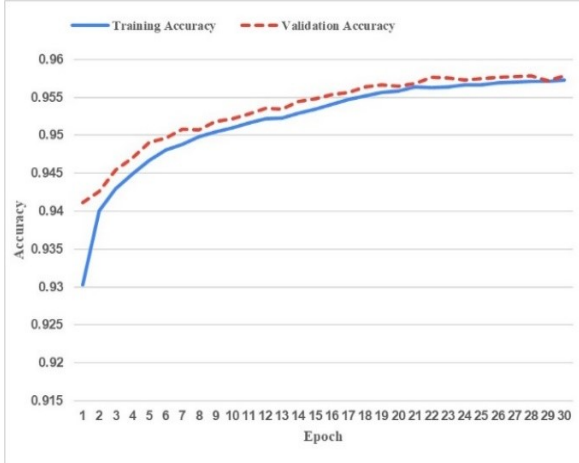

Fig. 5 (b) Graph showing variation in accuracy of the MLP model

The below figure 6 shows the variation in loss, where it can be concluded that the loss in the validation is lower than the training.

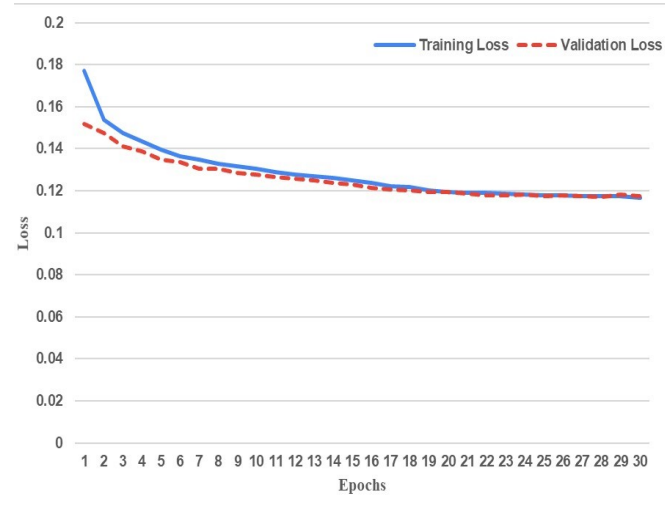

Fig. 6 Graph showing variation in loss of the MLP model

4.2 CNN

The proposed model of CNN is pre-trained on the eight-disease dataset and tested on COVID-19 dataset to find the similarity percentage of the sequence of the nucleotide. We have trained the model with 20 epochs. The batch size is 64. This CNN model has achieved an accuracy of 98.45%, f1 score of 93.73%, a precision of 94.71%, and a recall of 92.80%.

The figure 7 (a) shows the variation in the precision value and 7 (b) shows the variation of f1 score of the CNN model while training and validation on the eight diseases dataset and COVID-19 dataset, respectively.

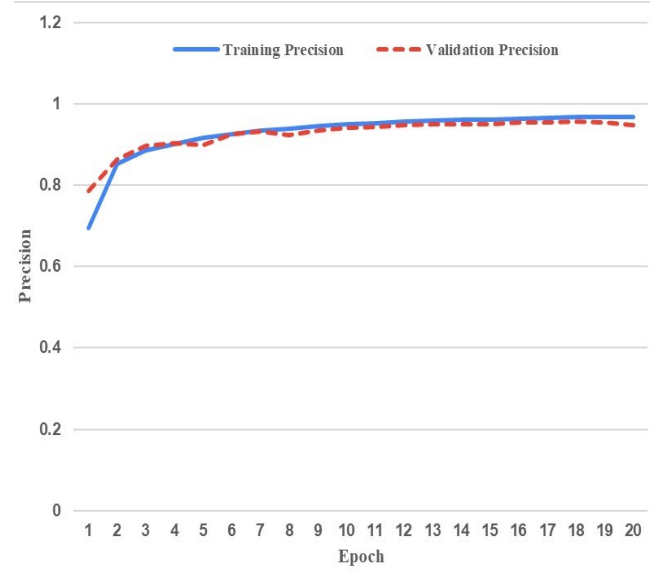

Fig. 7 (a)Graph showing variation in precision of the CNN model

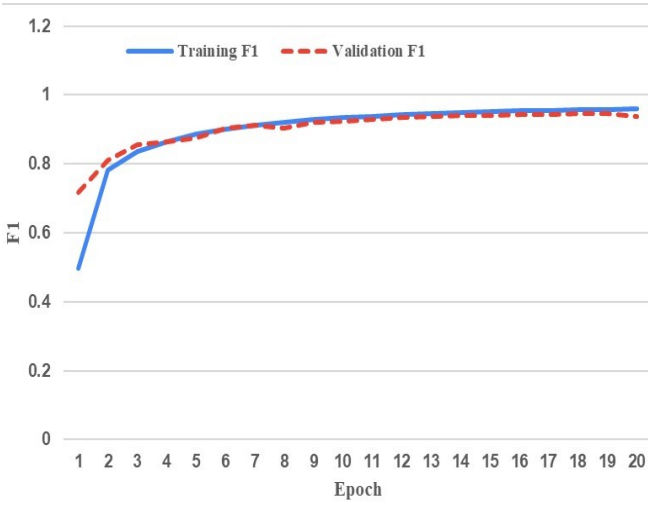

Fig. 7 (b) Graph showing variation in f1 score of the CNN model

The figure 8 (a) shows the variation in the values of recall and 8 (b) shows variation of accuracy in training and validation phase of the network.

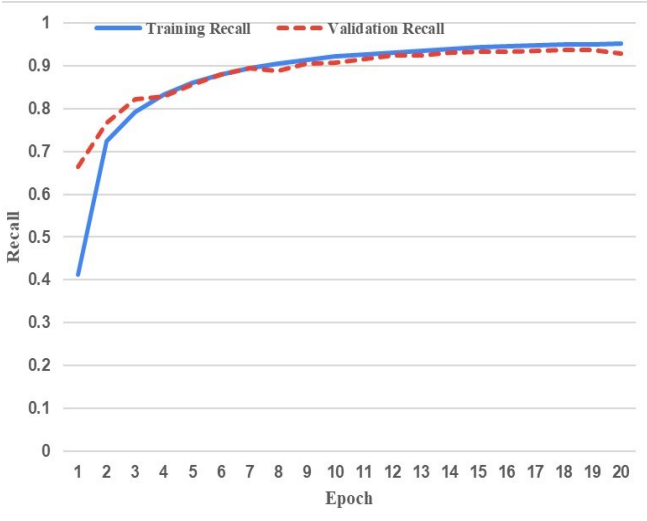

Fig. 8 (a) Graph showing variation in recall of the CNN model

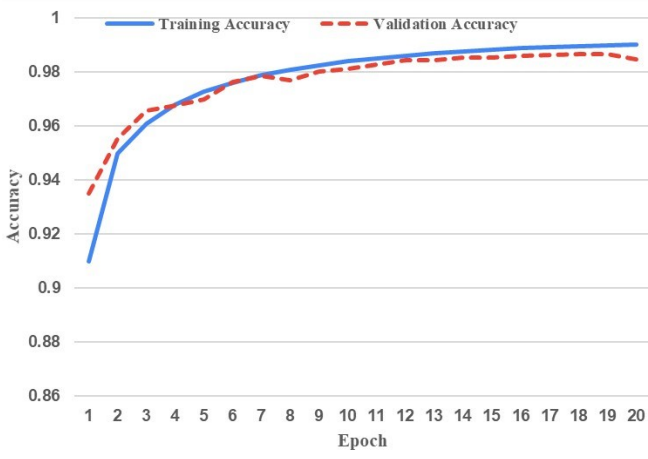

Fig. 8 (b) Graph showing variation accuracy score of the CNN model

The graph in figure 9 shows the loss variation loss in higher validation. This model has shown the best result in comparison to all the models and has given the state-of-the-art results. This CNN model has achieved an accuracy of 98.45%, f1 score of 93.73%, a precision of 94.71%, and a recall of 92.80%.

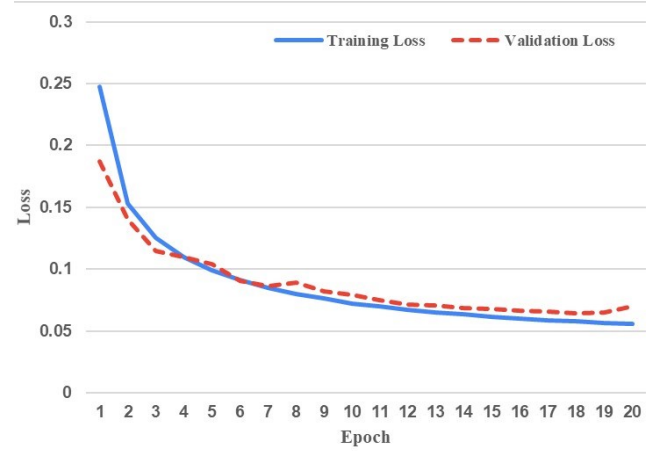

Fig. 9 Graph showing variation in loss in the CNN model

This model has achieved the best accuracy and has proved that a lack of data could be overcome with the transfer learning method.

4.3 LSTM

LSTM model is pretrained on the eight diseases dataset and tested on the COVID-19 dataset. The aim is to find the similarity percentage in the COVID-19 nucleotide sequence with regards to the eight diseases. This model is trained with 20 epochs and a batch size of 256.

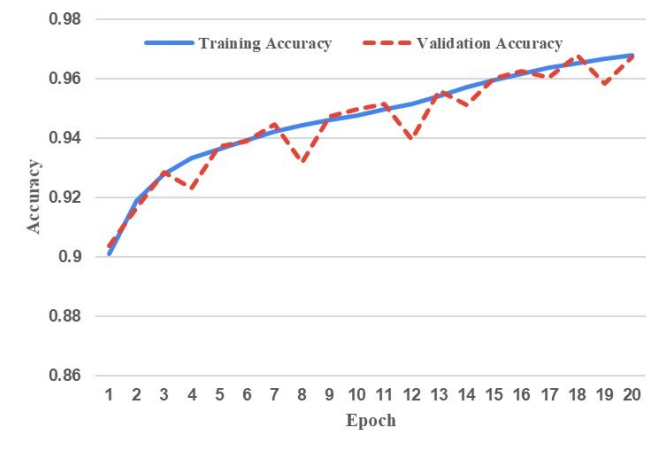

Fig. 10 (a) Graph showing variation in accuracy of the LSTM model

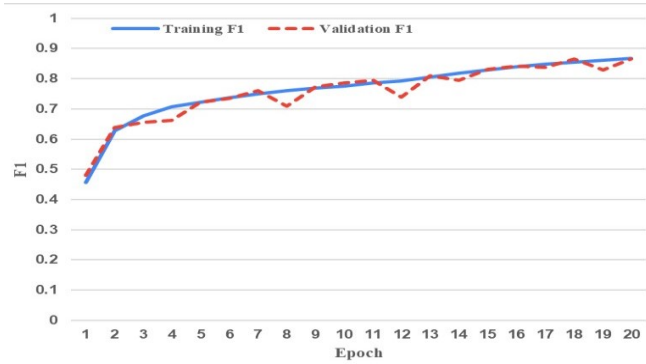

Fig. 10 (b) Graph showing variation in f1 score of LSTM model

The figure 10 (a) shows the graph of the variation in the f1 score and 10 (b) shows in variation in accuracy of the training and validation phase in the LSTM model. This model has outperformed the MLP model, and the model has achieved an accuracy of 96.73%, f1 score of 86.37%, precision of 89.70%, and recall of 83.38%.

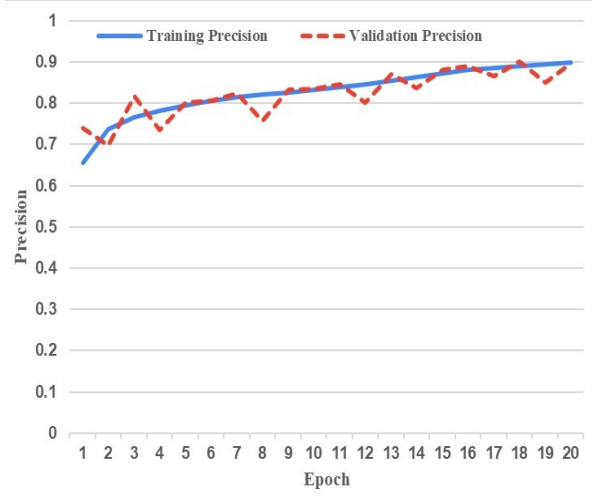

Fig. 11 (a) Graph showing variation in precision of the LSTM model

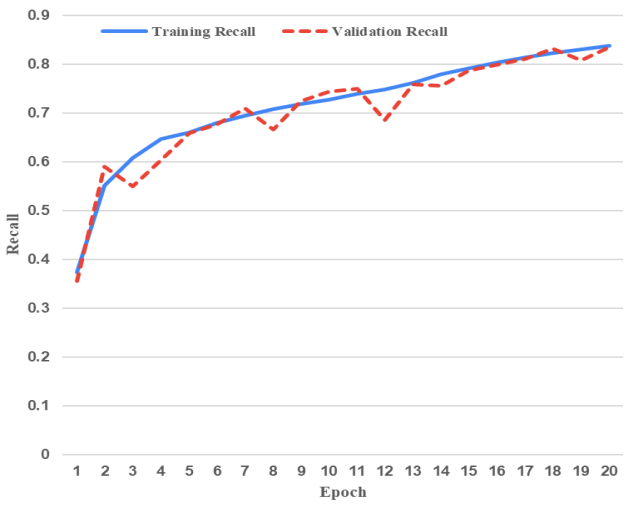

Fig. 11 (b) Graph showing variation in recall score of the LSTM model

The figure 11 (a) of the graph shows the variation in the LSTM model with the epochs in recall and 11 (b) shows the variation in precision.

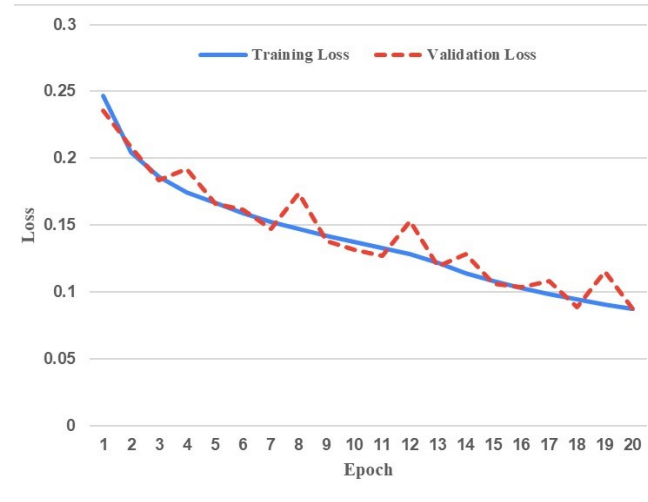

Fig. 12 Graph showing variation in loss in the LSTM model

The figure 12 shows the graph of loss variation in the training and validation. This model has performed well and has shown that LSTM could be used in the genomic sequence analysis.

5 Conclusion

In this paper we have experimented on transfer learning-based model whose main aim is to find the sequence matching percentage of COVID-19 corresponding to the eight diseases. We have proposed training the three machine learning models (CNN, LSTM, and MLP) with the transfer learning technique. We have compared with the eight diseases and COVID-19 dataset provided by NCBI websites from different countries. The optimum accuracy we have achieved is 98.45 by CNN, while the other networks have accuracy of 95.71, 96.73 (MLP and LSTM respectively). Top three diseases are found to be most common are MERS, Alphacorona, and Ebola, respectively with COVID-19 sequence. This work will help in solving the lack of understanding and data of genomic sequence of COVID-19 helping to find the solution to it.

References

[1] Drosten, C., Günther, S., Preiser, W., Van der Werf, S., Brodt, H. R., Becker, S., Rabenau, H., Panning, M., Kolesnikova, L., Fouchier, R. A. M., Berger, A., Burguière, A. M., Cinatl, J., Eickmann, M., Escriou, N., Grywna, K., Kramme, S., Manuguerra, J. C., Müller, S., … Doerr, H. W. (2003). Identification of a novel coronavirus in patients with severe acute respiratory syndrome. *New*

England Journal of Medicine, *348*(20), 1967–1976. https://doi.org/10.1056/NEJMoa030747

[2] Zaki, A. M., Van Boheemen, S., Bestebroer, T. M., Osterhaus, A. D. M. E., & Fouchier, R. A. M. (2012). Isolation of a novel coronavirus from a man with pneumonia in Saudi Arabia. *New England Journal of Medicine*, *367*(19), 1814–1820. https://doi.org/10.1056/NEJMoa1211721

[3] Cui, J., Li, F., & Shi, Z. L. (2019). Origin and evolution of pathogenic coronaviruses. *Nature Reviews Microbiology*, *17*(3), 181–192. https://doi.org/10.1038/s41579-018-0118-9

[4] Paper, M. D., Insights, T., & Paper, L. (2020). *COVID-19 Overview of information available to support the development of medical countermeasures and interventions against COVID-19*. 1–79.

[5] Zhou, P., Yang, X. Lou, Wang, X. G., Hu, B., Zhang, L., Zhang, W., Si, H. R., Zhu, Y., Li, B., Huang, C. L., Chen, H. D., Chen, J., Luo, Y., Guo, H., Jiang, R. Di, Liu, M. Q., Chen, Y., Shen, X. R., Wang, X., … Shi, Z. L. (2020). A pneumonia outbreak associated with a new coronavirus of probable bat origin. *Nature*, *579*(7798), 270–273. https://doi.org/10.1038/s41586-020-2012-7

[6] Li, X., Geng, M., Peng, Y., Meng, L., & Lu, S. (2020). Molecular immune pathogenesis and diagnosis of COVID-19. *Journal of Pharmaceutical Analysis*, *19*(xxxx), 1–7. https://doi.org/10.1016/j.jpha.2020.03.001

[7] Santosh, K. C. (2020). AI-Driven Tools for Coronavirus Outbreak: Need of Active Learning and Cross-Population Train/Test Models on Multitudinal/Multimodal Data. *Journal of Medical Systems*, *44*(5), 1–5. https://doi.org/10.1007/s10916-020-01562-1

[8] Xu, Q., & Yang, Q. (2011). A Survey of Transfer and Multitask Learning in Bioinformatics. *Journal of Computing Science and Engineering*, *5*(3), 257–268. https://doi.org/10.5626/jcse.2011.5.3.257

[9] *[dataset] NCBI Virus*. (n.d.). Retrieved April 10, 2020, from https://www.ncbi.nlm.nih.gov/labs/virus/vssi/#/virus?SeqType_s=Nucleotide&HostLineage_ss=Homo sapiens (human), taxid:9606&Completeness_s=comp

[10] *[dataset] NCBI Virus*. (n.d.). Retrieved April 10, 2020, from https://www.ncbi.nlm.nih.gov/labs/virus/vssi/#/virus?SeqType_s=Nucleotide&HostLineage_ss=Homo sapiens (human), taxid:9606&Flags_csv=complete&Completeness_s=complete&VirusLineage_ss=Ebolavirus, taxid:186536

[11] *[dataset] NCBI Virus*. (n.d.). Retrieved April 10, 2020, from https://www.ncbi.nlm.nih.gov/labs/virus/vssi/#/virus?SeqType_s=Nucleotide&HostLineage_ss=Homo sapiens (human), taxid:9606&Flags_csv=complete&Completeness_s=complete&VirusLineage_ss=H5N1 subtype, taxid:102793

[12] *[dataset] NCBI Virus*. (n.d.). Retrieved April 10, 2020, from https://www.ncbi.nlm.nih.gov/labs/virus/vssi/#/virus?SeqType_s=Nucleotide&HostLineage_ss=Homo sapiens (human), taxid:9606&Flags_csv=complete&Completeness_s=complete&VirusLineage_ss=H1N1 subtype, taxid:114727

[13] *[dataset] NCBI Virus*. (n.d.). Retrieved April 10, 2020, from https://www.ncbi.nlm.nih.gov/labs/virus/vssi/#/virus?SeqType_s=Nucleotide&HostLineage_ss=Homo sapiens (human), taxid:9606&Flags_csv=complete&Completeness_s=complete&VirusLineage_ss=Alphacoronavirus, taxid:693996

[14] *[dataset] NCBI Virus*. (n.d.). Retrieved April 10, 2020, from https://www.ncbi.nlm.nih.gov/labs/virus/vssi/#/virus?SeqType_s=Nucleotide&HostLineage_ss=Homo sapiens (human), taxid:9606&Completeness_s=complete&VirusLineage_ss=Human immunodeficiency virus 1 (HIV-1), taxid:11676

[15] *[dataset] NCBI Virus*. (n.d.). Retrieved April 10, 2020, from https://www.ncbi.nlm.nih.gov/labs/virus/vssi/#/virus?SeqType_s=Nucleotide&HostLineage_ss=Homo sapiens (human), taxid:9606&Completeness_s=complete&VirusLineage_ss=Human immunodeficiency virus 1 (HIV-1), taxid:11676

[16] *[dataset] NCBI Virus*. (n.d.). Retrieved April 10, 2020, from https://www.ncbi.nlm.nih.gov/labs/virus/vssi/#/virus?SeqType_s=Nucleotide&HostLineage_ss=Homo sapiens (human), taxid:9606&Flags_csv=complete&Completeness_s=complete&VirusLineage_ss=Picornavirales, taxid:464095

[17] *[dataset] NCBI Virus*. (n.d.). Retrieved April 10, 2020, from https://www.ncbi.nlm.nih.gov/labs/virus/vssi/#/virus?SeqType_s=Nucleotide&HostLineage_ss=Homo sapiens (human), taxid:9606&Completeness_s=complete&VirusLineage_ss=Pneumoviridae, taxid:11244

[18] Yue, T., & Wang, H. (2018). *Deep Learning for Genomics: A Concise Overview*. 1–40. http://arxiv.org/abs/1802.00810

[19] Zeng, H., Edwards, M. D., Liu, G., & Gifford, D. K. (2016). Convolutional neural network architectures for predicting DNA–protein binding. *Bioinformatics*, *32*(12), i121–i127. https://doi.org/10.1093/bioinformatics/btw255

[20] Cao, R., Freitas, C., Chan, L., Sun, M., Jiang, H., & Chen, Z. (2017). ProLanGO: Protein function

prediction using neural machine translation based on a recurrent neural network. *Molecules*, *22*(10). https://doi.org/10.3390/molecules22101732

[21] Sønderby, S. K., Sønderby, C. K., Nielsen, H., & Winther, O. (2015). Convolutional LSTM networks for subcellular localization of proteins. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *9199*, 68–80. https://doi.org/10.1007/978-3-319-21233-3_6

[22] Moon, S., Kim, S., & Wang, H. (n.d.). *Multimodal Transfer Deep Learning with Applications in Audio-Visual Recognition*.

[23] Ciresan, D. C., Meier, U., & Schmidhuber, J. (2012). Transfer learning for Latin and Chinese characters with deep neural networks. *Proceedings of the International Joint Conference on Neural Networks*, 10–15. https://doi.org/10.1109/IJCNN.2012.6252544

[24] Tampuu, A., Bzhalava, Z., Dillner, J., & Vicente, R. (2019). ViraMiner: Deep learning on raw DNA sequences for identifying viral genomes in human samples. *PLoS ONE*, *14*(9). https://doi.org/10.1371/journal.pone.0222

Arpita Gupta is pursuing Doctor of Philosophy from Department of Computer Applications from National Institute of Technology, Tiruchirappalli. She has received Master of Technology in Computer Science and Engineering from Indian Institute of Information Technology, Srirangam in 2017. She has received her Bachelor of Engineering from Technocrats Institute of Technology (Excellence), Bhopal in 2014. Her area of research is transfer learning, computer vision, etc.

Anirudh Krishnan is pursuing B-Tech with major in Chemical Engineering and minor in Computer Science and Engineering at National Institute of Technology, Tiruchirappalli.

Archana Ganesh is pursuing B-Tech with major in Instrumentation and Control Engineering and minor in Computer Science and Engineering at National Institute of Technology, Tiruchirappalli

Sriram Venkatesan is pursuing B-Tech with major in Instrumentation and Control Engineering and minor in Computer Science and Engineering at National Institute of Technology, Tiruchirappalli

Dr. (Mr.) RAMADOSS Balakrishnan received the M.Tech degree in Computer science and Engineering from the Indian Institute of Technology, Delhi in 1995 and earned Ph.D. degree in Applied Mathematics from Indian Institute of Technology, Bombay in 1983. Currently, he is working as a Professor of Computer Applications at National Institute of Technology, Tiruchirappalli. His research interests include: Software Testing Methodologies, Security and Privacy in Big Data and Cloud, Data Mining, Predictive Analytics in Big Data & Multimedia Mining.