2021 43rd Annual International Conference of the
IEEE Engineering in Medicine & Biology Society (EMBC)
Oct 31 - Nov 4, 2021. Virtual Conference

# Classification of Influenza Hemagglutinin Protein Sequences using Convolutional Neural Networks

Charalambos Chrysostomou[1*], Floris Alexandrou[1], Mihalis A. Nicolaou[1] and Huseyin Seker[2]

*Abstract*— **The Influenza virus can be considered as one of the most severe viruses that can infect multiple species with often fatal consequences to the hosts. The Hemagglutinin (HA) gene of the virus can be a target for antiviral drug development realised through accurate identification of its sub-types and possible the targeted hosts. This paper focuses on accurately predicting if an Influenza type A virus can infect specific hosts, and more specifically, Human, Avian and Swine hosts, using only the protein sequence of the HA gene. In more detail, we propose encoding the protein sequences into numerical signals using the Hydrophobicity Index and subsequently utilising a Convolutional Neural Network-based predictive model. The Influenza HA protein sequences used in the proposed work are obtained from the Influenza Research Database (IRD). Specifically, complete and unique HA protein sequences were used for avian, human and swine hosts. The data obtained for this work was 17999 human-host proteins, 17667 avian-host proteins and 9278 swine-host proteins. Given this set of collected proteins, the proposed method yields as much as 10% higher accuracy for an individual class (namely, Avian) and 5% higher overall accuracy than in an earlier study. It is also observed that the accuracy for each class in this work is more balanced than what was presented in this earlier study. As the results show, the proposed model can distinguish HA protein sequences with high accuracy whenever the virus under investigation can infect Human, Avian or Swine hosts.**

## I. INTRODUCTION

Pathogens, including viruses, can cause infectious diseases to spread within populations. These pathogens can be transmitted in multiple ways, with high transmission rates in most circumstances. Identification and diagnosis of infectious diseases are vital, especially in novel pathogens, to prevent and control global pandemics, such as the current COVID-19 pandemic [1].

Influenza viruses are part of the Orthomyxoviridae family of viruses that have negative-sense, single-stranded, segmented RNA genomes, with the majority of the virus burden being associated with influenza viruses type A and B [2]. Influenza viruses capable of infecting humans were introduced from birds and swine [3]. Their introduction to humans has begun global pandemics with the 1918 "Spanish flu" and 2009 "Swine flu" pandemics. Influenza viruses are responsible for more than 500,000 deaths worldwide and affect around 5–15% of the population each year [4]. The evolution of influenza viruses enables them to infect individuals who have previously gained immunity through vaccination or previous infections. Furthermore, Influenza viruses can be effectively transmitted within populations from direct contact, respiratory droplets, objects, or materials in the environment, such as clothes, utensils, and furniture. Currently, vaccines are the most efficient and practical method in limiting the influenza virus's spread and stop Influenza epidemics. These vaccines must be renewed periodically to maintain their effectiveness against Influenza viruses.

The Influenza virus genome contains eight gene segments that include: PB2 (polymerase subunit) responsible for encoding RNA; PB1 (polymerase subunit) responsible for encoding RNA and includes the PB1-F2 protein, which causes cell death; PA (polymerase subunit) responsible for encoding RNA and includes the PA-X protein, responsible for host transcription shutoff function; NP (nucleoprotein); M1 and M2 (matrix proteins); NS1 and NEP (non-structural proteins); NA (neuraminidase) which facilitates the release of new viruses from the infected host cell; HA (hemagglutinin) responsible for binding and entry of the virus to the host cell. As this study's primary aim is to classify the capability of an Influenza virus to infect different hosts, we will focus primarily on the HA gene. Influenza viruses are classified into subtypes based on the organisation of HA and NA glycoproteins on their surfaces. Currently, 18 HA subtypes and 11 NA subtypes exist in the wild [5] with the majority affecting Avian species. Only three subtypes are known to adapt and circulate in Humans H1N1, H2N2 and H3N2. Seasonal epidemics are caused by two of these Influenza subtypes H1N1 and H3N2.

In the literature, various methods were developed and used for analysing and characterising protein sequences. More specifically, for the classification and characterisation of Influenza subtypes, [6], [7], [8], [9], [10], [11], methods such as Complex resonant recognition model in analysing influenza subtype protein sequences [7], CISAPS: Complex informational spectrum for the analysis of protein sequences [10], and Structural Classification of protein sequences based on signal processing and support vector machines [11], have been proposed and developed.

In this paper, a new method is proposed based on deep learning methodologies and, more specifically, Convolutional Neural Networks (CNN) to classify the Influenza protein sequences based on their capability to infect the specific host by utilising only the amino acid sequence of the protein sequence to accomplish unprecedented accuracy. The paper is organised as follows: Section II presents the methods and

[1]Charalambos Chrysostomou, Floris Alexandrou and Mihalis Nicolaou are with Computation-based Science and Technology Research Center, The Cyprus Institute, Nicosia, Cyprus c.chrysostomou@cyi.ac.cy, f.alexandrou@cyi.ac.cy, m.nicolaou@cyi.ac.cy

[2]Huseyin Seker is with the Faculty of Computing, Engineering and the Built Environment, Birmingham City University, Birmingham, B5 5JU The United Kingdom hseker1970@yahoo.co.nz

* Corresponding Author

TABLE I
NUMBER OF HA PROTEIN SEQUENCES

| HA Subtype | Human | Avian | Swine |
|---|---|---|---|
| H1* | 8426 | 594 | 6626 |
| H2* | 81 | 416 | 2 |
| H3* | 9134 | 1316 | 2429 |
| H4* | 0 | 1001 | 5 |
| H5* | 220 | 3653 | 24 |
| H6* | 0 | 1485 | 2 |
| H7* | 100 | 1310 | 3 |
| H8* | 0 | 149 | 0 |
| H9* | 14 | 5009 | 20 |
| H10* | 2 | 621 | 1 |
| H11* | 0 | 654 | 1 |
| H12* | 0 | 290 | 0 |
| H13* | 0 | 276 | 0 |
| H14* | 0 | 19 | 0 |
| H15* | 0 | 10 | 0 |
| H16* | 0 | 140 | 0 |
| mixed | 22 | 724 | 165 |
| **Total** | **17999** | **17667** | **9278** |

TABLE II
HYDROPHOBICITY AMINO ACID INDEX USED TO ENCODE PROTEIN SEQUENCES FROM ALPHABETICAL TO NUMERICAL CHARACTERS FOR ANALYSIS

| Amino acid | Hydrophobicity Values | |
|---|---|---|
| | Original | Normalised |
| Alanine (A) | 0.61 | 0.23 |
| Arginine (R) | 0.60 | 0.23 |
| Asparagine (N) | 0.06 | 0.02 |
| Aspartic acid (D) | 0.46 | 0.17 |
| Cysteine (C) | 1.07 | 0.40 |
| Glutamine (Q) | 0.00 | 0.00 |
| Glutamic acid (E) | 0.47 | 0.18 |
| Glycine (G) | 0.07 | 0.03 |
| Histidine (H) | 0.61 | 0.23 |
| Isoleucine (I) | 2.22 | 0.84 |
| Leucine (L) | 1.53 | 0.58 |
| Lysine (K) | 1.15 | 0.43 |
| Methionine (M) | 1.18 | 0.45 |
| Phenylalanine (F) | 2.02 | 0.76 |
| Proline (P) | 1.95 | 0.74 |
| Serine (S) | 0.05 | 0.02 |
| Threonine (T) | 0.05 | 0.02 |
| Tryptophan (W) | 2.65 | 1.00 |
| Tyrosine (Y) | 1.88 | 0.71 |
| Valine (V) | 1.32 | 0.50 |
| Other | 0.00 | 0.00 |

materials developed and used, while Section III presents the results obtained. Finally, concluding remarks are outlined in Section IV.

## II. METHODS AND MATERIALS

### A. Protein Sequences

The Influenza HA protein sequences used in the proposed work are obtained from The Influenza Research Database (IRD) [12]. Specifically, complete and unique Hemagglutinin (HA) protein sequences were used for avian, human and swine hosts, which are the primary hosts affected with the virus. The data obtained for this work was 17999 human-host proteins, 17667 avian-host proteins and 9278 swine-host proteins. The complete list of proteins, including HA subclasses, used in this study can be found in Table I, and Figure 1 illustrates the percentage of HA proteins per class and species.

The proposed analysis was performed directly to the protein sequence using a Hydrophobicity index [13] to encode the protein sequences from alphabetical to numerical characters for analysis. Before encoding the protein sequences, the amino acid index values were re-normalised to 0-1. Any character beyond the standard 20 amino acids was encoded using the value 0. The complete list of the Hydrophobicity amino acid index can be found in Table II.

The collected protein sequences have variable sizes, with 576 being the maximum number of amino acids. To encode the sequences and use the proposed model, proteins with a lower number of amino acids were padded with 0's at the end of the sequence to reach the mentioned number. The data was further encoded with one-hot encoding for the multi-class classification problem.

### B. Proposed Model

The proposed work is based on a Convolutional Neural Network (CNN). CNN's are a subtype of deep feed-forward artificial neural networks that have been applied and used to analyse visual representations but recently have been used in other domains, including classification and characterisation of protein sequences [14], [15], [16]. CNN's utilise a variation of multi-layer perceptrons created to minimise the pre-processing required [17], thus in comparison, CNN's require relatively minimal pre-processing of data in relation to other classification methodologies and translates to a substantial advantage where prior knowledge and expertise of any given problem is not available or unknown.

As shown in Figure 2, the proposed model structure consists of three convolutional layers of 32, 3x3 kernels, followed by a max-pooling layer and a rectified nonlinear activation function (Leaky ReLU) [18] transforming the feature space from 32x576 to 32x72. The fourth block consists of a Flatten layer that transforms the feature space

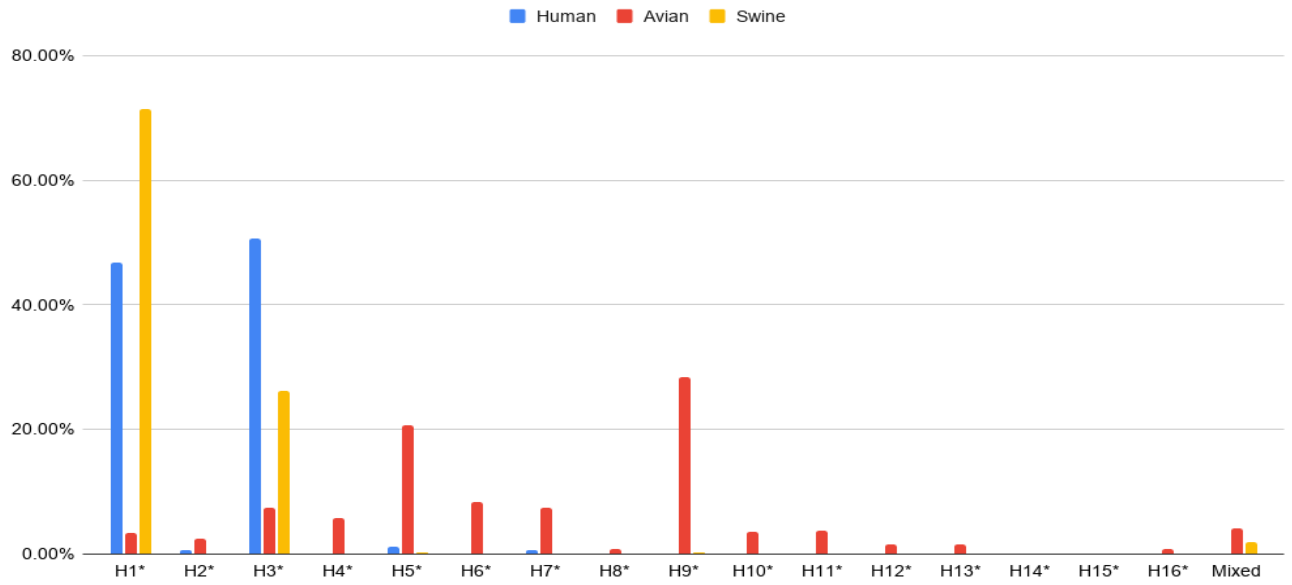


Fig. 1. Percentage of HA proteins per class and species

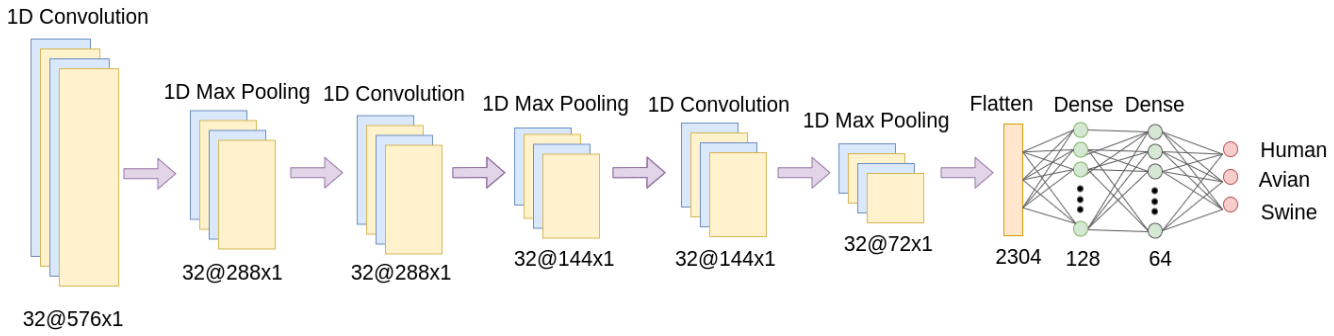


Fig. 2. Structure of the Proposed Model based on Convolutional Neural Networks

from 32x72 to 2304x1. The fifth and sixth layers are fully connected layers of 128 and 64 neurons, respectively, with Leaky ReLU as the activation function. The output layer consists of 3 neurons that correspond to each of the available species and utilises the softmax activation function [17]. The model is trained using the "Adam" optimiser [19] and the "categorical cross-entropy" loss function [17]. The proposed model and hyperparameters were selected based on a trial and error process to maximise the training and validation accuracy. The proposed methodology source code is available at `https://gitlab.com/charalambos.chrysostomou/embc21_influenza.git`

## III. Results and Discussion

In this paper, a classification model is presented, based on Deep Learning and Convolutional Neural Networks (CNN), for the characterisations and classification Influenza type A based upon the ability to infect a specific host, more specifically human, avian and swine hosts, by solely using the HA protein sequence.

To ensure that the proposed model is accurate and the results can be generalised, the model was trained 100 times for 1000 epochs, with different random subsets chosen uniformly and assigned to training, validation and testing sets. For each training cycle, the best weights were saved based on the validation error. To evaluate the proposed method's efficacy, the average of the test set, for the total accuracy, Matthews correlation coefficient (MCC) [20] and F-score (F1) [21] scores, were considered. The predictive model's average accuracy is 99.36% $\pm$ 0.11%, 98.84% $\pm$ 0.20% and 98.78% $\pm$ 0.22% for the training, validation and test sets, respectively. For the MCC score, the results are 0.990 $\pm$ 0.002, 0.982 $\pm$ 0.003 and 0.981 $\pm$ 0.003 for the training, validation and test sets, respectively. Finally, for the F1 score, the results are 0.994 $\pm$ 0.001, 0.988 $\pm$ 0.002 and 0.988 $\pm$ 0.002 for the training, validation and test sets, respectively. Detailed results can be found in Tables III and IV. The results based on the test sets per species were also calculated with 98.74% $\pm$ 0.32%, 99.52% $\pm$ 0.25% and 97.19% $\pm$ 0.75% for Human, Avian and Swine species, respectively. The proposed

TABLE III
RESULTS

| Set | Accuracy | MCC Score | F1 Score |
|---|---|---|---|
| Training | 99.36% ± 0.11% | 0.990 ± 0.002 | 0.994 ± 0.001 |
| Validation | 98.84% ± 0.20% | 0.982 ± 0.003 | 0.988 ± 0.002 |
| Test | 98.78% ± 0.22% | 0.981 ± 0.003 | 0.988 ± 0.002 |

TABLE IV
RESULTS BASED ON THE TEST SETS PER SPECIES

| Set | Accuracy |
|---|---|
| Human | 98.74% ± 0.32% |
| Avian | 99.52% ± 0.25% |
| Swine | 97.19% ± 0.75% |

method yields almost 10%, 5% and 2% higher accuracy for Avian, Human and Swine, respectively, than those of the earlier study [22]. It is also observed that the accuracy for each class presented in our work is more balanced than what was presented in this earlier study. The final overall accuracy is also found to be as much as 5% higher than that of the earlier study.

## IV. CONCLUSIONS

The paper presents a highly successful predictive model to classify Influenza viruses based on their capability to infect different hosts, including Human, Avian and Swine Hosts, as the results show compared to existing methods. For this study, the protein sequence of the HA gene was used, which is responsible for attaching to the host's cell and can be considered a promising antiviral drug candidate. The collected protein sequences were encoded using a normalised hydrophobicity amino acid index.

As published in the literature, more than 600 unique amino acid indices exist that describes a physicochemical feature of the protein [10]. Future studies are needed to identify potential alternative amino acid index or set of indices capable of better representing HA proteins and delivering even more reliable results. As the recent events of the COVID-19 pandemic have shown, a computational tool capable of identifying novel and potentially dangerous viruses in the wild that have acquired the capability to infect Human hosts will be crucial and needed to predict and control future outbreaks.

## REFERENCES


[1] M. Nicola, Z. Alsafi, C. Sohrabi, A. Kerwan, A. Al-Jabir, C. Iosifidis, M. Agha, and R. Agha, "The socio-economic implications of the coronavirus and covid-19 pandemic: a review," *International journal of surgery*, 2020.

[2] M. C. Zambon, "Epidemiology and pathogenesis of influenza," *Journal of Antimicrobial Chemotherapy*, vol. 44, no. suppl_2, pp. 3–9, 1999.

[3] C. A. Russell, P. M. Kasson, R. O. Donis, S. Riley, J. Dunbar, A. Rambaut, J. Asher, S. Burke, C. T. Davis, R. J. Garten, *et al.*, "Science forum: improving pandemic influenza risk assessment," *Elife*, vol. 3, p. e03883, 2014.

[4] K. Stöhr, "Influenza—who cares," *The Lancet infectious diseases*, vol. 2, no. 9, p. 517, 2002.

[5] S. Tong, X. Zhu, Y. Li, M. Shi, J. Zhang, M. Bourgeois, H. Yang, X. Chen, S. Recuenco, J. Gomez, *et al.*, "New world bats harbor diverse influenza a viruses," *PLoS pathog*, vol. 9, no. 10, p. e1003657, 2013.

[6] C. Chrysostomou, H. Partaourides, and H. Seker, "Prediction of influenza a virus infections in humans using an artificial neural network learning approach," in *2017 39th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*. IEEE, 2017, pp. 1186–1189.

[7] C. Chrysostomou, H. Seker, N. Aydin, and P. I. Haris, "Complex resonant recognition model in analysing influenza a virus subtype protein sequences," in *Information Technology and Applications in Biomedicine (ITAB), 2010 10th IEEE International Conference on*. IEEE, 2010, pp. 1–4.

[8] C. J. Carmona, C. Chrysostomou, H. Seker, and M. del Jesus, "Fuzzy rules for describing subgroups from influenza a virus using a multi-objective evolutionary algorithm," *Applied Soft Computing*, vol. 13, no. 8, pp. 3439–3448, 2013.

[9] C. Chrysostomou, H. Seker, and N. Aydin, "Effects of windowing and zero-padding on complex resonant recognition model for protein sequence analysis," in *Engineering in Medicine and Biology Society, EMBC, 2011 Annual International Conference of the IEEE*. IEEE, 2011, pp. 4955–4958.

[10] ——, "Cisaps: complex informational spectrum for the analysis of protein sequences," *Advances in bioinformatics*, vol. 2015, 2015.

[11] C. Chrysostomou and H. Seker, "Structural classification of protein sequences based on signal processing and support vector machines," in *Engineering in Medicine and Biology Society (EMBC), 2016 IEEE 38th Annual International Conference of the*. IEEE, 2016, pp. 3088–3091.

[12] Y. Zhang, B. D. Aevermann, T. K. Anderson, D. F. Burke, G. Dauphin, Z. Gu, S. He, S. Kumar, C. N. Larsen, A. J. Lee, *et al.*, "Influenza research database: An integrated bioinformatics resource for influenza virus research," *Nucleic acids research*, vol. 45, no. D1, pp. D466–D474, 2017.

[13] P. Argos, J. M. Rao, and P. A. Hargrave, "Structural prediction of membrane-bound proteins," *European Journal of Biochemistry*, vol. 128, no. 2-3, pp. 565–575, 1982.

[14] H. Zeng, M. D. Edwards, G. Liu, and D. K. Gifford, "Convolutional neural network architectures for predicting dna–protein binding," *Bioinformatics*, vol. 32, no. 12, pp. i121–i127, 2016.

[15] S. Wang, J. Peng, J. Ma, and J. Xu, "Protein secondary structure prediction using deep convolutional neural fields," *Scientific reports*, vol. 6, no. 1, pp. 1–11, 2016.

[16] Y. Zhao, N. He, Z. Chen, and L. Li, "Identification of protein lysine crotonylation sites by a deep learning framework with convolutional neural networks," *Ieee Access*, vol. 8, pp. 14 244–14 252, 2020.

[17] I. Goodfellow, Y. Bengio, A. Courville, and Y. Bengio, *Deep learning*. MIT press Cambridge, 2016, vol. 1, no. 2.

[18] A. L. Maas, A. Y. Hannun, and A. Y. Ng, "Rectifier nonlinearities improve neural network acoustic models," in *Proc. icml*, vol. 30, no. 1, 2013, p. 3.

[19] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[20] B. W. Matthews, "Comparison of the predicted and observed secondary structure of t4 phage lysozyme," *Biochimica et Biophysica Acta (BBA)-Protein Structure*, vol. 405, no. 2, pp. 442–451, 1975.

[21] Y. Sasaki *et al.*, "The truth of the f-measure. 2007," 2007.

[22] F. F. Sherif, N. Zayed, and M. Fakhr, "Classification of host origin in influenza a virus by transferring protein sequences into numerical feature vectors," *Int J Biol Biomed Eng*, vol. 11, 2017.