Contents lists available at ScienceDirect

Computers in Biology and Medicine

journal homepage: www.elsevier.com/locate/compbiomed

A novel data augmentation approach for influenza A subtype prediction based on HA proteins

Mohammad Amin Sohrabi [a], Fatemeh Zare-Mirakabad [b], Saeed Shiri Ghidary [c], Mahsa Saadat [b], Seyed-Ali Sadegh-Zadeh [c,*]

[a] Department of Mathematics and Computer Science, Amirkabir University of Technology, Tehran, Iran
[b] Computational Biology Research Center (CBRC), Department of Mathematics and Computer Science, Amirkabir University of Technology, Tehran, Iran
[c] Department of Computing, School of Digital, Technologies, and Arts, Staffordshire University, Stoke-On-Trent, UK

ARTICLE INFO

Keywords:
Supervised data augmentation
Protein data augmentation
Pre-trained transformer
NLP
PreIS

ABSTRACT

Influenza, a pervasive viral respiratory illness, remains a significant global health concern. The influenza A virus, capable of causing pandemics, necessitates timely identification of specific subtypes for effective prevention and control, as highlighted by the World Health Organization. The genetic diversity of influenza A virus, especially in the hemagglutinin protein, presents challenges for accurate subtype prediction. This study introduces PreIS as a novel pipeline utilizing advanced protein language models and supervised data augmentation to discern subtle differences in hemagglutinin protein sequences. PreIS demonstrates two key contributions: leveraging pre-trained protein language models for influenza subtype classification and utilizing supervised data augmentation to generate additional training data without extensive annotations. The effectiveness of the pipeline has been rigorously assessed through extensive experiments, demonstrating a superior performance with an impressive accuracy of 94.54% compared to the current state-of-the-art model, the MC-NN model, which achieves an accuracy of 89.6%. PreIS also exhibits proficiency in handling unknown subtypes, emphasizing the importance of early detection. Pioneering the classification of HxNy subtypes solely based on the hemagglutinin protein chain, this research sets a benchmark for future studies. These findings promise more precise and timely influenza subtype prediction, enhancing public health preparedness against influenza outbreaks and pandemics. The data and code underlying this article are available in https://github.com/CBRC-lab/PreIS.

1. Introduction

Influenza, a viral respiratory illness, afflicts millions of individuals worldwide annually [1]. Influenza viruses are classified into three phylogenetically distinct types: A, B, and C [2]. Influenza C primarily affects humans, has relatively few antigenic variations, and does not cause severe disease. Influenza B only occurs naturally in humans. In contrast, influenza A infects humans and a number of other mammals, including pigs, horses, and many poultry species. Influenza A virus (IAV) has much greater amino acid sequence variation than influenza B [3]. IAVs are the most widespread and virulent, capable of causing pandemics and major public health disruptions [4]. As a result, the World Health Organization (WHO) consistently emphasizes the importance of monitoring and tracking virus variations to promptly detect new subtypes, develop vaccines, and prevent catastrophic pandemics. Hence, this study specifically focuses on investigating IAV.

IAVs have 8 RNA segments which encode for the following proteins: Polymerase Subunit 1 (PB1), Polymerase Subunit 2 (PB2), Polymerase Acidic (PA), Nucleoprotein (NP), Matrix Proteins (which includes M1 and M2), Non-Structural Proteins (which includes NS1 and NEP), Neuraminidase (NA), and Hemagglutinin (HA) [5]. Out of all the segments, the HA and NA proteins are the two surface glycoproteins that undergo significant and are the primary targets for immunological detection [5]. Continuous antigenic drifts and occasional antigenic shifts in viral surface glycoproteins lead to difficulties in predicting and controlling epidemics [6]. To date, 18 HA subtypes and 11 NA subtypes have been identified [7–10].

IAV subtypes are named by combining the numbers H (HA) and N (NA), for example, H1N1 and H3N2. These IAV subtypes are currently prevalent in humans with varying sensitivities to antiviral drugs, so rapid classification of these viruses is becoming increasingly important [6].

[*] Corresponding author.
E-mail address: ali.sadegh-zadeh@staffs.ac.uk (S.-A. Sadegh-Zadeh).

https://doi.org/10.1016/j.compbiomed.2024.108316
Received 14 August 2023; Received in revised form 24 February 2024; Accepted 12 March 2024
Available online 13 March 2024

The Polymerase Chain Reaction (PCR) based approach [11,12] is a traditional lab method used for subtyping IVAs. This method utilizes a series of oligonucleotide primers in a single reverse transcription-PCR. A limitation of current subtyping techniques is that many assays are required to cover the broad range of circulating subtypes, making them costly and time-consuming.

With the advancement of sequencing technologies, the sequencing of influenza proteins has become the standard practice in influenza studies. Databases such as the National Center for Biotechnology Information (NCBI) Influenza Viruses database [13] provide convenient access to recently sequenced influenza protein sequences. Most computational methods for predicting and controlling epidemics focus on either the HA or NA protein sequences to predict 18 HA subtypes or 11 NA subtypes, respectively [14,15]. However, in the real-world scenarios, subtypes are defined based on variations in both HA and NA proteins. This is because a host infected with two different viral genotypes can produce a hybrid virus progeny with recombinant genotypes [16]. For example, coinfection with strains HxNy and HwNz, where $x, w \in \{1, \ldots, 18\}$ and $y, z \in \{1, \ldots, 11\}$, can produce hybrid strains HxNz and HwNy in addition to the parental types. This phenomenon is known as antigenic shift. Although the HA protein of IAV is known to mutate at a faster rate than the NA protein [17], it is generally recognized that for accurately predicting influenza subtypes, both the HA and NA proteins of each strain are essential [14]. Specifically, when two subtypes share the same HA protein subtype, such as HxNy and HxNz, a notable similarity between the HA proteins within these subtypes becomes apparent. Consequently, predicting the HxNy subtype of influenza can be particularly challenging, especially given that the NCBI often provides sequences for either the HA or NA proteins of a given strain but not both simultaneously.

In this study, our objective is to evaluate the potential of protein language models, as defined in protein sequences [18–22], commonly referred to as transformer-based approaches, in enhancing the ability of predictor models to discern between HxNy and HxNz subtypes. This investigation focuses on leveraging the nuanced distinctions within the HA protein sequences for subtype differentiation. To achieve this, we have designed an innovative pipeline named PreIS (Prediction of Influenza Subtype). Given the subtle distinctions between HA sequences in HxNy and HxNz subtypes, reflecting antigenic drift, we introduce a novel supervised data augmentation technique known as SDA (Supervised Data Augmentation). The SDA method simulates the antigenic drift process during training. By incorporating SDA, the model's performance is significantly enhanced, generating additional training data and thereby improving the diversity and quality of the dataset. Notably, our pipeline utilizes RITA [23] as a pre-trained protein language model for embedding, a model not previously explored for enhancing the accuracy and efficiency of IAV subtype prediction. These embedded sequences are then input into a multi-layer perceptrons (MLP) to predict the HxNy subtype.

We comprehensively evaluate the PreIS pipeline from two distinct perspectives. Initially, we present several versions of the PreIS pipeline to assess the performance of its individual steps, including data augmentation and embedding processes. Notably, the task of classifying HxNy subtypes has not been exclusively defined based on the single chain of the HA protein, resulting in a lack of previous models available for direct comparison. To establish a benchmark, we implement the MC-NN approach [14] and train it on our data to evaluate the performance of our pipeline. Remarkably, the MC-NN model achieves an accuracy of 89.6% on the test data, whereas our pipeline excels with an accuracy of 94.54%. Additionally, we scrutinize the model's capability to handle unknown classes (subtypes absent in the dataset), underscoring the importance of identifying new virus classes to preempt the emergence of potentially harmful ones.

2. Materials and methods

In this section, our objective is to introduce PreIS as an innovative pipeline designed to predict HxNy subtypes using only HA protein sequences. PreIS integrates a pre-trained transformer-based model with data augmentation techniques to enhance prediction accuracy. To provide a comprehensive introduction to PreIS, we begin by presenting the IAV subtypes utilized in this study. Subsequently, we introduce the dataset and define the predicting IAV subtypes problem, followed by a detailed explanation of the pipeline's steps.

2.1. *Influenza A virus subtype*

In this section, we introduce the significant IVA subtypes that form the basis of our study. Table 1 shows the importance of H1N1, H3N2, H5N1 and H7 subtypes extracted from Refs. [24–31].

2.2. *Data*

In order to gather the essential data for our study, we obtained the HA protein sequences of human, avian, and swine influenza viruses from the NCBI database [13]. The inclusion of avian and swine hosts in our study is motivated by the fact that these viruses have the potential to facilitate the transmission of avian and human influenza, which can lead to severe pandemics [32]. The NCBI database currently houses an extensive collection of over 900,000 protein sequences associated with the influenza virus. The specific query details used to extract the data for our study are outlined in Table 2.

The HA protein sequences are categorized into seven distinct classes as specified in Table 3. For each class, we randomly choose 1300 samples, resulting in a combined total of 9100 sequences. For each class, we select approximately 60 protein sequences. Given our pipeline's new data augmentation approach simulating antigenic drift, there's no need for additional data for training, validation, and set aside the remaining 8680 sequences exclusively for evaluating the performance of the model.

2.3. *Problem definition*

The problem of predicting IAV subtypes can be formulated as a computational problem in the following manner:

- **Input:** An HA protein sequence, $X = x_1 \ldots x_m$, where m represents the length of the sequence and x_i represents the i^{th} amino acid in sequence X.

Table 1
The importance of IAV subtypes.

IAV subtype	The importance of IAV subtypes
H1N1	The 1918 'Spanish Flu,' attributed to the H1N1 subtype, resulted in an estimated 50 million deaths [27,31].
H1N2	H1N2 features an HA component closely resembling that found in recent H1N1 strains and an NA component closely resembling that present in the presently circulating H3N2 strains. It seems this novel subtype emerged through the reassortment of these two human viruses [24–26].
H3N2	The 1968 'Hong Kong Flu' caused by the H3N2 subtype was relatively milder, yet it is estimated to have led to around 100,000 deaths. It's worth mentioning that the H3N2 subtype of influenza A has been the predominant cause of human infections and illnesses over the last four decades [27].
H5N1	The H5N1 subtype of IAV is a worldwide issue with the potential to trigger a pandemic [28,30].
H7	Instances of H7 avian influenza outbreaks in poultry are required to be reported to regulatory agencies due to the fact that specific strains of H7 subtypes can result in significant mortality among specific avian species and domestic fowl [29,30].

Table 2
The query used for data retrieval.

Option	Value
Sequence Type	Protein
Type	A
Host	Human, Avian, Swine
Country	All
Collapse Identical Sequences	✓
Subtype	H1N1, H1N2, H3, H5, H7

Table 3
IAV subtypes are defined into seven distinct classes.

Class	Subtype
0	H1N1
1	H1N2
2	H3N2
3	H3 — H3N2
4	H5N1
5	H5 — H5N1
6	H7

- **Output:** The task is to predict the IAV subtype, which is categorized into seven distinct classes as outlined in Table 3.

2.4. PreIS pipeline

We have developed a pipeline called PreIS for accurately predicting IAV subtypes. Fig. 1 illustrates the PreIS pipeline, outlining each step of the process in detail.

2.4.1. Supervised data augmentation

In the first step of PreIS pipeline (Fig. 1), after loading a mini-batch comprising an HA sequence and its corresponding label, we initiate the supervised data augmentation (SDA) function. Our data augmentation approach leverages labeled data to guide the generation of augmented samples, ensuring that the augmented data aligns more closely with the actual distribution of protein sequences. Through supervised control over the generation of augmented data, our objective is to maintain the antigenic drift characteristics inherent in HA protein sequences. The SDA function employs both global and local manipulations on the data, ensuring the model's output avoids the generation of incongruent samples (illustrated in Fig. 2).

To augment sequence X fetched from the training set, we take the following steps:

1. Selecting two random sequences S_g and S_l from the train set which have the same label as X.
2. For the global manipulation, we select a substring from the sequence S_g to substitute in the sequence X as follows:
   a. $m = min(|S_g|, |X|)$;
   b. Selecting randomly a consecutive of indexes from α to $\alpha + \lceil \gamma_g m \rceil$ where γ_g is obtained 0.4 by trial and error and α is a random integer in range 1 to $m - \lceil \gamma_g m \rceil$;
   c. Substituting subsequence of X in the range of α to $\alpha + \lceil \gamma_g m \rceil$ with the corresponding subsequence of S_g.

3. For local manipulation, we select a subsequence from the sequence S_l to substitute in the sequence X as follows:
   a. $m = min(|S_l|, |X|)$;
   b. Selecting $\lfloor \gamma_l m \rfloor$ random integers from 1 to m, where γ_l is obtained 0.1 by trial and error;
   c. Substituting items of X in $\lfloor \gamma_l m \rfloor$ randomly selected positions with corresponding items in S_l.

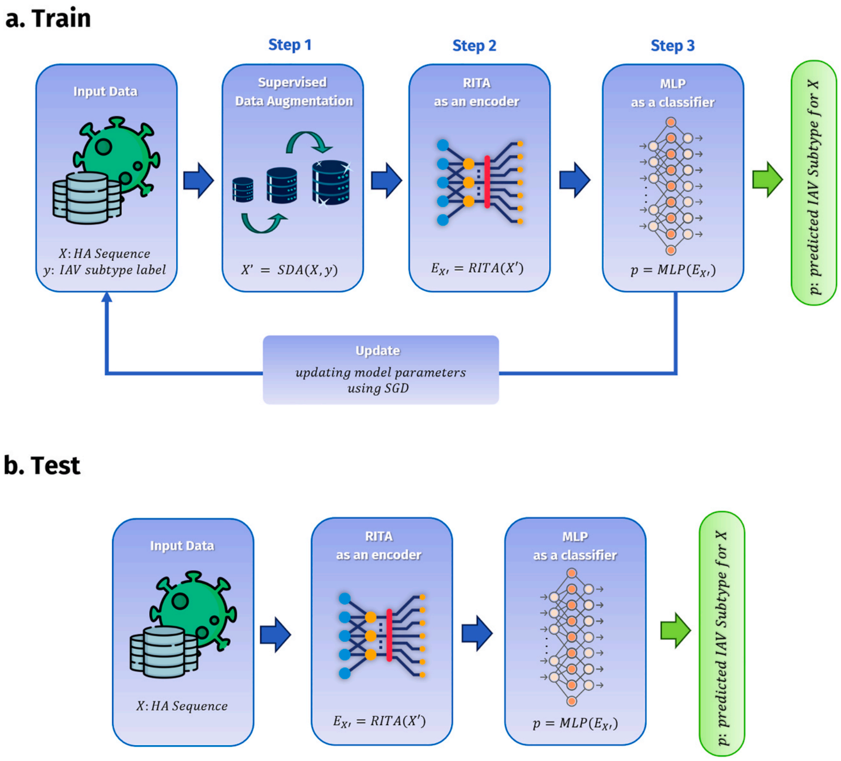

Fig. 1. The PreIS pipeline is comprised of two main parts: Train (part a) and Test (part b). In part a, an HA protein sequence X is loaded with its corresponding class label (y), and passed through the supervised data augmentation function to generate an augmented sequence X'. This augmented sequence is then processed through a pre-trained RITA model, producing a vector $E(X')$ that serves as input for the multi-layer perceptrons (MLP) to perform classification. In part b, an HA protein sequence X is loaded, encoded using the RITA model, and given as input to the MLP to predict its corresponding label.

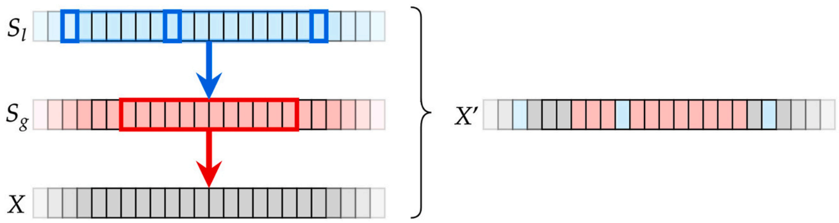

Fig. 2. A simple visualization of how SDA function works. Sequence X is augmented by two random training samples, S_l and S_g, with the same label as X, for the local and global manipulations, respectively to obtain $X^{'}$.

2.4.2. RITA as an encoder

Protein language models hold the potential to revolutionize our comprehension of proteins by unveiling the inherent rules that govern the language of proteins, relying solely on raw protein sequences [33]. A noteworthy recent development in protein representation learning is a transformer-based approach known as RITA [23]. In the second step of the PreIS pipeline, each augmented sequence generated by the SDA function undergoes processing through RITA [23]. As far as we are aware, RITA has not been utilized to tackle the IAV subtype problem, despite its established effectiveness in protein sequence classification. RITA employs a transformer decoder block architecture in an autoregressive manner and has shown that the model's capabilities on downstream tasks are correlated with its size. In our experiments, we use RITA as a pre-trained model. For each protein sequence $X = x_1 \dots x_m$ of length m, we utilize the hidden representation of the final layer of the RITA model to generate a token embedding vector for each amino acid x_i denoted as $e_i = [e_i^1, \dots, e_i^{768}]$.

To obtain a vector representation for sequence X in the same dimension as token embeddings, we employ global average pooling (GAP) as our combination strategy. It can be formulated as follows:

$$E_X = \left[E_X^1, .., E_X^{768}\right],$$

where $E_X^j = \frac{1}{m}\sum_{i=1}^m e_i^j, j = 1 \dots 768$.

Using GAP offers several advantages over a flatten layer. It acts as a structural regularizer, reducing the risk of overfitting [34], and it handles diverse sequence lengths more effectively.

2.4.3. Using multi-layer perceptrons as a classifier

In the third step of PreIS pipeline, a two-layer multi-layer perceptrons (MLP) is utilized as a classifier for predicting the IAV subtype (Fig. 3). Within our pipeline, the MLP serves as the classification head and is composed of a solitary hidden layer featuring 256 neurons. We opted for the Tanh function as the activation function. Our experiments have demonstrated that increasing the number of layers and neurons does not lead to significant changes in the results.

To ensure the predicted probability distribution aligns closely with the actual distribution, a cross-entropy loss function is defined as follows:

$$L(y,p) = -\sum_{k=0}^{6} y(k) logp(k),$$

where k shows the label of IAV subtype, y represents the actual influenza subtype, and p displays the predicted influenza subtype. The goal is to minimize the difference between the predicted and actual distributions through the loss function.

Finally, we optimize the performance of the pipeline and minimize the loss using the stochastic gradient descent (SGD) optimization algorithm. SGD is a widely used iterative optimization algorithm that aims to find the minimum of a given loss function. It achieves this by updating the model's parameters in the direction of steepest descent, considering small batches of training examples at each iteration. This approach ensures computational efficiency while effectively guiding the model towards better performance.

3. Results

In this section, we assess the performance of the PreIS pipeline in predicting IAV subtypes, taking into account the adjustment of

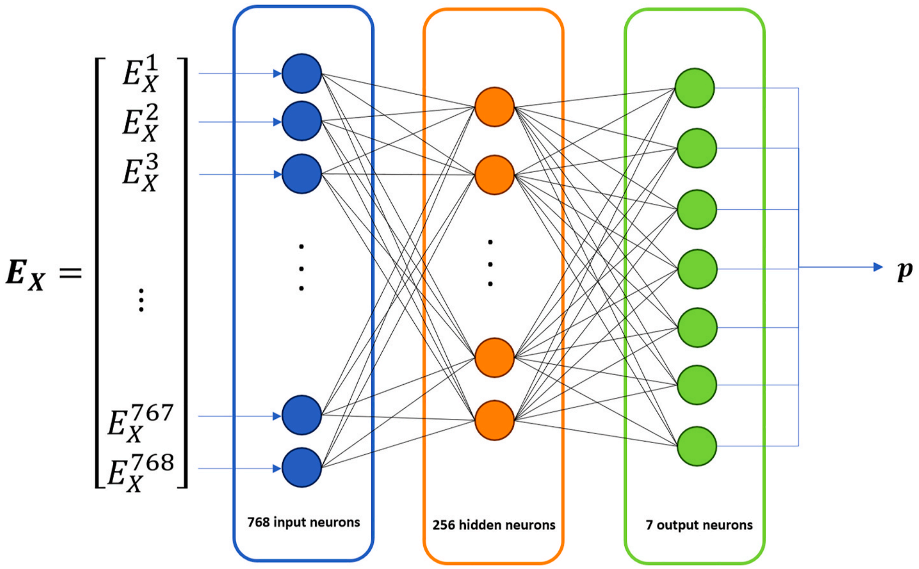

Fig. 3. MLP network architecture as a classifier. Illustration of the 7-dimensional vector p representing the probability distribution of input E_X (an embedding of protein sequence X) across the 7 classes.

hyperparameters. We investigate the impact of supervised data augmentation on the model's performance, along with assessing the effectiveness of employing a fine-tuned transformer called RITA instead of a pre-trained one.

3.1. *Experimental setup*

In the PreIS pipeline, we employ a compact version of the RITA model, consisting of 85 million parameters, along with an MLP-based classification head. To optimize model performance, we set the dropout probability to 10% and determine that a learning rate of 4e-5, in conjunction with a scheduler, yields favourable outcomes. The batch size is configured to 8, utilizing 2 accumulation steps, and we select the best model based on its performance on the validation set for subsequent testing. During training, the PreIS pipeline undergoes 200 epochs while concurrently fine-tuning the RITA model. Alternatively, when RITA is solely utilized for feature extraction, the pipeline undergoes 400 epochs.

Accuracy is a prevalent metric in machine learning used to assess the performance of a classification model. It measures the proportion of correct predictions made by the model out of the total number of predictions. The accuracy score is the percentage of predictions that are correct.

To comprehensively evaluate the performance of the classification model, we employ a confusion matrix. This table is utilized to compare the predicted labels with the actual labels, offering a detailed breakdown of the model's predictions for each class. By carefully analysing the confusion matrix, we can gain insights into the strengths and weaknesses of the model and identify areas for improvement.

3.2. *Using RITA as a pre-trained transformer*

In the following, two versions of the PreIS pipeline are introduced that utilize the pre-trained RITA to extract an embedding as a feature vector (FV) for each protein sequence:

- FV-SDA: This version includes the SDA function for supervised data augmentation.
- FV-NoSDA: This version relies solely on the real data without any data augmentation.

According to Table 4, the FV-SDA model achieves an accuracy of 90.70%. On the other hand, the FV-NoSDA model struggles with overfitting, resulting in a significantly lower accuracy of 45.46%, despite achieving a train accuracy of approximately 98%. This disparity in performance emphasizes the significance of addressing data scarcity when utilizing static sequence embeddings. Without sufficient data and fine-tuning of the RITA model, overfitting is likely to occur. However, the inclusion of an augmentation strategy like SDA can effectively mitigate this issue.

Table 4
The accuracy of test set. FV stands for the feature vector, and FT refers to fine-tuning of RITA. We also compare different data augmentation methods: NoSDA (without supervised data augmentation), LSDA (local supervised data augmentation), GSDA (global supervised data augmentation), and UnSDA (unsupervised data augmentation).

Versions of PreIS	Accuracy (%)
FV-NoSDA	45.46
FV-SDA	90.70
FT - NoSDA	91.72
FT - SDA	94.54
FT - UnSDA	90.05
FT – LSDA	92.44
FT – GSDA	93.19

3.3. *Using the fine-tuning RITA*

Here, two versions of PreIS pipeline are defined based on fine-tuning (FT) RITA along with the parameters of the MLP:

- FT-SDA: This version includes fine-tuning the RITA leveraging the SDA method.
- FT-NoSDA: This version fine tunes the RITA without utilizing SDA.

According to Table 4, the FT-SDA model exhibits a 2.82% increase in accuracy compared to the FT-NoSDA model. To gain further insights into the performance of both models, we present the confusion matrices in Fig. 4, which provide a comprehensive overview of the classification results by illustrating the distribution of predicted labels against the true labels. As indicated in Table 4, the FT-SDA model achieves an accuracy of 94.54%, while the model trained without any data augmentation achieves an accuracy of 91.72%. This suggests that fine-tuning the RITA model as a pre-trained model can enhance the generalizability of the model compared to a feature-based training approach. Additionally, it is evident that the incorporation of the SDA method on protein sequences has a significant positive impact on the model's performance within the fine-tuning approach.

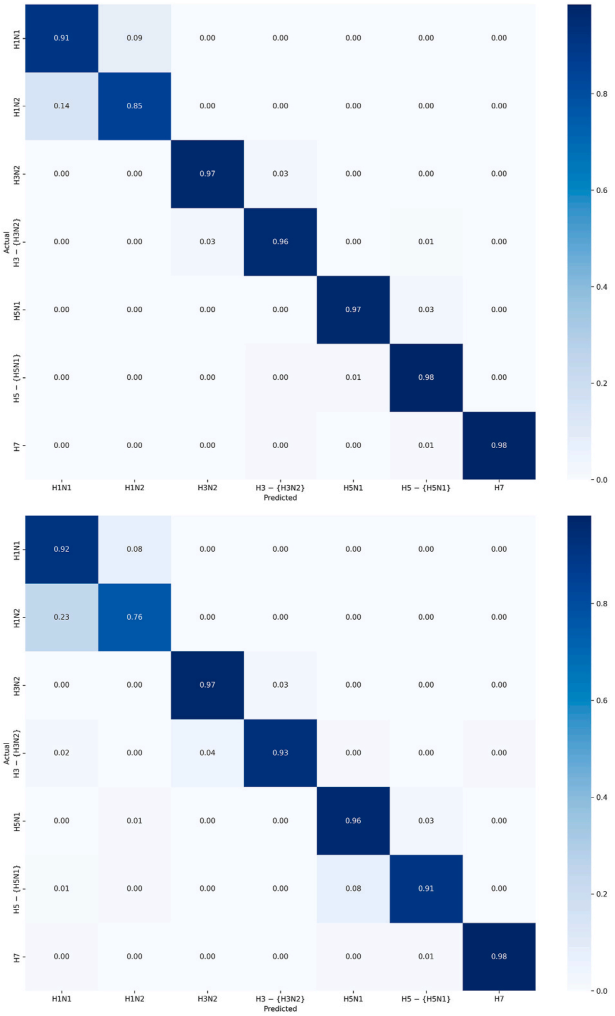

Fig. 4. Confusion matrices for FT-SDA (up) and FT-NoSDA (down). FT refers to the fine-tuning of RITA, SDA denotes the absence of supervised data augmentation, and NoSDA refers to the condition without any supervised data augmentation.

3.4. Comparing supervised and unsupervised data augmentation

In this section, we demonstrate the effective performance of supervised data augmentation in simulating antigenic drifts in HA protein. Based on supervised and unsupervised models for data augmentation three versions of PreIS pipeline are defined based on fine-tuning RITA:

- FT-SDA: This version includes SDA function for supervised data augmentation.
- FT-UnSDA: This version alternates SDA function to augment data based on unsupervised function. This function changes step 1 of SDA function where S_g and S_l are randomly selected from the training set without considering their labels.
- FT-NoSDA: This version does not augment data.

Table 4 shows that the FT-UnSDA pipeline exhibits a decrease in performance compared to the FT-NoSDA pipeline. This outcome emphasizes the importance of supervision in the data augmentation process. This supervision actually simulates the antigenic drifts in the HA protein, while the unsupervised mode simulates the antigenic shifts. As demonstrated in the results, the lack of supervision in the augmentation process resulted in a performance of 90.05%, which is even lower than the model without augmentation. These findings highlight the effectiveness of preserving the distribution of the data through supervision in enhancing the generalizability of the model.

Fig. 5 showcases the impact of data augmentation and presents the outcomes of applying Principal Component Analysis (PCA) to each vector E_X, where X is derived from the test set. The results indicate that even when SDA is employed, the various influenza subtypes remain distinguishable. This observation suggests that the augmented data retains the distinctive features of each class, indicating that the data augmentation technique used is successful in preserving the unique characteristics of the different influenza subtypes.

3.5. Local and global manipulation in data augmentation

Based on local and global manipulation in SDA function for data augmentation three versions of PreIS pipeline are compared based on fine-tuning the RITA:

- FT-SDA: This version includes SDA function for supervised data augmentation in local and global manipulation ($\gamma_g = 0.4$, $\gamma_l = 0.1$).
- FT-LSDA: This version includes SDA that is solely based on the local manipulation ($\gamma_g = 0$, $\gamma_l = 0.1$).
- FT-GSDA: This version includes SDA that is solely based on the global manipulation ($\gamma_g = 0.4$, $\gamma_l = 0$).

The results indicate that the FT-SDA version, which combines both local and global data manipulation, achieves the highest level of success. This version achieves a classification accuracy of 94.54%. On the other hand, the FT-LSDA version, which only utilizes local data manipulation, achieves an accuracy of 92.44%. Similarly, the FT-GSDA version, which solely employs global augmentation, attains an accuracy of 93.19%.

The superior performance of the FT-SDA version can be attributed to the fact that global and local augmentations attempt to replicate distinct aspects of protein sequence evolution. They mimic changes such as insertion and deletion, amino acid mutations, and antigenic drift. These factors contribute to their effectiveness in improving classification accuracy.

4. Discussion

In this section, we compare SDA part of the pipeline to the similar substitution (SS) method [35]. Since we have defined the problem of using HA to predict HxNy for the first time, there is no existing state-of-the-art model for direct comparison. As a result, we adapt the architecture of the most recent model designed for predicting Hx and Ny subtypes individually [14] to predict HxNy subtype. Then we compare its performance with our own pipeline. Finally, we examine a scenario where an unknown subtype is treated as a class, representing subtypes that are not included in the seven classes represented in the dataset.

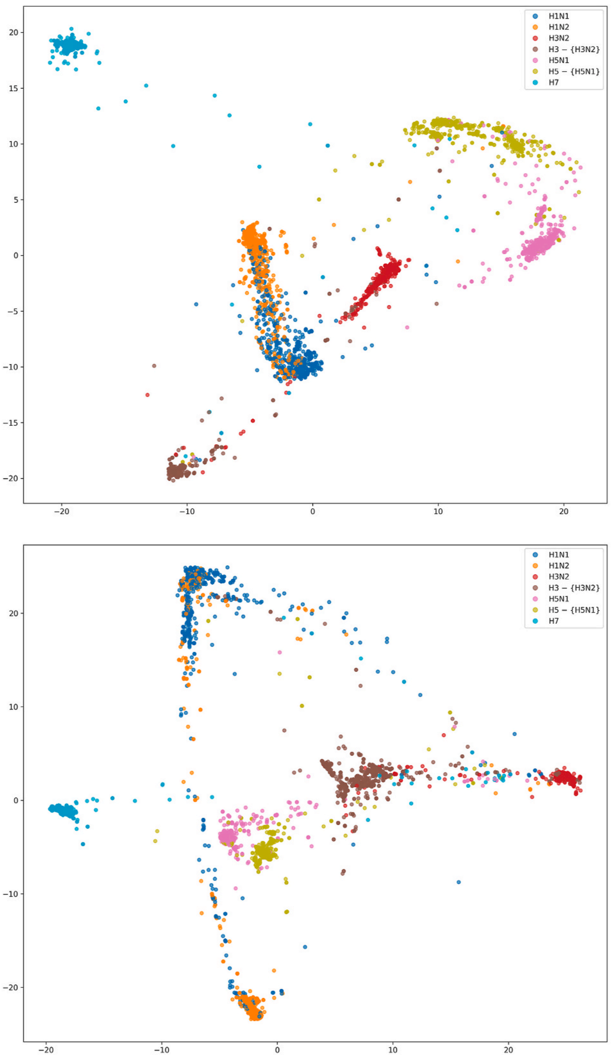

Fig. 5. The effectiveness of data augmentation in preserving the unique characteristics of each class can be observed through the extracted features shown in FT-SDA (up) and FT-NoSDA (down). The preserved features in the augmented data suggest that supervised data augmentation can help improve the quality and quantity of data for machine learning models.

4.1. Comparing the amino acid composition of real and synthetic data

In this section, the amino acid composition (AAC) of both synthetic and real data is examined. Fig. 6 presents density plots for eight amino acids, comparing the real data with data generated using two methods: the SDA function and SS method. We define three different versions of SDA function using combinations of local and global manipulation factors. The SS method in the work of [35], called replacement dictionary, involves randomly substituting amino acids in a primary sequence with similar amino acids based on substitution rules. Each amino acid is independently replaced with a probability denoted as p. In their study, they found the best substitution as follows: [[A,V], [S,T], [F,Y], [K,R], [C,M], [D,E], [N,Q], [V,I]]. We have adopted the same mapping for our study.

Fig. 6 displays the kernel density estimate (KDE) diagram of eight

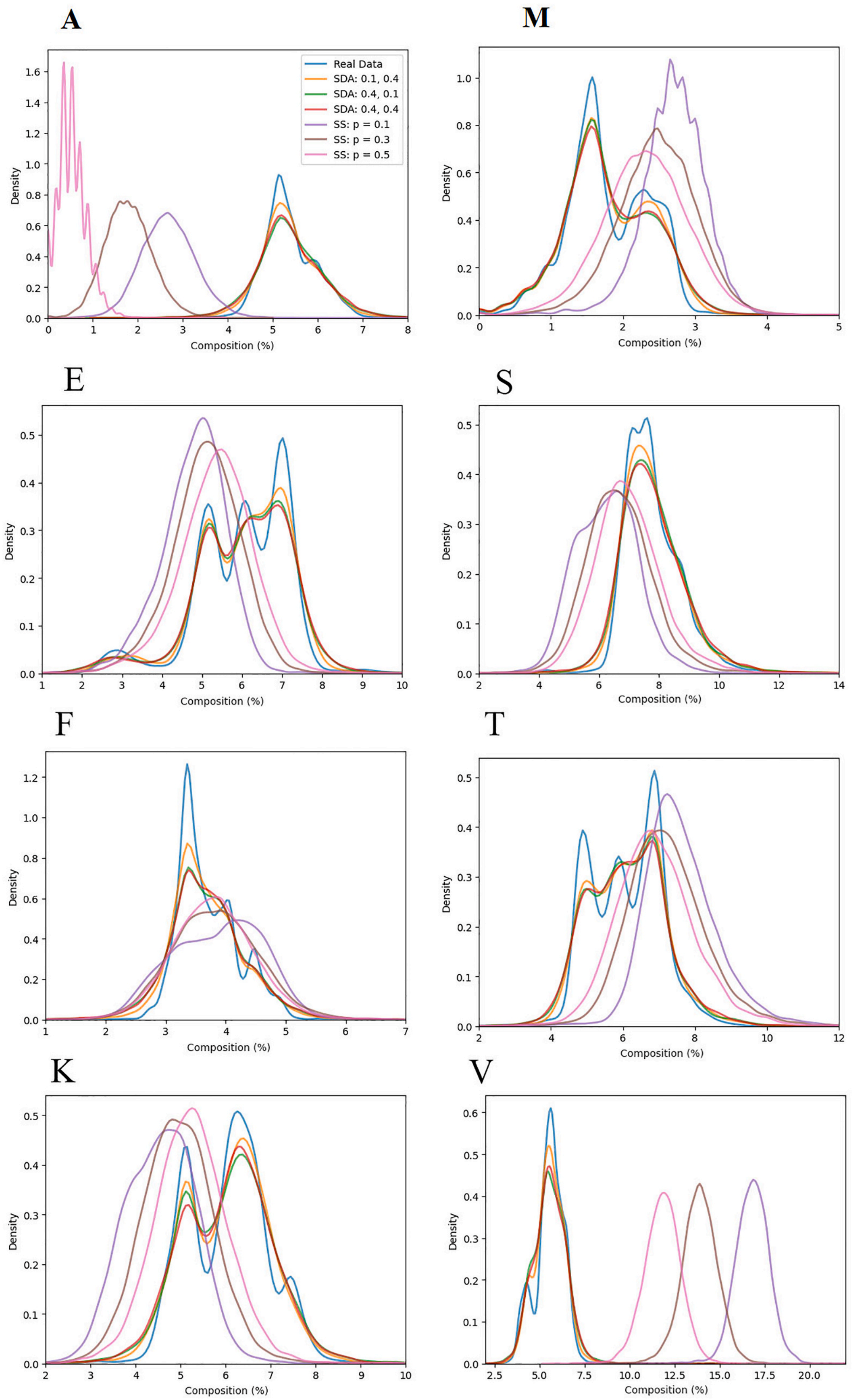

Fig. 6. KDE plots for amino acid composition, real vs synthetic data. In the case of SDA, the first value represents γ_l, while the second value represents γ_g. On the other hand, SS refers to the similar substitution method, with a probability denoted as p. The x-axis shows the percentage of presences of the amino acid in each protein sequence, and the y-axis rep-resents the probability density function of sequences that include the amino acid.

amino acids: A, M, A, S, F, T, K, and D. A KDE plot is a way to visualize the distribution of data using a continuous probability density curve, similar to a histogram. According to Fig. 6, SDA allows to generate synthetic data based on data-driven assumptions, ensuring the preservation of the data distribution. On the other hand, the SS method generates data that completely alters the distribution of the real data.

4.2. *Comparison with state-of-art*

As we have introduced the novel challenge of utilizing HA protein sequences for the prediction of HxNy subtypes, there are no established state-of-the-art models available for direct benchmarking. Consequently, we customize the architecture of the latest model, originally designed for predicting Hx and Ny subtypes separately [14], to address the specific task of HxNy subtype prediction. In this work, the researchers employed a transformer encoder block, which was followed by dense layers. Notably, they utilized a 3-g tokenization approach and incorporated sinusoidal positional encoding. According to our results, PreIS achieved an accuracy of 94.54%, while this model yielded 89.6%.

4.3. *Evaluation using unknown class*

The level of certainty expressed by a machine learning model in its predictions can be measured by a critical metric known as the confidence score. This metric plays a crucial role in evaluating the reliability and accuracy of the model's performance, serving as an indicator of the confidence level the model holds in its predictions. This score is often employed to identify cases where the model may display uncertainty or require further refinement. Typically ranging between 0 and 1, the confidence score represents a probability value that reflects the likelihood of a prediction being correct.

The main objective of prediction models is to facilitate the rapid and accurate diagnosis of unknown subtypes of the influenza A virus. These models play a critical role in enhancing influenza surveillance and controlling its transmission by enabling the precise identification of virus hosts and subtypes.

In order to evaluate the effectiveness of the PreIS pipeline in identifying new virus subtypes, a special class called ‘unknown’ is included during the testing phase of the model. This class consists of HA protein sequences whose subtypes are not included in the 7 classes used for training the model. Fig. 7 illustrates a graph displaying confidence score values, which provides insights into the model's performance in recognizing the unknown class within the test data. It is observed that when the SDA approach is employed during model training, the model exhibits improved performance in identifying the unknown class accurately. Specifically, the model trained using the SDA approach exhibits relatively low confidence scores when faced with unknown data. Conversely, this is not observed in the model that does not employ SDA. In the model trained Using SDA, whenever it assigns a low confidence score to a specific sequence, it could potentially indicate the presence of a novel subtype and open up new possibilities for investigation.

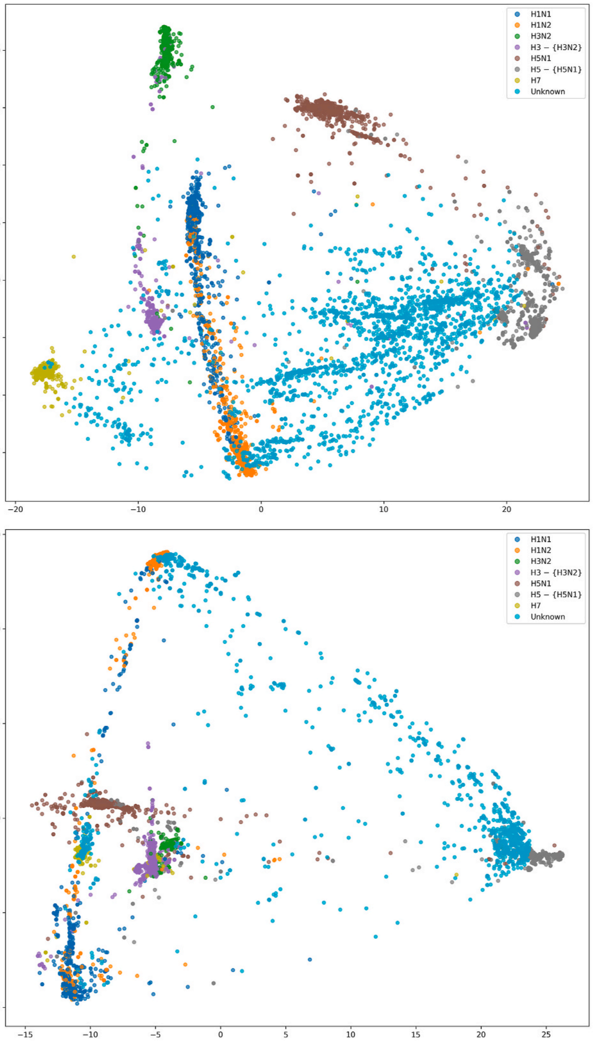

Fig. 8. Extracted features in FT-SDA (up) and FT-NoSDA (down) demonstrate data augmentation's effectiveness in preserving class-specific characteristics. To facilitate better visualization, unknown samples with 42 different labels were plotted using a single color.

Fig. 8 shows the results of applying PCA to each E_X vector where X is derived from the test set. It provides evidence of the effectiveness of the SDA approach in detecting unknown subtypes when extracting features using RITA, without the need for fine-tuning. Additionally, the data from the model that incorporates SDA demonstrates a reduction in overlap among different classes compared to the model that does not utilize SDA.

5. Conclusion

Previous computational methods for predicting IAV subtypes have

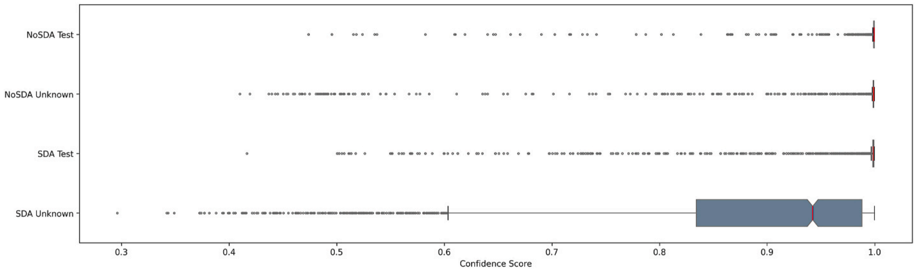

Fig. 7. The confidence score values are presented in four different tests, including the presence and absence of data augmentation for two test groups: one with the unknown class and the other without the unknown class.

focused on predicting either HA or NA subtypes, not both, as in the case of HxNy. This limitation arises from the fact that the NCBI often provides sequences for either the HA or NA proteins of a given strain, but not both simultaneously. However, in real-world scenarios, subtypes are defined based on variations in both HA and NA proteins.

The primary challenge addressed in this study was predicting HxNy subtypes in IAV solely using HA protein sequences. Specifically, we investigated whether protein language models, defined on protein sequences, could enable predictor models to distinguish between HxNy and HxNz subtypes based on the minor differences in HA protein sequences. To address this challenge, we introduced a novel pipeline named PreIS. Our pipeline incorporates a new supervised data augmentation method designed to generate additional training data, thereby enhancing dataset diversity and quality during model training. Additionally, we utilized a pre-trained protein language model for protein sequence embedding, named RITA. Different versions of PreIS were defined to assess the impact of each step on IAV subtype prediction. The results indicate that supervised data augmentation and fine-tuning RITA significantly improve IAV subtype prediction.

While our study demonstrated the effectiveness of supervised data augmentation in simulating antigenic drift in HA proteins, there is still room for enhancing the efficiency of each pipeline step for more accurate IAV subtype prediction. Considering the rapid advancements in deep learning within the healthcare domain [36–38] and recognizing the significance of overcoming data limitations through data augmentation, we propose exploring advanced deep learning approaches beyond the current MLP step for IAV classification.

CRediT authorship contribution statement

Mohammad Amin Sohrabi: Writing – original draft, Visualization, Validation, Software, Resources, Methodology, Formal analysis, Data curation, Conceptualization. **Fatemeh Zare-Mirakabad:** Writing – review & editing, Writing – original draft, Validation, Investigation, Formal analysis, Data curation, Conceptualization. **Saeed Shiri Ghidary:** Writing – original draft, Methodology, Conceptualization. **Mahsa Saadat:** Writing – review & editing, Visualization, Conceptualization. **Seyed-Ali Sadegh-Zadeh:** Writing – original draft, Supervision, Methodology.

Declaration of competing interest

We declare that there is no conflict of interest regarding the research presented in this paper. No financial or personal relationships have influenced the design, execution, or interpretation of the study. This work is solely based on our objective investigation and analysis of the subject matter, and we have no affiliations with any organizations or individuals that could be perceived as affecting the integrity of our findings.

References

[1] K.E. Lafond, et al., Global burden of influenza-associated lower respiratory tract infections and hospitalizations among adults: a systematic review and meta-analysis, PLoS Med. 18 (3) (Mar. 2021) e1003550, https://doi.org/10.1371/journal.pmed.1003550.

[2] R.A. Lamb, R.M. Krug, in: D.M. Knipe, P.M. Howley (Eds.), Orthomyxoviridae: the Viruses and Their Replication, fourth ed.Fields Virology, Lippincott Williams & Wilkins, Philadelphia, 2001, pp. 1487–1532.

[3] Steven A. Frank, Experimental Evolution: Influenza," in *Immunology and Evolution Of Infectious Disease*, Princeton University Press, Princeton and Oxford, 2002, pp. 205–229.

[4] C.E. Mills, J.M. Robins, M. Lipsitch, Transmissibility of 1918 pandemic influenza, Nature 432 (7019) (2004) 904–906, https://doi.org/10.1038/nature03063.

[5] C. Chrysostomou, F. Alexandrou, M.A. Nicolaou, H. Seker, Classification of influenza hemagglutinin protein sequences using convolutional neural networks, in: 2021 43rd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC), IEEE, Nov. 2021, pp. 1682–1685, https://doi.org/10.1109/EMBC46164.2021.9630673.

[6] Y. Wang, J. Bao, J. Du, Y. Li, Rapid detection and prediction of influenza A subtype using deep convolutional neural network based ensemble learning, in: Proceedings of the 2020 10th International Conference on Bioscience, Biochemistry and Bioinformatics, New York, NY, USA: ACM, Jan. 2020, pp. 47–51, https://doi.org/10.1145/3386052.3386053.

[7] S.A. Valkenburg, et al., Stalking influenza by vaccination with pre-fusion headless HA mini-stem, Sci. Rep. 6 (1) (Mar. 2016) 22666, https://doi.org/10.1038/srep22666.

[8] C.-Y. Wu, et al., Influenza A surface glycosylation and vaccine design, Proc. Natl. Acad. Sci. USA 114 (2) (Jan. 2017) 280–285, https://doi.org/10.1073/pnas.1617174114.

[9] Y. Peng, et al., A universal computational model for predicting antigenic variants of influenza A virus based on conserved antigenic structures, Sci. Rep. 7 (1) (Feb. 2017) 42051, https://doi.org/10.1038/srep42051.

[10] M.G. Joyce, et al., Vaccine-induced antibodies that neutralize group 1 and group 2 influenza A viruses, Cell 166 (3) (Jul. 2016) 609–623, https://doi.org/10.1016/j.cell.2016.06.043.

[11] M. Vinikoor, J. Stevens, J. Nawrocki, K. Singh, Influenza A virus subtyping: paradigm shift in influenza diagnosis, J. Clin. Microbiol. 47 (9) (Sep. 2009) 3055–3056, https://doi.org/10.1128/JCM.01388-09.

[12] K.E. Wright, G.A. Wilson, D. Novosad, C. Dimock, D. Tan, J.M. Weber, Typing and subtyping of influenza viruses in clinical samples by PCR, J. Clin. Microbiol. 33 (5) (May 1995) 1180–1184, https://doi.org/10.1128/jcm.33.5.1180-1184.1995.

[13] E.W. Sayers, et al., Database resources of the national center for biotechnology information, Nucleic Acids Res. 50 (D1) (Jan. 2022) D20–D26, https://doi.org/10.1093/nar/gkab1112.

[14] Y. Xu, D. Wojtczak, MC-NN: an end-to-end multi-channel neural network approach for predicting influenza A virus hosts and antigenic types, SN Comput Sci 4 (5) (Jun. 2023) 435, https://doi.org/10.1007/s42979-023-01839-5.

[15] P.K. Attaluri, Z. Chen, A.M. Weerakoon, G. Lu, Integrating decision tree and hidden markov model (hmm) for subtype prediction of human influenza A virus, Communications in Computer and Information Science 35 (2009) 52–58, https://doi.org/10.1007/978-3-642-02298-2_8.

[16] C. Scholtissek, Source for influenza pandemics, Eur. J. Epidemiol. 10 (4) (Aug. 1994) 455–458, https://doi.org/10.1007/BF01719674.

[17] T.K.W. Cheung, L.L.M. Poon, Biology of influenza A virus, in: Annals of the New York Academy of Sciences, Blackwell Publishing Inc., 2007, pp. 1–25, https://doi.org/10.1196/annals.1408.001.

[18] N. Brandes, D. Ofer, Y. Peleg, N. Rappoport, M. Linial, ProteinBERT: a universal deep-learning model of protein sequence and function, Bioinformatics 38 (8) (Apr. 2022) 2102–2110, https://doi.org/10.1093/bioinformatics/btac020.

[19] M. Filipavicius, M. Manica, J. Cadow, M.R. Martinez, Pre-training Protein Language Models with Label-Agnostic Binding Pairs Enhances Performance in Downstream Tasks, Dec. 2020 [Online]. Available: http://arxiv.org/abs/2012.03084.

[20] A. Elnaggar, et al., ProtTrans: toward understanding the language of life through self-supervised learning, IEEE Trans. Pattern Anal. Mach. Intell. 44 (10) (Oct. 2022) 7112–7127, https://doi.org/10.1109/TPAMI.2021.3095381.

[21] A. Rives, et al., Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences, Proc. Natl. Acad. Sci. USA 118 (15) (Apr. 2021), https://doi.org/10.1073/pnas.2016239118.

[22] R. Rao, et al., Evaluating protein transfer learning with TAPE, Adv. Neural Inf. Process. Syst. 32 (Dec. 2019) 9689–9701.

[23] D. Hesslow, N. Zanichelli, P. Notin, I. Poli, D. Marks, RITA, A Study on Scaling up Generative Protein Sequence Models, May 2022 [Online]. Available: http://arxiv.org/abs/2205.05789.

[24] N. Komadina, J. McVernon, R. Hall, K. Leder, A historical perspective of influenza A(H1N2) virus, Emerg. Infect. Dis. 20 (1) (Jan. 2014) 6–12, https://doi.org/10.3201/eid2001.121848.

[25] R. Sah, A. Mohanty, R. Rohilla, B.K. Padhi, A recent outbreak of human H1N2 infection: correspondence, Int. J. Surg. 109 (3) (Mar. 2023) 604–605, https://doi.org/10.1097/JS9.0000000000000185.

[26] J.-R. Yang, et al., Human infection with a reassortant swine-origin influenza A (H1N2)v virus in Taiwan, 2021, Virol. J. 19 (1) (Dec. 2022) 63, https://doi.org/10.1186/s12985-022-01794-2.

[27] S. Vemula, J. Zhao, J. Liu, X. Wang, S. Biswas, I. Hewlett, Current approaches for diagnosis of influenza virus infections in humans, Viruses 8 (4) (Apr. 2016) 96, https://doi.org/10.3390/v8040096.

[28] H.S. Abd Raman, S. Tan, J.T. August, A.M. Khan, Dynamics of Influenza A (H5N1) virus protein sequence diversity, PeerJ 7 (May 2020) e7954, https://doi.org/10.7717/peerj.7954.

[29] S. Su, Y. Bi, G. Wong, G.C. Gray, G.F. Gao, S. Li, Epidemiology, evolution, and recent outbreaks of avian influenza virus in China, J. Virol. 89 (17) (Sep. 2015) 8671–8676, https://doi.org/10.1128/JVI.01034-15.

[30] Y. Poovorawan, S. Pyungporn, S. Prachayangprecha, J. Makkoch, Global alert to avian influenza virus infection: from H5N1 to H7N9, Pathog. Glob. Health 107 (5) (Jul. 2013) 217–223, https://doi.org/10.1179/2047773213Y.0000000103.

[31] D.M. Morens, J.K. Taubenberger, Influenza cataclysm, 1918, N. Engl. J. Med. 379 (24) (Dec. 2018) 2285–2287, https://doi.org/10.1056/NEJMp1814447.

[32] P.K. Attaluri, Z. Chen, G. Lu, Applying neural networks to classify influenza virus antigenic types and hosts, in: 2010 IEEE Symposium on Computational Intelligence in Bioinformatics and Computational Biology, CIBCB 2010, 2010, pp. 279–284, https://doi.org/10.1109/CIBCB.2010.5510726.

[33] A. Behjati, F. Zare-Mirakabad, S.S. Arab, A. Nowzari-Dalini, Protein sequence profile prediction using ProtAlbert transformer, Comput. Biol. Chem. 99 (Aug. 2022) 107717, https://doi.org/10.1016/j.compbiolchem.2022.107717.

[34] M. Lin, Q. Chen, S. Yan, Network in network [Online]. Available: http://arxiv.org/abs/1312.4400, Dec. 2013.

[35] S. French, B. Robson, What is a conservative substitution? J. Mol. Evol. 19 (2) (Mar. 1983) 171–175, https://doi.org/10.1007/BF02300754.

[36] G. Gupta, W. Salehi, A prospective and comparative study of machine and deep learning techniques for smart healthcare applications, in: Mobile Health : Advances in Research and Applications, Nova Science Publishers, Inc., 2021, pp. 163–189.

[37] A. Waleed Salehi, P. Baglat, G. Gupta, Review on machine and deep learning models for the detection and prediction of Coronavirus, Mater. Today Proc. 33 (2020) 3896–3901, https://doi.org/10.1016/j.matpr.2020.06.245.

[38] Bharti Thakur, Nagesh Kumar, Gaurav Gupta, Machine learning techniques with ANOVA for the prediction of breast cancer, International Journal of Advanced Technology and Engineering Exploration 9 (87) (Feb. 2022), https://doi.org/10.19101/IJATEE.2021.874555.