Bioinformatics, YYYY, 0–0
doi: 10.1093/bioinformatics/xxxxx
Advance Access Publication Date: DD Month YYYY
Original Paper

Gene Expression

Forecasting symptomatic influenza A infection based on pre-exposure gene expression using a Deep Learning approach

Anna Zan[1], Zhong-Ru Xie[2], Yi-Chen Hsu[1], Yu-Hao Chen[1], Tsung-Hsien Lin[1], Yong-Shan Chang[1] and Kuan Y. Chang [1,*]

[1]Computational Biology Laboratory, Department of Computer Science and Engineering, National Taiwan Ocean University, Keelung, Taiwan, ROC, [2]Computational Drug Discovery Laboratory, School of Electrical and Computer Engineering, College of Engineering, University of Georgia, Athens, GA, USA

*To whom correspondence should be addressed.

Associate Editor: XXXXXXX

Received on 11 February 2021; revised on XXXXX; accepted on XXXXX

Abstract

Motivation: Not everyone gets sick after an exposure to influenza A viruses (IAV). It was unclear whether forecasting symptomatic flu infection based on pre-exposure host gene expression would be possible. To examine this problem, we applied the state-of-the-art deep learning approach on the human gene expression data infected with IAV subtype H1N1 or H3N2 viruses.

Results: The results indicate that such forecast is possible and, in other words, gene expression can reflect the strength of one's immune system against IAV infection. Here we have developed DeepFlu—based on deep neural networks—to predict who would catch the flu prior to an exposure to IAV. We validated DeepFlu both on a leave-one-person-out cross-validation and an independent external cohort. Better than the models using convolutional neural network, random forest, and support vector machine, DeepFlu achieved 71.4% accuracy, an area under the receiver operating characteristics curve (AUROC) of 0.700, and an area under the precision-recall curve (AUPR) of 0.723 on the external H1N1 data (healthy n=26, ill n=12), and 73.5% accuracy, an AUROC of 0.732, and an AUPR of 0.749 on the external H3N2 data (healthy n=13, ill n=8). DeepFlu worked the best when trained only by the pre-exposure data than by other time spans. Besides, mixed training data of H1N1 and H3N2 did not necessarily enhance prediction. The findings suggest that deep learning models can recognize individuals susceptible to flu to help prevent the spread of IAV.

Availability and Implementation: A python implementation of DeepFlu is available at https://github.com/ntou-compbio/DeepFlu.

Contact: kchang@ntou.edu.tw (KC)

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

Influenza is a highly contagious and deadly viral respiratory disease second only to coronavirus disease 2019 (covid-19). There are about 3 to 5 million severe cases of influenza worldwide every year, resulting in approximately 291,000 to 646,000 deaths annually (Iuliano *et al.*, 2018; WHO, 2018). Influenza A caused by influenza A virus (IAV) accounts for about 75% of all influenza infections (Heikkinen *et al.*, 2014). IAV is also responsible for 1918, 1957, 1968, and 2009 influenza pandemics (Paules and Subbarao, 2017). About 2.3% of adults have contracted IAV (Jayasundara *et al.*, 2014; Park and Ryu, 2018), but not everyone gets sick after an exposure to IAV (Iwasaki and Pillai, 2014). Slightly more than half of young adults would get infected and develop symptoms of acute upper respiratory tract infection after a controlled IAV exposure; the rest remain either asymptomatic or uninfected (Fourati *et al.*, 2018; Zaas *et al.*, 2009).

Due to the heavy socioeconomic and health burden of influenza (Cassini *et al.*, 2018; Gordon and Reingold, 2018; Putri *et al.*, 2018), it is important to identify and anticipate who will contract influenza and

become ill. Delaying the recognition of infected persons with IAV will increase the viral transmission (Poon *et al.*, 2016) and sometimes worsen the infection, thereby increasing the death toll. Although vaccination is recommended to prevent the influenza and improve public health in schools, hospitals, communities, and societies (Houser and Subbarao, 2015), a flu vaccine does not necessarily work for everyone (Nakaya *et al.*, 2011). Instead, foreseeing who will succumb to IAV infection prior to viral exposure can be an alternative to prevent the spread of IAV and reduce the death toll from the IAV and its related complications.

However, it remains unclear whether host susceptibility to IAV is predictable based on pre-viral-exposure host gene expression. Forecasting the onset of viral infection prior to viral exposure has been only demonstrated on respiratory syncytial virus (RSV) (Fourati *et al.*, 2018). Two prediction models have shown promising results for forecasting host susceptibility to RSV. One is an epsilon support vector regression model (Chang and Lin, 2011) with a radial basis function kernel (Fourati *et al.*, 2018) learning directly from host gene expression. The other is a LASSO regularized regression model (Fourati *et al.*, 2018) based on modulation of biological pathways, deriving from the Hallmark gene sets (Liberzon *et al.*, 2015) of the Molecular Signature Database (MSigDB). No one yet has demonstrated whether forecasting symptomatic IAV infection based on pre-viral-exposure host gene expression is feasible and, if so, how accurate the forecast is.

Forecasting susceptibility to IAV differs from detecting IAV infection during the onset of influenza symptoms based on host gene expression. Although both host gene expression are influenced by the same factors such as host genetics (Casanova and Abel, 2004), immunological memory (Zhang *et al.*, 2019), seasonal and circadian changes (Dopico *et al.*, 2015), age (Qi *et al.*, 2014), and sex (Andres-Terre *et al.*, 2015; Cutolo *et al.*, 2004), they attend to different times. Forecasting susceptibility to IAV addresses the time prior to an IAV exposure, but IAV-infected hosts can be detectable only after the viral exposure (Andres-Terre *et al.*, 2015; Woods *et al.*, 2013; Zaas *et al.*, 2009).

It has been a success to detect IAV infection based on host gene expression. The peripheral blood gene expression signatures of post IAV infection different from those of bacterial and other respiratory viral infections can be used to detect symptomatic IAV infection (Andres-Terre *et al.*, 2015; Zaas *et al.*, 2009), which have been applied to the real-world cases of 2009 H1N1 pandemic with over 90% accuracy. In fact, comparing top 50 discriminative genes (Woods *et al.*, 2013), the infection signatures across different IAV sub-strains are so similar (Woods *et al.*, 2013) that some have exhibited cross-strain detectability (Chen *et al.*, 2011). Remarkably, a multi-cohort analysis further demonstrates that detecting symptomatic subjects infected with IAVs can count on only 11 genes derived from an influenza meta-signature (Andres-Terre *et al.*, 2015). Whether such success can occur in forecasting host susceptibility to IAV is yet to be unraveled.

To our knowledge, this is the first study to attempt the-state-of-the-arts deep learning to forecast host susceptibility to IAV prior to a viral exposure. In this study, we report the development of a deep learning model called DeepFlu.

2 Materials and Methods

2.1 Datasets Description

The datasets used in this study were the human peripheral blood gene expression data obtained from Gene Expression Omnibus (GEO) database. RMA normalization (Bullard *et al.*, 2010) was performed on each of the datasets. GEO GSE52428 (Woods *et al.*, 2013) which was done by the Affymetrix Human Genome U133A 2.0 array with 22,277 probes was the primary dataset for training and internal testing. 24 and 17 human subjects who aged 19 to 41 were inoculated intranasally with influenza A H1N1 (A/Brisbane/59/2007) or H3N2 (A/Wisconsin/67/2005), respectively. All subjects were healthy prior to the inoculation and monitored about every 8 hours over the course of 7 days. Consequently, 9 of the 24 H1N1 subjects and 8 of the 17 H3N2 subjects became symptomatic IAV-infected. 3 H1N1 subjects who developed symptoms but not infected by the H1N1 viruses were not classified as symptomatic IAV-infected. The rest were either asymptomatic or uninfected. Consequently, a total of 382 H1N1 and 267 H3N2 expression records were collected (Fig 1).

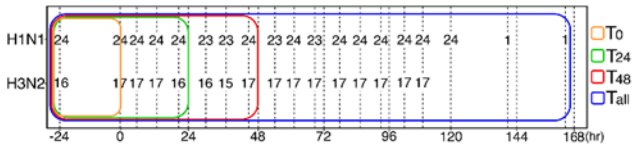

Fig 1. Datasets in four different time spans. T_0, T_{24}, T_{48}, and T_{all} represent the data taking from 1 day before influenza exposure up to immediately before exposure, 24h, 48h, and 168h after exposure, respectively. Each number in the dash line is the number of subjects at that time point.

Being symptomatic IAV-infected is defined as the presence of both clinical flu symptoms and microbiological infection. Technically speaking, a symptomatic infected subject should hold the modified Jackson score≧6 for at least 5 consecutive days and a positive viral culture or RT-PCR viral test (Boggild and McGeer, 2010) for at least 2 consecutive days. The modified Jackson score (Jackson and Dowling, 1959) assesses each of eight symptoms on a severity scale of 0 to 3, including sneeze, cough, malaise, runny nose, stuffy nose, sore throat, myalgia, and headache, to determine the presence or absence of flu symptoms. The viral culture or RT-PCR test is used to confirm the existence of IAV infection. Once the subject was identified as symptomatic IAV-infected, all of the subject's records were marked as symptomatic infected.

The external validation dataset came from GEO GSE73072 (Liu *et al.*, 2016). The experiments of DEE2 and DEE3 in this dataset were the same as those of GSE52428, but they were done by the Affymetrix Human Genome U133A 2.0 array with 12,023 genes instead. Therefore, the validation set were DEE4 and DEE5, which challenged another 19 and 21 subjects with the H1N1 and H3N2 viruses respectively.

2.2 Leave-one-person-out cross-validation

Due to the limited gene expression data, a leave-one-person-out (L1PO) cross-validation strategy on all subjects was applied to the primary dataset to evaluate how well the models perform for predicting whether a previously unknown subject would develop symptoms and become IAV-infected upon to viral exposure. This performance test involves a rotation of all individuals partitioning into one-person validation set and the remaining set. Models which were trained using the observations of the remaining set were tested on the validation set. Because of stochastic nature of the prediction models, we repeated the procedure 100 times to estimate the overall model performance.

2.3 DeepFlu Layout

Fig 2 illustrates the architecture of DeepFlu designed for predicting symptomatic influenza infection. DeepFlu is based on Deep Neural Network (DNN), also known as multilayer perceptron, which is a

feedforward neural network that can distinguish linear non-separable patterns (Cybenko, 1989). A series of model assessment was taken (See Supplement). DeepFlu is a 6-layer DNN, consisting four fully connected hidden layers each with 100 nodes, one input layer to extract the gene expression information, and one output layer to predict the presence of symptomatic flu infection.

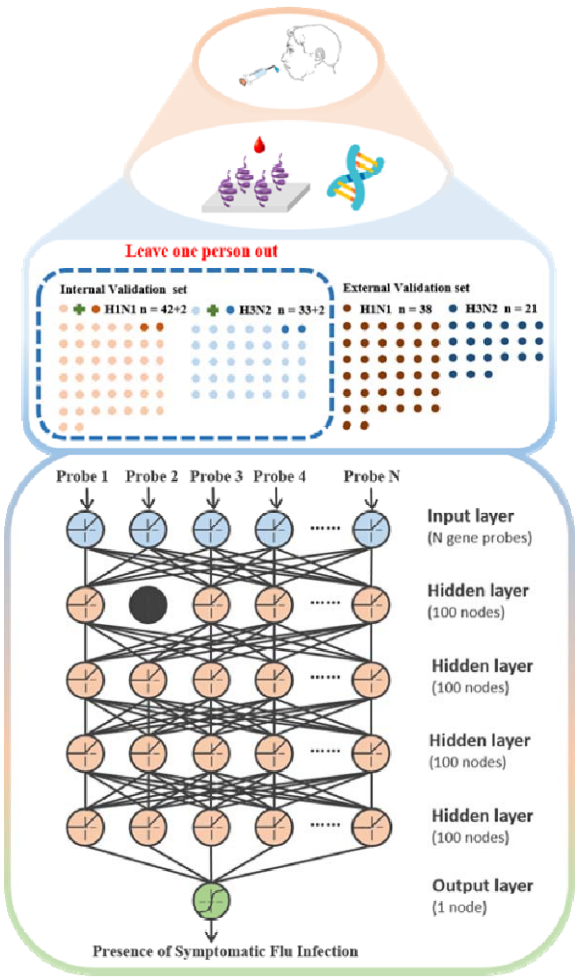

Fig. 2. DeepFlu Architecture. DeepFlu is a 6-layer DNN. The input layer consists of N nodes corresponding to the number of given gene expression probes. The four fully connected hidden layers consist of 100, 100, 100, and 100 nodes respectively. The black node in the hidden layer represents the dropout. The output layer utilizes single node to generate a binary outcome.

The following parameters and methods were used in DeepFlu. Both the input and hidden layers used a rectified linear unit (ReLU) function (Nair and Hinton, 2010) as the activation function, for it could overcome the vanishing gradient for nonnegative inputs. The output layer adopted a sigmoid function for a binary decision, whose outcome was strictly increasing and [0, 1]. To minimize the loss measured by binary cross-entropy, adaptive moment estimation (ADAM) optimization function (Kingma and Ba, 2014), which used the moving averages of the past gradient and squared gradient to update model weights, was applied during backpropagation, for ADAM could work effectively for high-dimensional parameter spaces. In addition, 150 training epochs were run with a learning rate of 0.001 without weight decay and 10% dropout as regularization was applied to the first hidden layer to avoid overfitting.

The performance of DeepFlu was evaluated by both L1PO cross-validation on the primary dataset and the external validation dataset.

2.4 Convolutional Neural Network

Convolutional neural networks (CNNs) (Zhang, 1988) are one of the most popular deep learning models, especially for image recognition. Many well-known image recognition models such as AlexNet (Iandola *et al.*, 2016), GoogLeNet (Szegedy *et al.*, 2015), an ResNet (He *et al.*, 2016) are based on CNNs. However, it remains unclear how CNNs perform in gene expression data.

We extended the 6-layer DeepFlu to include two convolutional layers and two pooling layers to form a 10-layer CNN model. CNNs, which can acquire spatial relations of data, can be divided into two sections. The first section is to extract features through alternate convolutional filters and pooling layers, and the second section is to learn through fully connected layers. Moreover, the activation function of the convolutional layers was a ReLU function (Nair and Hinton, 2010) and the filter of the pooling layers was a max pooling function. The rest of parameters and methods used in the CNN model were adopted from those of DeepFlu. Because the CNN model had more layers and parameters, the CNN model would have more epochs than DeepFlu.

The CNN models can be described simply as follows: the input is a tensor X ∈ , where *w*, *h*, and *d* are the image width, height, and depth, respectively, and the depth refers to the number of channels. The convolution transforms the input X using *m* filter tensors K ∈ with *q* zero-paddings and *s* strides into *m* feature maps M ∈ , where *n* is the order of filter kernels, *d* keeps the same number of channels as the input, and *m* which determines the number of the output channels is the number of filter kernels. To effectively reduce the dimensions of feature maps, the pooling layer adopts a filter kernel. The pooling layer transforms the feature map using a pooling filter tensor P ∈ into a pooled feature map, where *q* is the order of pooling kernels and *m* is the channel number of feature maps. Excluding the convolutional and pooling layers, the makeup of the CNN models used in this study is the same as that of DeepFlu.

2.5 Random Forest

Random Forest (RF) is a classification method using an ensemble of bagged decision trees with randomly selected features (Ho, 1995). The RF algorithm embeds the concept of random subset feature selection into bagged decision trees which are trained on random samples with replacement to avoid overfitting. To determine the class of the predicted case, each decision tree which is best split by a small fraction of randomly selected features casts a vote independently under majority rule. In this work, the parameters and methods used for the RF models were adopted from the Chang study (Chang and Yang, 2013). Each RF classifier had 1,000 decision trees. Gini impurity (Ellerman, 2018) was used to measure the quality of a split by a selected feature and when looking for a best split, the number of tried features was set to be square root of the total features (Breiman, 2001).

2.6 Support Vector Machine

SVM is also a classification method that searches an optimal hyperplane to separate two classes with maximum margin (Cortes and Vapnik, 1995). In this work, a radial basis function kernel (Vert and Tsuda, 2004) was used for our SVM models, whose cost and gamma parameters

(Chang and Lin, 2011) for the kernel were set as 1 and one over the number of features.

2.7 Performance Comparison

The performances of the predictive models were measured by sensitivity, specificity, precision, accuracy, an area under the receiver operating characteristic curve (AUROC), and an area under the precision recall curve (AUPR). These evaluation measurements were calculated as follows:

$$\text{Sensitivity (recall)} = \frac{TP}{TP+FN} \quad (1)$$

$$\text{Specificity} = \frac{TN}{TN+FP} \quad (2)$$

$$\text{Precision} = \frac{TP}{TP+FP} \quad (3)$$

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN} \quad (4)$$

where TP, TN, FP, and FN represent true positives, true negatives, false positives, and false negative, respectively. Sensitivity, specificity, precision, and accuracy are the percentage of the correct predictions on positive data, negative data, predicted positive data, and all the data, respectively.

Like accuracy, AUROC and AUPR are used to evaluate the overall performance of a binary classifier (Davis and Goadrich, 2006; Fawcett, 2006). AUROC and AUPR are based on dynamic positive-negative prediction thresholds instead of a static threshold like accuracy. AUROC is the area under the curve which plots true positive rates (sensitivity) against false positive rates using different prediction thresholds; similar to AUROC, AUPR is the area under the curve by plotting precision against recall rates (sensitivity). AUROC and AUPR are in the range [0,1]. The higher AUROC or AUPR, the better the classifier. A perfect classifier has AUROC or AUPR equal to 1. However, unlike AUROC, AUPR only focuses on positive cases. The AUROC value tends to be overly optimistic for an imbalance dataset, but AUPR does not (Yu and Park, 2014).

3 Results

3.1 Best Learning Method for the Gene Expression Data

By randomly dividing the influenza gene expression data into 80% training and 20% testing data, Table 1 shows the performances of four different predictive modeling approaches using the full 22,277 array probes. The setting of the training and test data used here, which were not exclusive to those prior to exposure, differed from most of the experiments throughout this study. All the four models, CNN, DNN, RF, and SVM, demonstrate that the special gene expression signals of the flu infected subjects can be learned. Among the four models, DNN generally performed the best, reaching averagely 96.1% accuracy, 0.957 AUROC, and 0.963 AUPR for H1N1; 95.4% accuracy, 0.954 AUROC, and 0.965 AUPR for H3N2. Thus, DNN was chosen as the backbone of DeepFlu for the following development.

3.2 Time Spans on Prediction Performance

Based on the leave-one-person-out (L1PO) cross-validation, Table 2 compares the prediction performances of DeepFlu trained under different time spans. T_0, T_{24}, T_{48} and T_{all} represent the expression data taking from the beginning (one day before exposure) to the moment of exposure, one day after exposure, two days after exposure, and one week after exposure, respectively (Fig 1). In other words, T_0 contains only the expression data without viral exposure. The results show that training by the data prior to viral exposure (T_0) was the best for DeepFlu to predict whether one would catch the flu and develop symptoms after an exposure, achieving 70.0% accuracy, 0.787 AUROC, and 0.758 AUPR for H1N1 and 73.8% accuracy, 0.847 AUROC, and 0.901 AUPR for H3N2 (Fig. 3A-B).

Table 1. Model comparison for distinguishing symptomatic flu-infected subjects

Method	Accuracy	Sensitivity	Specificity	Precision	AUROC	AUPR
H1N1						
CNN	0.955	**0.951**	0.960	0.929	0.955	0.949
DNN	**0.961**	0.936	**0.978**	**0.964**	**0.957**	0.963
RF	0.911	0.835	0.960	0.927	0.898	**0.969**
SVM	0.841	0.639	0.891	0.881	0.789	0.859
H3N2						
CNN	0.937	0.962	0.914	0.940	0.946	0.951
DNN	**0.954**	**0.975**	**0.932**	**0.942**	**0.954**	0.965
RF	0.924	0.930	0.921	0.927	0.925	**0.984**
SVM	0.867	0.851	0.888	0.898	0.866	0.913

AUROC: area under the receiver operating characteristic curve; AUPR: area under the precision-recall curve; CNN: convolutional neural network; DNN: deep neural network; RF: random forest; SVM: support vector machine

Table 2. Performance comparison of DeepFlu trained under different time spans by L1PO cross-validation

Span	Accuracy	Sensitivity	Specificity	Precision	AUROC	AUPR
H1N1						
$\mathbf{T_{All}}$	0.637	0.556	0.668	0.618	0.661	0.638
$\mathbf{T_{48}}$	0.658	0.602	0.689	0.662	0.713	0.701
$\mathbf{T_{24}}$	0.669	0.613	0.715	0.679	0.725	0.712
$\mathbf{T_0}$	**0.700**	**0.616**	**0.822**	**0.718**	**0.787**	**0.758**
H3N2						
$\mathbf{T_{All}}$	0.638	0.707	0.548	0.649	0.668	0.688
$\mathbf{T_{48}}$	0.661	0.683	0.635	0.688	0.679	0.744
$\mathbf{T_{24}}$	0.689	0.689	0.723	0.759	0.759	0.806
$\mathbf{T_0}$	**0.738**	**0.722**	**0.756**	**0.770**	**0.847**	**0.901**

AUROC: area under the receiver operating characteristic curve; AUPR: area under the precision-recall curve; L1PO: leave one person out

3.3 Feature Sets on Prediction Performance

Based on the L1PO cross-validation test focusing on host gene expression prior to exposure (T_0), Table 3 compares the performances of the four predictive models using three different feature sets for forecasting symptomatic flu-infected subjects. Besides DeepFlu, the three other predictive models are CNN, RF, and SVM; the three feature sets are the complete probe set of the microarray (22,277 features), the Hallmark genes (Liberzon *et al.*, 2015) of MSigDB (4,164 corresponding features), and the top discriminative genes for distinguishing IAV post infection (64 and 61 corresponding features for H1N1 and H3N2 respectively) (Woods *et al.*, 2013). The results indicate that the prediction models using more features generally performed better and DeepFlu achieved the overall best performance using the full 22,277 features for both

H1N1 and H3N2. For H1N1, DeepFlu had 70.0% accuracy, 0.787 AUROC, and 0.758 AUPR; for H3N2, it reached 73.8% accuracy, 0.849 AUROC, and 0.901 AUPR.

Compared to using the full 22,277 features, DeepFlu gave a comparable but weaker performance by using 4,164 features. Still it outperformed the other three models.

For using the top discriminative genes of IAV post infection, all the four models performed poorly. All the four models using the limited gene set gave the worst AUROC and AUPR values for H1N1. For H3N2, the models performed slightly better than a random guess.

Table 3. Performance comparison using different feature sets trained under T_0 by L1PO cross-validation

Feature #	Method	Accuracy	Sensitivity	Specificity	Precision	AUROC	AUPR
H1N1							
64	**CNN**	**0.450**	**0.647**	0.269	**0.443**	**0.495**	**0.432**
64	**DeepFlu**	0.336	0.430	0.245	0.341	0.298	0.364
64	**RF**	0.293	0.268	**0.319**	0.281	0.165	0.337
64	**SVM**	0.135	0.120	0.151	0.120	0.083	0.319
*4225	**CNN**	0.651	0.629	**0.665**	**0.646**	0.723	0.633
4164	**DeepFlu**	**0.691**	**0.835**	0.573	0.642	**0.762**	**0.732**
4164	**RF**	0.513	0.498	0.529	0.514	0.517	0.466
4164	**SVM**	0.620	0.640	0.600	0.615	0.631	0.564
*22500	**CNN**	0.588	0.470	0.664	0.486	0.636	0.624
22277	**DeepFlu**	**0.700**	**0.616**	**0.822**	**0.718**	**0.787**	**0.758**
22277	**RF**	0.519	0.495	0.543	0.520	0.545	0.482
22277	**SVM**	0.585	0.586	0.585	0.617	0.627	0.604
H3N2							
64	**CNN**	0.548	0.628	0.468	0.572	**0.563**	**0.611**
61	**DeepFlu**	0.486	0.461	0.494	0.565	0.422	0.492
61	**RF**	**0.615**	**0.689**	**0.506**	**0.670**	0.553	0.546
61	**SVM**	0.528	0.574	0.476	0.552	0.522	0.524
*4225	**CNN**	0.550	0.550	0.550	0.581	0.556	0.534
4164	**DeepFlu**	0.653	0.615	0.696	0.696	**0.766**	**0.821**
4164	**RF**	**0.690**	0.602	**0.818**	**0.828**	0.712	0.719
4164	**SVM**	0.624	**0.794**	0.432	0.612	0.723	0.760
*22500	**CNN**	0.705	**0.739**	0.667	0.715	0.812	0.863
22277	**DeepFlu**	**0.738**	0.722	**0.756**	0.770	**0.847**	**0.901**
22277	**RF**	0.730	0.723	0.742	**0.803**	0.782	0.808
22277	**SVM**	0.702	0.655	0.755	0.750	0.739	0.757

AUROC: area under the receiver operating characteristic curve; AUPR: area under the precision-recall curve; CNN: convolutional neural network; L1PO: leave one person out; RF: random forest; SVM: support vector machine. *Zero paddings are utilized to satisfy the requirement of CNN

3.4 Mixed Training Data on Prediction Performance

Because the post infection signatures of H1N1 and H3N2 were similar, sharing up to 90% of top 50 discriminative genes (Woods *et al.*, 2013), and the H3N2 signature could be used to detect H1N1 infection (Chen *et al.*, 2011), it was speculated that extending the training set by mixed data of H1N1 and H3N2 might enhance prediction performance. Instead of training by the expression data for the same IAV strain as seen in Table 3, Table 4 compares the performance of the L1PO cross validation of the predictive models trained by the mixed T0 data of H1N1 and H3N2. Table 4 shows that despite a drop in performance, DeepFlu is still the best for forecasting symptomatic influenza A-infected subjects prior to viral exposure. For H1N1, DeepFlu using the full 22,277 features of the mixed training data achieved 72.1% accuracy, 0.752 AUROC, and 0.676 AUPR, not better than that trained only by the H1N1 data with 70.0% accuracy, 0.787 AUROC, and 0.758 AUPR. Similarly, as for H3N2, DeepFlu using the mixed data dropped performance than using the H3N2 data alone, but retained a high performance, keeping more than 0.790 AUROC and AUPR. The results indicate that the mixed H1N1 and H3N2 training data neither improved identifying who were susceptible to H1N1 nor those to H3N2 infection.

Table 4. Performance comparison of DeepFlu trained using different feature sets by mixed H1N1 and H3N2 T_0 data by L1PO cross-validation

Feature #	Accuracy	Sensitivity	Specificity	Precision	AUROC	AUPR
H1N1						
64	0.529	0.545	0.517	0.427	0.488	0.389
4164	0.682	**0.787**	0.610	0.574	0.751	0.660
22277	**0.721**	0.625	**0.782**	**0.656**	**0.752**	**0.676**
H3N2						
61	0.688	0.611	**0.775**	0.759	0.522	0.652
4164	0.696	**0.785**	0.596	0.687	0.748	0.772
22277	**0.707**	0.650	0.771	**0.764**	**0.796**	**0.855**

AUROC: area under the receiver operating characteristic curve; AUPR: area under the precision-recall curve; L1PO: leave one person out

3.5 External Validation on DeepFlu

To examine the robustness of DeepFlu, an independent validation set of 38 H1N1 and 21 H3N2 records from 19 H1N1 and 21 H3N2 subjects were used. Table 5 shows the performance of DeepFlu on the external validation data. Here DeepFlu achieved 71.4% accuracy, 0.700 AUROC, and 0.723 AUPR for H1N1; 73.5% accuracy, 0.732 AUROC, and 0.749 AUPR for H3N2 (Fig. 3C-D).

Table 5. External validation of DeepFlu

	Accuracy	Sensitivity	Specificity	Precision	AUROC	AUPR
H1N1	0.714	0.733	0.767	0.700	0.700	0.723
H3N2	0.735	0.714	0.750	0.667	0.732	0.749

AUROC: area under the receiver operating characteristic curve; AUPR: area under the precision-recall curve

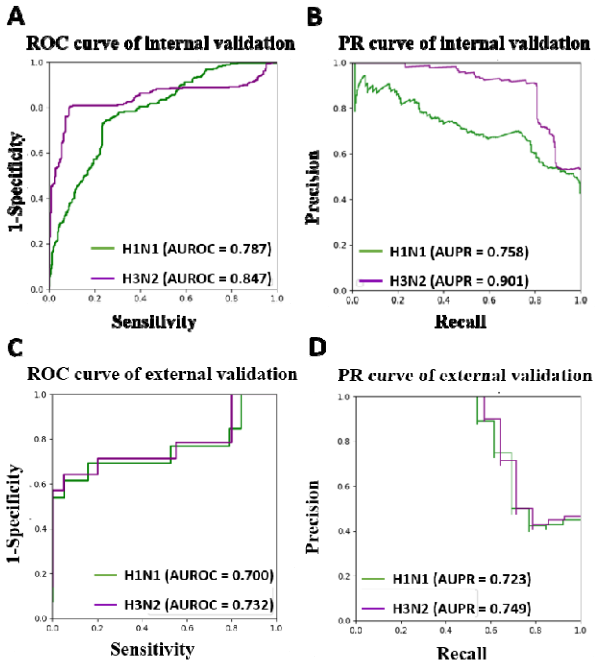

Fig. 3. Receiver operating characteristic curve and precision-recall curves of DeepFlu. (A, B) Internal validation by leave-one-person-out cross-validation (C, D) Independent external validation.

4 Discussion

According to the L1PO cross-validation and independent test results, we have demonstrated that it is possible to forecast whether an individual would catch the flu and develop symptoms immediately after an exposure to IAV based on pre-IAV-exposure host gene expression. The results are consistent with the previous finding on the predictability of RSV susceptibility (Fourati *et al.*, 2018)—both studies demonstrate the possibility of forecasting susceptibility of respiratory viral infections.

We have also demonstrated that deep learning can work with small dataset too. The gene expression dataset with 382 H1N1 samples of 24 subjects and 267 H3N2 samples of 17 persons used in this study is considered a small dataset, compared with the traditional big data. To avoid overfitting with the small training data, we added dropout and did not build overly complex models. Besides, we have validated our deep learning model using the external independent test. DeepFlu could handle this particular 'large *p* small *n*' problem well. For the independent test, DeepFlu achieved 71.4% accuracy, an AUROC of 0.700, and an AUPR of 0.723 on the H1N1 data, and 73.5% accuracy, an AUROC of 0.732, and an AUPR of 0.749 on the external H3N2 data. To our knowledge, this is the first study to demonstrate that deep learning can work for such forecast.

Feature selection is important for predicting IAV susceptibility. Our results generally show that the larger the feature size, the better the prediction performance. In fact, DeepFlu with the complete 22,274 features is the optimal model. The DNN model which used the Hallmark genes of MSigDB, about only one-fifth of the complete features, could still reach a comparable performance, indicating that there must exist a limited number of candidate genes closely related to one's susceptibility to IAV. However, the underlying biological mechanism might be too complex to learn well with the selected features. Instead entire features became a better alternative.

The top 50 discriminative post-IAV-infection genes are inadequate to be used as features for this problem. The models using the limited number of IAV infection genes all consistently performed poorly compared with the others. This poor result suggests that the IAV infection genes, which are stimulated after viral exposure, may significantly differ from the genes relating to one's susceptibility to IAV, which may regulate and control subsequent immune responses. Furthermore, the biological mechanisms behind IAV post infection, such as inflammatory responses (Iwasaki and Pillai, 2014), may differ from those behind IAV immunity.

The results also indicate that DeepFlu performs better for recognizing gene expression patterns of IAV immunity than the other three models, including the CNN models. DeepFlu outperformed the CNN model in both random setting (80% training and 20% test data) and L1PO setting, perhaps because gene expression profiles did not conform to the CNN properties—neither subsampling gene chips would hold the original meaning nor the same expression patterns could happen in different regions of gene chips.

Why forecasting H3N2 susceptibility works better than forecasting H1N1 susceptibility is unsettled. We suspect the length of latency period may be a factor, for the maximal symptoms of H3N2 and H1N1 patients peak at day two and day four, respectively (Fourati *et al.*, 2018). The H3N2 models foresee clearer, perhaps because T_0 is closer to the symptom onset of H3N2 than that of H1N1.

Moreover, using which time span to train is also important. The results support that the best outcome happened when both the training and test data came from the same time span—prior to exposure (T_0). Having other time spans might bring in other signals or noises, thus lowering the performance.

The performance of forecasting may still have room to improve. However, by mixing H1N1 and H3N2 records to expend the training data may not be the way to enhance prediction performance. Using the mixed data, the performance of DeepFlu dropped for identifying the susceptible to H1N1 or H3N2, suggesting there may exist distinct differences in susceptibility to different IAV strains.

There are some limitations in this study. First, limited influenza gene expression data are available. Although DeepFlu excels even with limited gene expression data, the deep learning model may improve further with more training data. Second, it is unknown if the subjects are immunologically naïve to IAV. The conditions in the gene expression experiments had been well controlled, but a rigorous setting may be needed for more proper comparisons. Third, the biological mechanisms behind IAV susceptibility or immunity remain unclear. Although the development of RSV infection has linked the presence of heme metabolism of hosts using an enrichment analysis of the target gene set (Fourati *et al.*, 2018), DeepFlu with full features, which does not rely on specific gene sets, may require more subtle approaches to study the mechanisms behind IAV susceptibility or immunity.

5 Conclusion

Gene expression can portend IAV susceptibility and deep learning can catch the expression signals to identify people at risk for the flu and alarm them in advance to reduce hospitalization and mortality caused by IAV.

Acknowledgements

We thank the groups of G. Ginsburg and A. Hero for making these microarray data publicly available.

Funding

This work has been supported by Ministry of Science and Technology, R.O.C. under MOST-108-2221-E-019-052 (https://www.most.gov.tw). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Conflict of Interest: none declared.

References

Andres-Terre, M., *et al.* (2015) Integrated, Multi-cohort Analysis Identifies Conserved Transcriptional Signatures across Multiple Respiratory Viruses. *Immunity*, **43**, 1199-1211.

Boggild, A.K. and McGeer, A.J. (2010) Laboratory diagnosis of 2009 H1N1 influenza A virus. *Critical Care Medicine*, **38**.

Breiman, L. (2001) Random Forests. *Machine Learning*, **45**, 5-32.

Bullard, J.H., *et al.* (2010) Evaluation of statistical methods for normalization and differential expression in mRNA-Seq experiments. *BMC Bioinformatics*, **11**, 94.

Casanova, J.L. and Abel, L. (2004) The human model: a genetic dissection of immunity to infection in natural conditions. *Nat Rev Immunol*, **4**, 55-66.

Cassini, A., *et al.* (2018) Impact of infectious diseases on population health using incidence-based disability-adjusted life years (DALYs): results from the Burden of Communicable Diseases in Europe study, European Union and European Economic Area countries, 2009 to 2013. *Euro Surveill*, **23**.

Chang, C.-C. and Lin, C.-J. (2011) Libsvm. *ACM Transactions on Intelligent Systems and Technology*, **2**, 1-27.

Chang, C.-C. and Lin, C.-J. (2011) LIBSVM: A library for support vector machines. *ACM Trans. Intell. Syst. Technol.*, **2**, Article 27.

Chang, K.Y. and Yang, J.R. (2013) Analysis and prediction of highly effective antiviral peptides based on random forests. *PLoS One*, **8**, e70166.

Chen, M., *et al.* (2011) Detection of viruses via statistical gene expression analysis. *IEEE Trans Biomed Eng*, **58**, 468-479.

Cortes, C. and Vapnik, V. (1995) Support-vector networks. *Machine Learning*, **20**, 273-297.

Cutolo, M., *et al.* (2004) Sex hormones influence on the immune system: basic and clinical aspects in autoimmunity. *Lupus*, **13**, 635-638.

Cybenko, G. (1989) Approximation by superpositions of a sigmoidal function. *Mathematics of Control, Signals and Systems*, **2**, 303-314.

Davis, J. and Goadrich, M. (2006) The relationship between Precision-Recall and ROC curves. *Proceedings of the 23rd international conference on Machine learning*, 233–240.

Dopico, X.C., *et al.* (2015) Widespread seasonal gene expression reveals annual differences in human immunity and physiology. *Nat Commun*, **6**, 7000.

Ellerman, D. (2018) Logical Entropy: Introduction to Classical and Quantum Logical Information Theory. *Entropy (Basel)*, **20**, 679.

Fawcett, T. (2006) An introduction to ROC analysis. *Pattern recognition letters*, **27**, 861-874.

Fourati, S., *et al.* (2018) A crowdsourced analysis to identify ab initio molecular signatures predictive of susceptibility to viral infection. *Nat Commun*, **9**, 4418.

Gordon, A. and Reingold, A. (2018) The Burden of Influenza: a Complex Problem. *Curr Epidemiol Rep*, **5**, 1-9.

He, K., *et al.* (2016) Deep Residual Learning for Image Recognition. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 770-778.

Heikkinen, T., Ikonen, N. and Ziegler, T. (2014) Impact of influenza B lineage-level mismatch between trivalent seasonal influenza vaccines and circulating viruses, 1999-2012. *Clin Infect Dis*, **59**, 1519-1524.

Ho, T.K. (1995) Random decision forests. *Proceedings of 3rd international conference on document analysis and recognition*, **1**, 278-282.

Houser, K. and Subbarao, K. (2015) Influenza vaccines: challenges and solutions. *Cell Host Microbe*, **17**, 295-300.

Iandola, F.N., *et al.* (2016) SqueezeNet: AlexNet-level accuracy with 50x fewer parameters and< 0.5 MB model size. *arXiv preprint arXiv:1602.07360*.

Iuliano, A.D., *et al.* (2018) Estimates of global seasonal influenza-associated respiratory mortality: a modelling study. *Lancet*, **391**, 1285-1300.

Iwasaki, A. and Pillai, P.S. (2014) Innate immunity to influenza virus infection. *Nat Rev Immunol*, **14**, 315-328.

Jackson, G.G. and Dowling, H.F. (1959) Transmission of the common cold to volunteers under controlled conditions. IV. Specific immunity to the common cold. *J Clin Invest*, **38**, 762-769.

Jayasundara, K., *et al.* (2014) Natural attack rate of influenza in unvaccinated children and adults: a meta-regression analysis. *BMC Infect Dis*, **14**, 670.

Kingma, D.P. and Ba, J. (2014) Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Liberzon, A., *et al.* (2015) The Molecular Signatures Database (MSigDB) hallmark gene set collection. *Cell Syst*, **1**, 417-425.

Liu, T.Y., *et al.* (2016) An individualized predictor of health and disease using paired reference and target samples. *BMC Bioinformatics*, **17**, 47.

Nair, V. and Hinton, G.E. (2010) Rectified linear units improve restricted boltzmann machines. *Proceedings of the 27th International Conference on International Conference on Machine Learning*, 807–814.

Nakaya, H.I., *et al.* (2011) Systems biology of vaccination for seasonal influenza in humans. *Nat Immunol*, **12**, 786-795.

Park, J.E. and Ryu, Y. (2018) Transmissibility and severity of influenza virus by subtype. *Infect Genet Evol*, **65**, 288-292.

Paules, C. and Subbarao, K. (2017) Influenza. *Lancet*, **390**, 697-708.

Poon, L.L., *et al.* (2016) Quantifying influenza virus diversity and transmission in humans. *Nat Genet*, **48**, 195-200.

Putri, W., *et al.* (2018) Economic burden of seasonal influenza in the United States. *Vaccine*, **36**, 3960-3966.

Qi, Q., *et al.* (2014) Diversity and clonal selection in the human T-cell repertoire. *Proc Natl Acad Sci U S A*, **111**, 13139-13144.

Szegedy, C., *et al.* (2015) Going deeper with convolutions. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 1-9.

Vert, J.-P. and Tsuda, K. (2004) A primer on kernel methods. *Kernel methods in computational biology*, **47**, 35-70.

WHO. (2018) Influenza (Seasonal) Facts. Retrieved from http://www.who.int/news-room/fact-sheets/detail/influenza-(seasonal)

Woods, C.W., *et al.* (2013) A host transcriptional signature for presymptomatic detection of infection in humans exposed to influenza H1N1 or H3N2. *PLoS One*, **8**, e52198.

Yu, W. and Park, T. (2014) AucPR: an AUC-based approach using penalized regression for disease prediction with high-dimensional omics data. *BMC Genomics*, **15 Suppl 10**, S1.

Zaas, A.K., *et al.* (2009) Gene expression signatures diagnose influenza and other symptomatic respiratory viral infections in humans. *Cell Host Microbe*, **6**, 207-217.

Zhang, A., *et al.* (2019) Original Antigenic Sin: How First Exposure Shapes Lifelong Anti-Influenza Virus Immune Responses. *J Immunol*, **202**, 335-340.

Zhang, W. (1988) Shift-invariant pattern recognition neural network and its optical architecture. *Proceedings of annual conference of the Japan Society of Applied Physics*.