Using Machine Learning to Distinguish Infected from Non-Infected Subjects at an Early Stage Based on Viral Inoculation

Ghanshyam Verma[1,2], Alokkumar Jha[1,2], Dietrich Rebholz-Schuhmann[1,2,3], and Michael G. Madden[1,2]

[1] Insight Centre for Data Analytics, National University of Ireland Galway, Ireland
[2] School of Computer Science, National University of Ireland Galway, Ireland
[3] ZB MED - Information Center for Life Sciences, University of Cologne, Germany
{ghanshyam.verma, alokkumar.jha, rebholz}@insight-centre.org,
michael.madden@nuigalway.ie

Abstract. Gene expression profiles help to capture the functional state in the body and to determine dysfunctional conditions in individuals. In principle, respiratory and other viral infections can be judged from blood samples; however, it has not yet been determined which genetic expression levels are predictive, in particular for the early transition states of the disease onset. For these reasons, we analyse the expression levels of infected and non-infected individuals to determine genes (potential biomarkers) which are active during the progression of the disease. We use machine learning (ML) classification algorithms to determine the state of respiratory viral infections in humans exploiting time-dependent gene expression measurements; the study comprises four respiratory viruses (H1N1, H3N2, RSV, and HRV), seven distinct clinical studies and 104 healthy test candidates involved overall. From the overall set of 12,023 genes, we identified the 10 top-ranked genes which proved to be most discriminatory with regards to prediction of the infection state. Our two models focus on the time stamp nearest to $t = 48$ hours and nearest to $t =$ "*Onset Time*" denoting the symptom onset (at different time points) according to the candidate's specific immune system response to the viral infection. We evaluated algorithms including k-Nearest Neighbour (k-NN), Random Forest, linear Support Vector Machine (SVM), and SVM with radial basis function (RBF) kernel, in order to classify whether the gene expression sample collected at early time point t is infected or not infected. The "*Onset Time*" appears to play a vital role in prediction and identification of ten most discriminatory genes.

Keywords: Machine learning · Respiratory viral infection · Prediction · Deferentially expressed genes.

1 Introduction

Respiratory viral infections are common diseases which are caused by a wide range of viruses, e.g., H1N1, H3N2, RSV and HRV, affecting the respiratory tract. While patients usually recover in a short period of time without any treatment, respiratory viral infections can lead to severe outcomes among individuals with other aggravating primary diseases, in particular, when these are deleterious to the function of the respiratory system. Such severe cases may increase the likelihood of death in elderly or immuno-compromised individuals [14]. Moreover, each influenza epidemic leads to an increase in healthcare costs through excessive hospitalizations apart from the need for substantial amounts of vaccines, and the spread of respiratory virus diseases affect all age groups and thus can lead to periodic epidemics [25]. Overall, the early identification of respiratory viral infections could be useful as a means to reduce large-scale outbreaks and periodic epidemics as well as achieving early intervention for individual patients [13].

In this paper, we investigated the changes in gene expression distinguishing infected individuals from non-infected ones. We use different ML methods to determine the most predictive changes comparing samples from healthy and infected individuals, using public data collected in seven different studies involving healthy individuals before and after inoculation of the viruses. This data (gene expression only) – generated from these seven challenge studies – has been released in 2016 and is available on Gene Expression Omnibus (GEO). In 2017, the label information (non-infected vs. infected) associated with this dataset also had been made available for open access to all. We use this label information as a ground-truth for labeling the whole data. ML solutions form a vital role in the identification of specific patterns, and subsequent functional annotation of the identified genes can explain the causality behind the exposed patterns. Gene expression changes often happen due to some regulatory markers, while other genes behave as housekeeping genes. Therefore, identification of relevant patterns and responsible regulatory markers at consistent time points should yields credible biomarkers in such cases. In this work we identify top ten such biomarkers which are found to be highly contributing in progression of respiratory viral infections at an early stage. The labeled data with code and build ML models are available here: https://github.com/GhanshyamVerma/DILS_2018.

2 The Respiratory Viral Data Sets

We conducted ML experiments on the data collected from 7 Respiratory Viral Challenge studies which is available for open access on GEO (accession number GSE73072)[4]. These respiratory viral challenge studies consist of a total of 151 human volunteers, each of whom was exposed to one of 4 viruses, summarised in Table 1 [12].

In Table 1, the first column represents the sub-study designation, the second column denotes the type of virus used in the challenge, the third and the fourth columns represent the year and the location of the conducted sub-study, respectively, the fifth column represents the DUHS IRB protocol number, the sixth column represents the duration of the sub-study in hours and the last two

[4] https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE73072.

Table 1. Details of the data collected in the seven respiratory virus challenge studies [12]

Challenge	Virus	Year	Location	IRB Protocol	Duration (hrs)	#Subjects	#Time-points
DEE1	RSV	2008	Retroscreen	Pro00002796	166	20	21
DEE2	H3N2	2009	Retroscreen	Pro00006750	166	17	21
DEE3	H1N1	2009	Retroscreen	Pro00018132	166	24	20
DEE4	H1N1	2010	Retroscreen	Pro00019238	166	19	21
DEE5	H3N2	2011	Retroscreen	Pro00029521	680	21	23
HRV UVA	HRV	2008	UoVirginia	Pro00003477	120	20	15
HRV Duke	HRV	2010	Duke Univ.	Pro00022448	136	30	19

columns denote the number of subjects and the number of time-points collected per subject, respectively [12].

All the participants were healthy when they enrolled for the study. After enrolment in the study, all subjects were inoculated with one of the 4 viruses (H1N1, H3N2, HRV, RSV). Their blood samples were taken at different pre-defined time-points, thus delivering samples from non-infected individuals as well as from infected ones. The samples from non-infected individuals were taken at two time-points before the inoculation of the virus, as shown in Fig. 1 (inspired by a figure by Liu et al. [12]). All the subjects were exposed to the virus immediately after taking the healthy blood sample (at time-point 0). During each study, blood samples were taken for twice before the inoculation of virus and at various time stamps after the inoculation of virus. The whole blood gene expression data was obtained using Affymetrix Human U133A 2.0 GeneChips. Additional details can be found on GEO (accession number GSE73072).

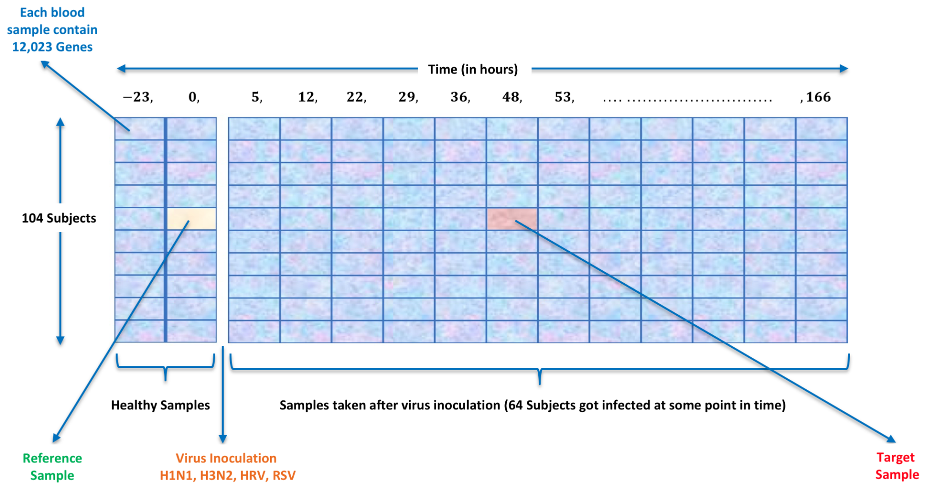

Each blood sample contain 12,023 Genes
Time (in hours)
−23, 0, 5, 12, 22, 29, 36, 48, 53, …. …………………….. ,166
104 Subjects
Healthy Samples
Samples taken after virus inoculation (64 Subjects got infected at some point in time)
Reference Sample
Virus Inoculation H1N1, H3N2, HRV, RSV
Target Sample

Fig. 1. Layout describing characteristics of the data. Every cell depicting blood sample taken at some point of time during the whole study and contains gene expression values of 12,023 human genes.

From the start, 151 subjects were enrolled in the 7 challenge studies, however, we had to exclude 47 subjects from the study. Among those 47 subjects, 44 subjects had inconsistencies between their declared symptomatic status and the measured shedding status (see Table 2). These 44 clinically ambiguous subjects were at some time either acutely infected but remained asymptomatic or were not infected but did turn acutely symptomatic [12], therefore, we must conclude that these 44 subjects data is inconsistent (faulty). We cannot draw any conclusions from faulty data. Moreover, faulty data can be misleading and harmful while model building. Apart from these 44 subjects, three more subjects have been excluded because there is no Affymetrix data available for them (subjects 6, 9 and 21 from the HRV Duke university sub study). We have identified those 47 ambiguous subjects whose data is faulty, removed them and the unambiguous labeled data with code and build ML models can be accessed using a link provided in the Introduction section.

Table 2. Detail of the ambiguous subjects those excluded due to inconsistencies between their declared symptomatic status and measured shedding status.

Sr. No.	**Challenge**	**Subject IDs (Ambiguous subjects)**	**Total (Ambiguous subjects)**
1	DEE1	13, 15, 16	3
2	DEE2	2, 4	2
3	DEE3	1, 2, 5, 7, 11, 15, 18, 21, 23	9
4	DEE4	5, 7, 8, 9, 10, 11, 12, 13, 17, 19	10
5	DEE5	3, 7, 15, 16, 17	5
6	UoVirginia	1, 10, 12, 17	4
7	Duke Univ.	3, 10, 11, 15, 18, 20, 25, 27, 28, 29, 30	11
			44

3 Experimental Design

The overall goal of our study is to analyze the ability of different ML algorithms to predict the state of health and disease soon after the disease onset time. As we do not have always data for each subject exactly the times we are interested in, we have taken nearest available time-point. We make the assumption that at $t = 0$, just before the inoculation, all the candidates are not sick and at $t = 48h$ (approximately) or at the onset time, the effect of virus inoculation should be visible, and thus exposed in the gene expression data. After first mining for expression patterns, we are also interested in finding the important genes/biomarkers which are highly likely to contribute to the progression of the respiratory viral infection.

After excluding the 47 ambiguous candidates, we were left with total 104 candidates, all of whom were healthy at $t = 0$. Out of these 104 healthy subjects, 64 became sick at some point of time after inoculation of the virus and the other 40 remained healthy during the whole study. There is no onset time for these 40 non-infected subjects, therefore, we took the average for the available onset values, which was 55.01 hours after inoculation.

Our main focus is on the gene expression levels when comparing the 40 subjects who did not become infected after inoculation with the 64 who did. We designed 4 different experiments: for each experiment we made different use of the number of subjects that got infected after inoculation and of the time-points (48 hours vs. onset time).

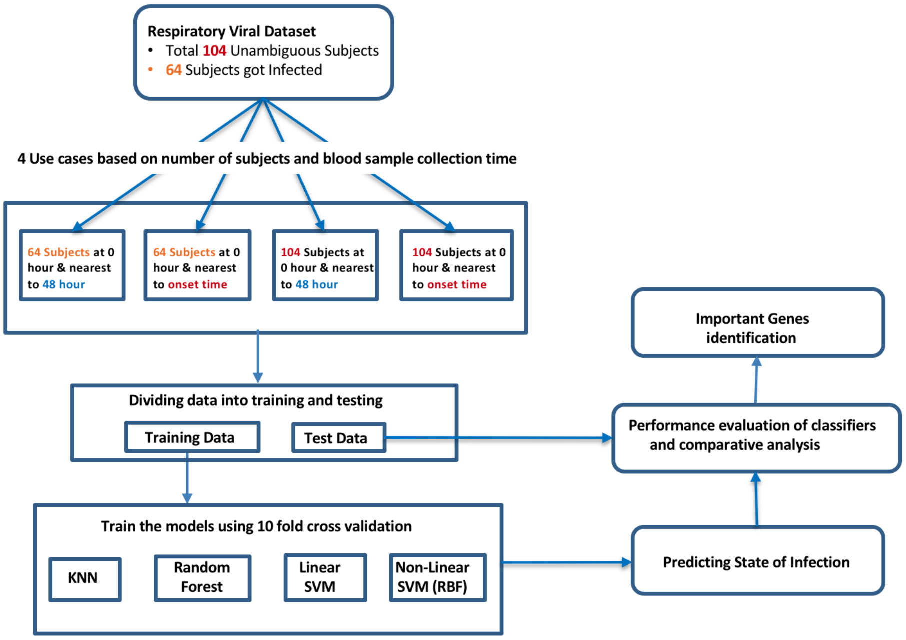

Respiratory Viral Dataset
• Total 104 Unambiguous Subjects
• 64 Subjects got Infected
4 Use cases based on number of subjects and blood sample collection time
64 Subjects at 0 hour & nearest to 48 hour
64 Subjects at 0 hour & nearest to onset time
104 Subjects at 0 hour & nearest to 48 hour
104 Subjects at 0 hour & nearest to onset time
Dividing data into training and testing
Training Data
Test Data
Train the models using 10 fold cross validation
KNN
Random Forest
Linear SVM
Non-Linear SVM (RBF)
Important Genes identification
Performance evaluation of classifiers and comparative analysis
Predicting State of Infection

Fig. 2. Experimental design for distinguishing infected subjects form non-infected subjects by exploiting ML algorithm's ability to learn pattern.

We believe that our experiments play a useful role in determining the involvement of particular genes in the states of infection at the early stage of the disease. We took the data of all the unambiguous subjects and divided it into four subsets as shown in Fig. 2. The details of the four experiments designed using these four subsets of data can be seen in Fig. 3 and Fig. 4. For each of the four subsets of the data, we partitioned the data into training and test sets, and then applied four well established ML approaches. The random sampling is done with preserving the class distribution to partition the whole data into training and test. In order to reduce the risk of overfitting we have applied 10-fold cross validation, repeated 3 times. The build model were then used to predict the state of infection for the kept test data. Finally, the performance evaluation and important gene identification steps have been carried out.

We identified 6 sets of data denoting different states, and labelled them State A, B, B', C, D, D' (see Fig. 3). State A contains the gene expression profile of all the 64 subjects which are healthy at time-point 0: these 64 subjects showed clear

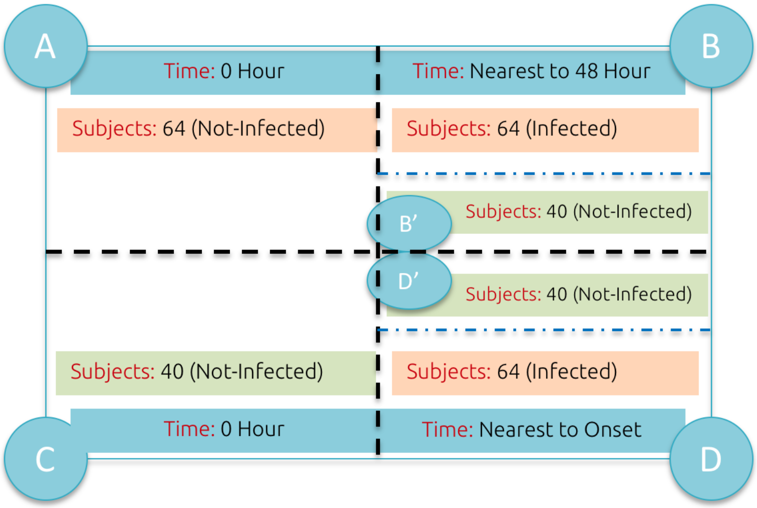


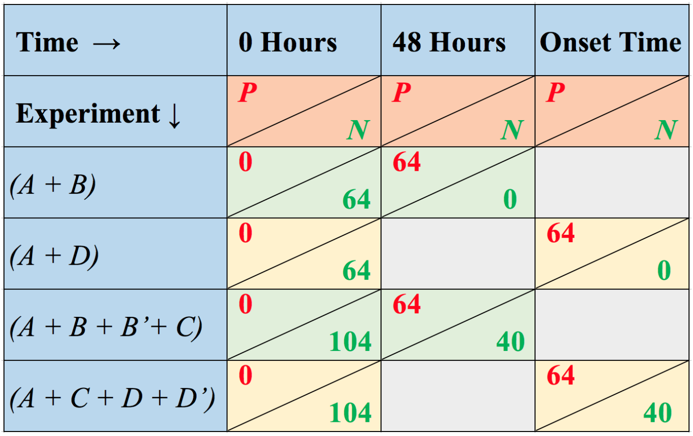
Fig. 3. A view of 4 experiments that comprises significant part of the overall experimental design.

Table 3. Detail of the experiments designed by combining two or more states of subjects status of gene expression profile.

Experiment No.	States	Description
1	$(A + B)$	64 subjects data collected at 0 and nearest to 48 hours
2	$(A + D)$	64 subjects data collected at 0 and nearest to onset time
3	$(A + B + B' + C)$	104 subjects data collected at 0 and nearest to 48 hours
4	$(A + C + D + D')$	104 subjects data collected at 0 and nearest to onset or average onset time.

signs of infection at some point of time after the inoculation of the virus. States B and D determine the gene expression profiles of the same 64 subjects at the time-point nearest to 48 hours or nearest to onset time, respectively. State C shows the gene expression profile of 40 subjects at 0 timestamp: these 40 subjects never get infected throughout the duration of the study. States B' and D' show the gene expression profiles of the same 40 subjects at nearest to 48 hours or at nearest to average onset time, respectively.

We carried out four experiments by combining two or more of the above states based on the number of infected and non-infected subjects and timestamps at which their blood samples are collected. These experiments are designed in such a way so that we can analyse the differences in disease prediction at two different early stages and find the most important Differentially Expressed Genes (DEG) across the different timestamps. The details of these experiments are shown in table 3. The numbers of positive and negative samples for each designed experiment at different time point are shown in Fig. 4.



Time →	0 Hours	48 Hours	Onset Time
Experiment ↓	*P* *N*	*P* *N*	*P* *N*
(A + B)	0 64	64 0	
(A + D)	0 64		64 0
(A + B + B'+ C)	0 104	64 40	
(A + C + D + D')	0 104		64 40

Fig. 4. Positive and negative sample counts for each experiment at different time points. Here P denotes positive samples (infected) and N denotes negative samples (non-infected).

4 Methodology

In this section we briefly explain the methodology used for classifying the state of health of any individual at any given time point t. It is well-known that no single ML algorithm is best for all kind of datasets, so we tested a selection of different ML approaches. In all experiments, 78% of the data is used for training the classifiers and the remaining 22% is kept as a hold-out test set. The stratified sampling is used to partition the whole data into training and test. To build the ML model for each algorithm we estimated model parameters over the training data using 10-fold cross validation, repeated 3 times.

First, we used the very simple baseline algorithm, k-NN which does not have any in-build capability to deal with high dimensional data [4], however, it can be used to set a base to compare the results and to see the improvements yielded by more complex algorithms. We also used the Random Forest algorithm which is an ensemble technique and has proven to be an efficient approach for the classification of microarray data as well as for gene selection [5]. We then employed both linear SVM [2] and SVM with RBF kernel which has inbuilt capability to learn pattern from high dimensional data [17]. We have used R programming language version 3.4.1 for coding [15].

4.1 *k*-Nearest Neighbour (*k*-NN)

k-NN has two stages, the first stage is the determination of the nearest neighbours i.e. the value of k and the second is the prediction of the class label using those neighbours. The "k" nearest neighbours are selected using a distance metric [4]. We have used Euclidean distance for our experiments. This distance metric is then used to determine the number of neighbours. There are various ways to use this distance metric to determine the class of the test sample. The most straightforward way is to assign the class that majority of k-nearest neighbours has. In the present work, the optimum value of k is searched over the range of $k = 1$ to 50. The best value of the parameter k obtained for each experiment can be found in Section 5.

4.2 Random Forest

Random Forest is often well-suited for microarray data. It can cope with noisy data and can be used when the number of samples is much smaller than the number of features. Furthermore, it can determine the relevance of variables in the decision process, which can be used for selecting the most relevant genes [5]. It is based on the ensemble of many classification trees [11,18]. Each classification tree is created by selecting a bootstrap sample from the whole training data and a random subset of variables with size denoted as $mtry$ are selected at each split. We have used the recommended value of $mtry$: $(mtry = \sqrt{(number\ of\ genes)})$ [5]. The number of trees in the ensemble is denoted as $ntree$. We have used $(ntree) = 10,001$ so that each variable can reach a sufficiently large likelihood to participate in forest building as well as in variable importance computations.

4.3 Support Vector Machine (SVM)

Assume that we have given a training set of instance-label pairs $(\boldsymbol{x}_i, y_i); \forall i \in \{1, 2, ..., l\}$ where $\boldsymbol{x}_i \in \mathbb{R}^n$ and $\boldsymbol{y} \in \{1, -1\}^l$, then the SVM [2, 7, 8] can be formulated and solved by the following optimization problem:

$$\begin{aligned} \min_{\boldsymbol{w}, b, \boldsymbol{\xi_i}} \quad & \frac{1}{2}\boldsymbol{w}^T\boldsymbol{w} + C\sum_{i=1}^{l}\xi_i, \\ \text{subject to} \quad & y_i\left(\boldsymbol{w}^T\phi\left(\boldsymbol{x}_i\right) + b\right) \geq 1 - \xi_i, \\ & \xi_i \geq 0. \end{aligned}$$

Here the parameter $C > 0$ is the penalty parameter of the error term [8] and $\xi_i \forall i \in \{1, 2, ..., l\}$ are positive slack variables [2]. For linear SVM, we did a search for best value of parameter C for a range of values $\left(C = 2^{-5}, 2^{-3}, ..., 2^{15}\right)$ and the one with the best 10-fold cross validation accuracy has finally been chosen.

We also used SVM with RBF kernel which is a non-linear kernel. There are four basic kernels that are frequently used: linear, polynomial, sigmoid, and RBF. We picked the RBF kernel, as recommended by Hsu et al. [8]. It has the following form:

$$K\left(\boldsymbol{x}_i, \boldsymbol{x}_j\right) = \exp\left(\tfrac{-\|\boldsymbol{x}_i - \boldsymbol{x}_j\|^2}{2\sigma^2}\right); \tfrac{1}{2\sigma^2} > 0.$$

We performed a grid-search over the values of C and σ using 10-fold cross validation. The different pairs of (C, σ) values are tried in the range of $(C = 2^{-5}, 2^{-3}, ..., 2^{15}; \sigma = 2^{-25}, 2^{-13}, ..., 2^{3})$ and the values with the best 10-fold cross validation accuracy are picked for the final model building (see Section 5).

5 Results

We experimentally obtained the 10-fold cross validation accuracy and hold-out test set accuracy using four algorithms including k-NN, Random Forest, linear SVM, and SVM with RBF Kernel. Based on the results obtained on the hold-out test set for all the four experiments, it can be concluded that the Random Forest model performs better than the rest of the algorithms (see Table 4, 5, 6 and 7). Random Forest gives the most stable and consistently highest accuracy on the hold-out test set. Moreover, the Random Forest has the additional capability to assign a relevance score to the variables (genes), hence, we have selected random forest for the determination of the genes playing the most important role in the development of the infection.

Table 4. Results on 64 infected subjects data at 0 and nearest to 48 hours (Experiment 1).

Sr. No.	**Algorithm**	**Model parameters**	**Accuracy 10-fold CV**	**Accuracy hold-out**
1	*k*-NN	$k = 23$	67.66%	53.57%
2	Random forest	$mtry = 109$ $ntree = 10,001$	**75.33%**	**64.29%**
3	Linear SVM	$C = 0.03125$	68.33%	64.29%
4	SVM with RBF kernel	$C = 5$ $\sigma = 3.051758 \times 10^{-5}$	73%	64.29%

Table 5. Results on 64 infected subjects data at 0 and nearest to the onset time (Experiment 2).

Sr. No.	**Algorithm**	**Model parameters**	**Accuracy 10-fold CV**	**Accuracy hold-out**
1	*k*-NN	$k = 24$	67.33%	53.57%
2	Random forest	$mtry = 109$ $ntree = 10,001$	73.33%	**82.14%**
3	Linear SVM	$C = 1$	75.33%	67.86%
4	SVM with RBF kernel	$C = 128$ $\sigma = 1.907349 \times 10^{-5}$	**76%**	71.43%

Table 6. Results on 104 subjects data at 0 and nearest to 48 hours (Experiment 3).

Sr. No.	**Algorithm**	**Model parameters**	**Accuracy 10-fold CV**	**Accuracy hold-out**
1	*k*-NN	$k = 3$	78.79%	77.78%
2	Random forest	$mtry = 109$ $ntree = 10,001$	81.99%	**80%**
3	Linear SVM	$C = 1$	77.39%	80%
4	SVM with RBF kernel	$C = 3$ $\sigma = 3.051758 \times 10^{-5}$	**82.83%**	77.78%

When the 10-fold cross-validation accuracy is considered, none of the algorithms uniformly outperform the others. The SVM with RBF Kernel is able to achieve highest 10-fold cross-validation accuracy for the last 3 experiments, however, the random forest also has similar performance for these last 3 experiments and is even better for the first experiment.

Table 7. Results on 104 subjects data at 0 and nearest to onset or average onset time (Experiment 4).

Sr. No.	**Algorithm**	**Model parameters**	**Accuracy 10-fold CV**	**Accuracy hold-out**
1	k-NN	$k = 4$	78.1%	77.78%
2	Random forest	$mtry = 109$ $ntree = 10,001$	84.26%	**77.78%**
3	Linear SVM	$C = 1$	81.77%	75.56%
4	SVM with RBF kernel	$C = 8$ $\sigma = 3.051758 \times 10^{-5}$	**85.45%**	75.56%

Overall, the results are best when "*Onset Time*" is considered for all 104 subjects (experiment 4) in comparison to the rest of the experiments. This is due to the significance of the "*Onset Time*" which shows that the blood samples collected at nearest to "*Onset Time*" is playing important role in discrimination of the infected samples from non-infected samples.

The highest accuracy obtained at nearest to 48 hours is 82.83% and at nearest to "*Onset Time*" is 85.45% which gives a positive sign that the prediction of respiratory viral infection at the early stage is possible with considerable accuracy.

6 Biomarker Identification

In this section, we show the top 10 important genes which are experimentally found to be the most important ones for the progression of respiratory viral infection and play an important role in the discrimination of infected samples from non-infected ones (see Table 8).

Table 8. The 10 most important genes with their overall importance score.

Sr. No.	**Probe IDs**	**Gene symbol**	**Overall importance score**
1	3434_at	IFIT1	100
2	23586_at	DDX58	92.9190292
3	5359_at	PLSCR1	77.908644
4	51056_at	LAP3	76.5473908
5	9111_at	NMI	74.5011703
6	23424_at	TDRD7	67.0779044
7	8743_at	TNFSF10	60.7319657
8	2633_at	GBP1	58.8176266
9	24138_at	IFIT5	53.9912712
10	4599_at	MX1	53.3913318

Random Forest also calculates the overall importance score for every feature. We used the caret package which calculates the overall importance score and scales it in a range from 0 to 100 [10]. We extracted the 109 genes which have highest overall importance score (100 to 15.97 in descending order) and plotted them to find the cut-off threshold to come up with 10 most important genes which contribute significantly in the progression of the disease (see Fig. 5).

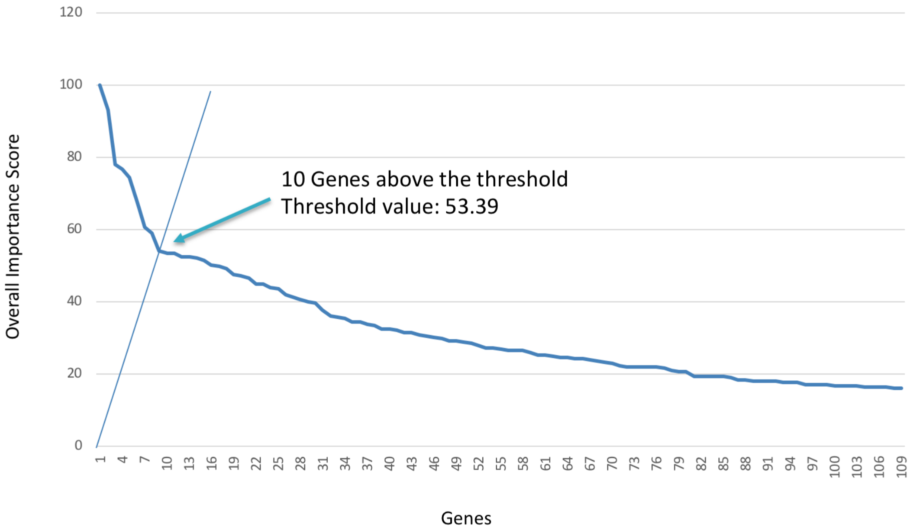


Fig. 5. Plot of overall importance score. 10 genes are above the cut-off threshold which are significantly the most important ones.

Moreover, we compared top 109 genes selected using random forest at nearest to 48 hours with top 109 genes selected at nearest to onset time and we found that these top 10 genes shown in Fig. 5 are common in both categories which shows that these top 10 genes are significantly important at both the early timestamps, i.e., nearest to 48 hours and nearest to onset time.

The Five-number summary of gene expression of input data for the identified top 10 genes at different timestamps can be seen in the form of boxplots shown in Fig. 6. The boxplots of the top 10 genes at timestamp 0, 48 and "*Onset Time*" support our findings. First, the boxplots support our claim that the top 10 genes reported by us are differentially expressed genes and contributing in progression of respiratory viral infection as their median value of gene expression at 0 hours and at "*Onset Time*" has a significant difference. Second, the boxplots also support the importance of genes, for example, gene IFIT1 has the highest importance score which can be seen in boxplot in terms of the highest difference in median gene expression values. Third, these plots also support our finding that the "*Onset Time*" is a better choice for learning the predictive models.

7 Discussion

We have identified 10 top genes from a set of 12,023 genes. To understand the mechanism associated with these genes we performed Gene Set Enrichment Analysis (GSEA) of these genes [19]. To further understand the association of retrieved disease mechanisms we performed Transcription Factor (TF) analysis [21]. During TF analysis we integrated TRANSFAC [24], BioGPS [26] and

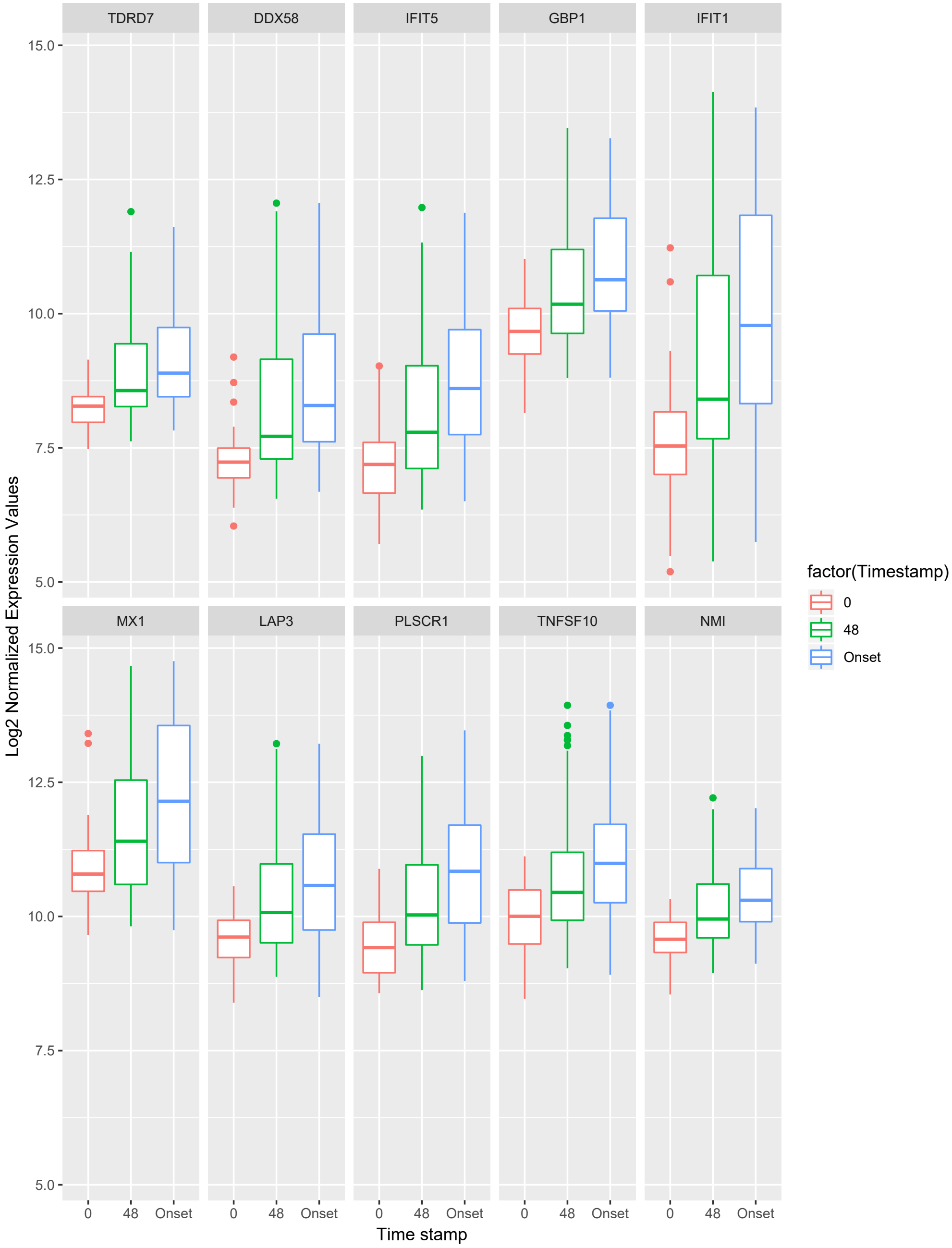


Fig. 6. Boxplots of the identified top 10 genes at 0 hours, 48 hours and "*Onset Time*".

JASPER database [16] and ran GSEA. The GSEA yielded the 441 associations against ten input genes. To understand the process associated with retrieved TFs and ten seed genes we performed functional annotation considering neighbouring genes [Table 9] and later without considering neighbouring genes [Table 10].

Table 9. Functional annotation and Disease enrichment analysis (DEA) with neighbour genes using Gene Set Enrichment Analysis (GSEA).

Gene Symbol	p-value	Geneset Friends	Total Friends	GO Annotation
IFIT1	1.32×10^{-18}	10	721	Interferon-induced protein with tetratricopeptide repeats 1
DDX58	4.24×10^{-17}	10	1020	DEAD(Asp-Glu-Ala-Asp) box polypeptide 58
IFIT5	1.89×10^{-15}	10	1491	Interferon-induced protein with tetratricopeptide repeats 5
GBP1	1.89×10^{-15}	10	1491	Guanylate binding protein 1, interferon-inducible
MX1	1.52×10^{-14}	10	1837	Myxovirus (influenza virus) resistance 1, interferon-inducible protein p78 (mouse)
PLSCR1	1.99×10^{-14}	10	1837	Phospholipid scramblase 1
TNFSF10	3.99×10^{-13}	10	2547	Tumour necrosis factor (ligand) superfamily, member 10
LAP3	5.67×10^{-11}	9	2505	Leucine aminopeptidase 3
NMI	9.54×10^{-11}	9	2655	N-myc (and STAT) interactor
TDRD7	7.75×10^{-10}	8	2018	Tumour domain containing 7

Here Geneset friends column explain how many genes contributed from the seed gene to establish the outcome. The MX1 gene has been known for its relevance to the intervention in the influenza virus infectious disease and it is known as the antiviral protein 1 [1, 22]. IFIT1, IFIT5 have interferon-induced protein with tetratricopeptide repeats 1 as annotation which indicates it's role in viral pathogenesis [6]. DDX58 is cytoplasmic viral RNA receptor, that is also known as DDX58 (DExD/H-box helicase 58). GBP1 induces infectious virus production in primary human macrophages [9, 27]. PLSCR1 are responsible for Hepatitis B virus replication with in-vitro and in-vivo both. LAP3 has already been predicted as principal viral response factor for all samples in H3N2 [3]. NMI has been reported as viral infection with Respiratory Syncytial Virus (RSV) infection and neuromuscular impairment (NMI) [23]. TDRD7 is known as interferons antiviral action and responsible for paramyxovirus replication [20]

To explore the effect of captured mechanism further we conducted a DEA using GSEA outcomes and as a result, most of the genes appear to be aligned against response to virus (83), immune response (467), innate immune response (105). GSEA, a measure to define the inhibition of a gene alongside its nearest neighbours and known interactions not only helped to understand the virology aspect of ten seed genes but also associated factors and genes. As we can observe from Table 10 most of the genes are involved in antiviral infection and their extended neighbours are against the response to the virus or process related to immune the body against the virus attack.

Table 10. Functional annotation without neighbour genes using Gene Set Enrichment Analysis (GSEA).

GO Biological Process	**p-value**	**GSEA Enriched GENES**
GO:0009615: response to virus (83)	$e^{-48.85}$	IRF7; PLSCR1; MX2; MX1; EIF2AK2; STAT1; BST2; IFIH1; TRIM22; IRF9; IFI35; DDX58; ISG15; RSAD2
GO:0006955: immune response (467)	$e^{-42.41}$	IFITM2; TAP1; IFITM3; IFI35; TNFSF10; GBP1; IFI6; CXCL10; IFI44L; OASL; OAS3; TRIM22; OAS2; PSMB9; OAS1; CXCL11; DDX58; IFIH1; SP110; PLSCR1
GO:0045087: innate immune response (105)	$e^{-11.45}$	IFIH1; DDX58; MX1; MX2; SP110

This provides a strong domain associated validation for these genes where core gene works as antiviral, and neighbour and interaction genes are immune and protective markers. This etiological discriminant prediction model and identified predictors is a potentially useful tool in epidemiological studies and viral infections.

8 Future work

We will be extending this work to establish identified genes for pathogen related infection. Findings could have diagnostic and prognostic implications by informing patient management and treatment choice at the point of care. Thus, further our efforts in this direction will establish the power of non-linear mathematical models to analyze complex biomedical datasets and highlight key pathways involved in pathogen-specific immune responses. The implemented classification methodology will support future database updates or largely integrated knowledge graphs to include new viral infection database to establish diagnostically strong biomarker with phenotype data, which will enrich the classifiers. The sets of identified genes can potentiate the improvement of the selectivity of non-invasive infection diagnostics. Currently, any type of viral data with labelled samples (i.e. case/control) can be used to discover small sets of biomarkers. In future we will also be focusing on the following aspects:

- Predictive performance assessed with an n-fold cross-validation scheme and simulation of a validation with unseen samples of multiple databases having integrated knowledge graphs (i.e. external validation).

- Biomarker extraction and inference of the predictive model by incorporating time series analysis performed on the data that includes all the different time-points.
- Permutation test to statistically validate the predictive performance of the model. On this point, currently we have already achieved the following:
  - The variable importance represents the contribution of each biomarker at an early stage within the predictive model.
  - The variable direction indicates how the change in values affect the overall prediction (e.g. probability of the disease to occur).

9 Conclusions

In this work, we aim to use a hybrid approach that harnesses the power of both ML and database integration to provide new insights and improve understanding of viral etiology, particularly related to the mechanism of viral diseases. To achieve this we conducted four different experiments to assess the capability of ML algorithms to predict the state of disease at the early stage by analyzing gene expression data. We establish that the prediction at an early stage is possible with considerable accuracy, 82.83% accuracy at nearest to 48 hours and 85.45% accuracy at nearest to onset time using 10-fold cross-validation, and accuracies of 80% and 82.14%, respectively on the hold-out test set. We got highest 10-fold cross-validation accuracy when all 104 subjects data are collected at 0 and nearest to onset or average onset time. This shows that for these kinds of studies if "*Onset Time*" is considered for learning the model then one can achieve considerably high accuracy in discrimination of infected from non-infected samples, however, it is observed that the accuracy on the hold-out test set is sometimes lower and sometimes higher than the 10-fold cross-validation accuracy, which means that the data has high variability and further analysis to capture this variability can improve the accuracy of prediction. The experiments indicate that the k-NN and linear SVM are not an ideal choice for these kinds of high dimensional datasets. By considering the fact that the Random Forest gives more stable and highest accuracy on unseen data (hold-out test set) for all the 4 experiments and due to its capability to assign importance score to variables, it is reasonable to choose Random Forest rather than the others. Moreover, we have identified top 10 most important genes which are having the maximum contribution in the progression of the respiratory viral infection at the early stage. The diagnosis and prevention of the respiratory viral infection at the early stage by targeting these genes can potentially improve the results than targeting the genes affected at the later stage of the infection.

Acknowledgements

This publication has emanated from research conducted with the financial
support of Science Foundation Ireland (SFI) under Grant Number
SFI/12/RC/2289, co-funded by the European Regional Development Fund.

References

1. Braun, B.A., Marcovitz, A., Camp, J.G., Jia, R., Bejerano, G.: Mx1 and mx2 key antiviral proteins are surprisingly lost in toothed whales. Proceedings of the National Academy of Sciences **112**(26), 8036–8040 (2015)
2. Burges, C.J.: A tutorial on support vector machines for pattern recognition. Data Mining and Knowledge Discovery **2**(2), 121–167 (Jun 1998)
3. Chen, M., Zaas, A., Woods, C., Ginsburg, G.S., Lucas, J., Dunson, D., Carin, L.: Predicting viral infection from high-dimensional biomarker trajectories. Journal of the American Statistical Association **106**(496), 1259–1279 (2011)
4. Cunningham, P., Delany, S.J.: k-nearest neighbour classifiers. Multiple Classifier Systems **34**, 1–17 (2007)
5. Díaz-Uriarte, R., De Andres, S.A.: Gene selection and classification of microarray data using random forest. BMC bioinformatics **7**(1), 3 (2006)
6. Fensterl, V., Sen, G.C.: Interferon-induced ifit proteins: their role in viral pathogenesis. Journal of virology pp. JVI–02744 (2014)
7. Guyon, I., Weston, J., Barnhill, S., Vapnik, V.: Gene selection for cancer classification using support vector machines. Machine Learning **46**(1), 389–422 (Jan 2002)
8. Hsu, C.W., Chang, C.C., Lin, C.J.: A practical guide to support vector classification (2010)
9. Krapp, C., Hotter, D., Gawanbacht, A., McLaren, P.J., Kluge, S.F., Stürzel, C.M., Mack, K., Reith, E., Engelhart, S., Ciuffi, A., et al.: Guanylate binding protein (gbp) 5 is an interferon-inducible inhibitor of hiv-1 infectivity. Cell host & microbe **19**(4), 504–514 (2016)
10. Kuhn, M.: Building predictive models in r using the caret package. Journal of Statistical Software, Articles **28**(5), 1–26 (2008)
11. Liaw, A., Wiener, M.: Classification and Regression by randomForest. R News **2**(3), 18–22 (2002), http://CRAN.R-project.org/doc/Rnews/
12. Liu, T.Y., Burke, T., Park, L.P., Woods, C.W., Zaas, A.K., Ginsburg, G.S., Hero, A.O.: An individualized predictor of health and disease using paired reference and target samples. BMC Bioinformatics **17**(1), 47 (Jan 2016)
13. McCloskey, B., Dar, O., Zumla, A., Heymann, D.L.: Emerging infectious diseases and pandemic potential: status quo and reducing risk of global spread. The Lancet Infectious Diseases **14**(10), 1001 – 1010 (2014)
14. Molinari, N.A.M., Ortega-Sanchez, I.R., Messonnier, M.L., Thompson, W.W., Wortley, P.M., Weintraub, E., Bridges, C.B.: The annual impact of seasonal influenza in the US: Measuring disease burden and costs. Vaccine **25**(27), 5086 – 5096 (2007)
15. R Core Team: R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria (2013), http://www.R-project.org/
16. Sandelin, A., Alkema, W., Engström, P., Wasserman, W.W., Lenhard, B.: Jaspar: an open-access database for eukaryotic transcription factor binding profiles. Nucleic acids research **32**(suppl_1), D91–D94 (2004)
17. Scholkopf, B., Sung, K.K., Burges, C.J.C., Girosi, F., Niyogi, P., Poggio, T., Vapnik, V.: Comparing support vector machines with Gaussian kernels to radial basis function classifiers. IEEE Trans. Signal Process. **45**(11), 2758–2765 (Nov 1997)
18. Statistics, L.B., Breiman, L.: Random forests. In: Machine Learning. pp. 5–32 (2001)

19. Subramanian, A., Tamayo, P., Mootha, V.K., Mukherjee, S., Ebert, B.L., Gillette, M.A., Paulovich, A., Pomeroy, S.L., Golub, T.R., Lander, E.S., et al.: Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proceedings of the National Academy of Sciences **102**(43), 15545–15550 (2005)
20. Subramanian, G., Kuzmanovic, T., Zhang, Y., Peter, C.B., Veleeparambil, M., Chakravarti, R., Sen, G.C., Chattopadhyay, S.: A new mechanism of interferons antiviral action: Induction of autophagy, essential for paramyxovirus replication, is inhibited by the interferon stimulated gene, tdrd7. PLoS pathogens **14**(1), e1006877 (2018)
21. Vaquerizas, J.M., Kummerfeld, S.K., Teichmann, S.A., Luscombe, N.M.: A census of human transcription factors: function, expression and evolution. Nature Reviews Genetics **10**(4), 252 (2009)
22. Verhelst, J., Parthoens, E., Schepens, B., Fiers, W., Saelens, X.: Interferon-inducible protein mx1 inhibits influenza virus by interfering with functional viral ribonucleoprotein complex assembly. Journal of virology **86**(24), 13445–13455 (2012)
23. Wilkesmann, A., Ammann, R.A., Schildgen, O., Eis-Hübinger, A.M., Müller, A., Seidenberg, J., Stephan, V., Rieger, C., Herting, E., Wygold, T., et al.: Hospitalized children with respiratory syncytial virus infection and neuromuscular impairment face an increased risk of a complicated course. The Pediatric infectious disease journal **26**(6), 485–491 (2007)
24. Wingender, E., Chen, X., Hehl, R., Karas, H., Liebich, I., Matys, V., Meinhardt, T., Prüß, M., Reuter, I., Schacherer, F.: Transfac: an integrated system for gene expression regulation. Nucleic acids research **28**(1), 316–319 (2000)
25. Woods, C.W., McClain, M.T., Chen, M., Zaas, A.K., Nicholson, B.P., Varkey, J., Veldman, T., Kingsmore, S.F., Huang, Y., Lambkin-Williams, R., et al.: A host transcriptional signature for presymptomatic detection of infection in humans exposed to influenza H1N1 or H3N2. PloS one **8**(1), e52198 (2013)
26. Wu, C., Orozco, C., Boyer, J., Leglise, M., Goodale, J., Batalov, S., Hodge, C.L., Haase, J., Janes, J., Huss, J.W., et al.: Biogps: an extensible and customizable portal for querying and organizing gene annotation resources. Genome biology **10**(11), R130 (2009)
27. Zhu, Z., Shi, Z., Yan, W., Wei, J., Shao, D., Deng, X., Wang, S., Li, B., Tong, G., Ma, Z.: Nonstructural protein 1 of influenza a virus interacts with human guanylate-binding protein 1 to antagonize antiviral activity. PloS one **8**(2), e55920 (2013)