Optimizing the Detection of Antiviral-Resistant Influenza-A Strains using Machine Learning

N. Shaltout, A. Rafea, A. Moustafa, M. Moustafa & M. ElHefnawi

***Abstract*—The research optimizes the detection of Influenza-A's resistance to Osetamivir (Tamiflu), an antiviral drug, by: (a) Determining the best number of Principle Component Analysis (PCA) features to generate accurate results. (b) Selecting the best classifier for detecting Oseltamivir-resistance, when using PCA features, by comparing the performance of neural networks (NNs) & decision trees (DTs) (c) Comparing findings to previous experiments conducted on detecting Adamantane-resistance in the H1N1 strain. (d) Noting the performance of using Information Gain (IG) when detecting Oseltamivir-resistance. Viral DNA sequences from the *NA* segment, belonging to the 2009 pH1N1 strain, were used; they possess a 90% elimination rate by Oseltamivir. Sequences understudy were further divided into Oseltamivir-resistant & Oseltamivir-susceptible. The performance measures used were accuracy, sensitivity, specificity, precision, & time. Using IG resulted in classifier overlearning when detecting Oseltamivir-resistance. NNs outperformed DTs, when using 40 PCA features from the *NA* segment to detect Oseltamivir-resistance, with an overall accuracy increase of 5%. In contrast DTs showed better performance when detecting Adamantane-resistance with just 3 PCA features from the *M1/M2 (M)* segment, due to the availability of a larger, more balanced training dataset. Using 40 PCA features additionally enhanced the detection of Adamantine-resistance on the *NA* segment by 5%. The findings can be used later for building a multilabel antiviral-resistance detector .**

***Index Terms*—Principle Component Analysis (PCA), Influenza-A, machine learning, Adamantane-resistance, Oseltamivir-resistance**

I. INTRODUCTION

Influenza-A's high mutation rate is the cause for its high morbidity/mortality rates during virulent pandemics. Antiviral drugs are sparsely used during outbreaks as the virus's mutation to an antiviral-resistant strain is unpredictable. Antiviral brands having lower success rates at eliminating the virus are excluded from consideration altogether e.g. Adamantane [1]. As a result numerous infected people do not get the required timely treatment. The virus additionally develops increasing resistance to drugs it is susceptible to via mutation. Thus a drug which was effective in solving one pandemic, can become ineffective in future pandemics [1]. Oseltamivir (Tamiflu) is one such drug; it targets the H1N1 viral strains. Oseltamivir's effectiveness rate was 98% during the 2007-2008 Influenza season, but was reduced to 90% during the 2008-2009 season, due to the virus's mutation.

This work will use cDNA/RNA (DNA) for virus analysis. Fig. 1. shows a simplified version of Influenza A's most important features. Influenza-A's genetic material is encoded in 8 RNA segments. The RNA segments eventual mutation causes the antiviral drugs to be ineffective against the resulting viral proteins . The mutation is further enhanced by antigenic shift [2], possibly increasing the virus's drug resistance. Currently there are at least 16 hemagglutinine/ *HA* (H) & 9 neuraminidase/ *NA* (N) known subtypes due to Influenza-A's rapid mutation rate [3]. Viral strains are coded using these subtypes. E.g. H1N1. By improving the classification of antiviral drug resistance using viral RNA, these mutations can be spotted early.

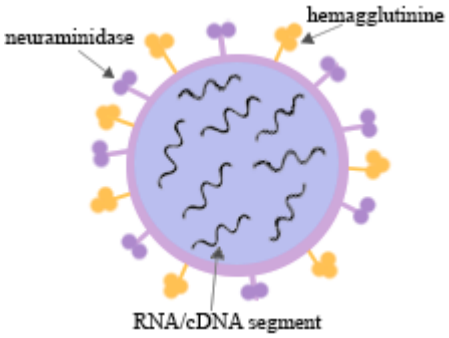

Fig. 1. Summary of Influenza-A's structure.

In some cases, the viral strains may mutate gaining resistance to both Oseltamivir & Adamantane respectively [4]. Improving the classification of Oseltamivir-resistance via machine learning can thus allow the prediction of the aforementioned scenario more rapidly. Previous studies on machine learning were conducted to classify virus's resistance to Adamantane. However, no studies were conducted yet on improving the classification of Oseltamivir-resistance. The 2008-2009 H1N1 strains were 10% susceptible to Oseltamivir. Thus more research is needed to account for Oseltamivir's biased training data compared to Adamantane's somewhat balanced viral dataset.

Principle Component Analysis (PCA) was chosen as a preprocessing technique, pre-evaluation of classifiers, due to its ability to compress data while retaining the most variable information. It is also predicted to prevent overlearning when using DNA sequences as follows. Influenza-A's aligned DNA sequences reach an average of 1500 nucleotides per sequence. With smaller datasets, the number

Manuscript received June 19, 2016; revised August 15, 2016.
Sponsored by The American University in Cairo.
N. Shaltout is with the research department of Kngine, Cairo, Egypt (e-mail: nermeena@gmail.com).
A. Rafea & M. Moustafa, & A. Moustafa are with The American University of Cairo, New Cairo, Cairo, 11835, Egypt.
M. ElHefnawi is with the Informatics & System Department and Biomedical Informatics & Chemoinformatics group, Division of Engineering Research and Centre of Excellence for Advanced Sciences. National Research Center Tahreer Street, Cairo, 12311, Egypt.

training instances can be smaller than the number of nucleotide per sequence used for training the classifier. The classifier can overlearn as a result. Compressing DNA sequences with PCA is predicted to solve this problem when detecting Oseltamivir-resistance.

The remainder of the paper is structured as follows: Section II discusses the related works; Section III describes the research aim. Section IV & Section V summarize the PCA algorithm & methodology. Lastly, Section VI & VII contain the experimental results & conclusion.

II. RELATED WORKS

To render machine learning efficient, feature selection is usually applied to Bioinformatics datasets having high dimensionality. Ref. [5] describes the pros & cons of three features selection techniques & their application to Bioinformatics problems: filter, embedded & wrapper methods. They point out that univariate filter methods are preferred for high dimensionality data analysis due to their speed, scalability & classifier independence. PCA is a filter technique which suits the high dimensionality of Influenza-A's data. Ref. [6] utilized PCA to select features prior to classifying Adamantane-resistance. The technique was not explored when detecting Oseltamivir-resistance.

Leung describes another filter method known as Information Gain (IG) when identifying the RNA biomarkers in the Hepatitis-B virus [7]. The biomarkers determined if the virus has the ability to cause hepatic cancer. IG selected the most informative nucleotide positions in the sequence that could differentiate two or more classes. Ref. [8] also applied IG to Influenza-A's cDNA sequences to classify hosts. Although classification performance was improved, the IG algorithm generated up to 100 features & the process of selecting features was relatively time consuming. Ref. [9] used IG to further reduce DNA sequence to 10 units, but the results varied on the viral subtype; the H1 subtype yielded more accurate results. In [6] a speedier feature selection technique using PCA was devised but it was only optimized for detecting Adamantane-resistance.

In order for DNA & protein sequences to be used in machine learning, they must be encoded in a format recognizable by the utilized classifiers. Attaluri contrasted the effect of different neural network (NN) encoding schemes on Influenza-A classification performance [10]. He shows that including the multiple sequence alignment (MSA) gaps in direct encoding schemes increased classification accuracy. He additionally determined the k-tuple frequencies that generated optimal results when using indirect encoding schemes for NNs. This work encodes DNA sequences directly with the former method prior to using PCA; the MSA gap is included to increase accuracy.

The following key researches on the use of machine learning for Influenza-A analysis were conducted by Attaluri, ElHefnawi *et al*. & Shaltout *et al* [6,8-12]. Ref. [11] used hidden Markov Models (HMMs) to classify Influenza-A based on hosts & subtypes. Protein sequences from the HA segment pertaining to subtypes H1, H2, H3, H4 & H5 were used for subtype classification. The virus sequences were further divided into human & nonhuman hosts. The research yielded an overall subtype classification accuracy of 100%, whereas the host classification accuracy ranged from 50% to 100%, depending on the viral subtype.

Ref. [12] compared the use of HMMs & decision trees (DTs) on extracted host associated protein signatures, which aided in increasing host identification accuracy. The experiments were conducted on the HA protein of various subtypes. DTs yielded higher host classification accuracies, ranging from 92%-100%, compared to HMMs. Both works [11, 12] did not explore classification performance at the RNA/ cDNA level.

Attaluri analyzed the use of various classifiers, namingly, NNs, DTs & support vector machines (SVMs) for identifying Influenza-A hosts & subtypes [10]. The experiments were conducted using both cDNA & Protein sequences. A subset of virus sequences belonging to the H1N1 strain was used for identifying hosts. Sequences belonging to the H1, H2, H3, N1, & N2 subtypes were used for subtype classification analysis. The overall classification accuracies, for both subtype & host classification were 96.5%, 96.2% & 95.1% when using DTs, SVMs & NNs respectively. Attaluri additionally integrated DTs & HMMs in a hybrid model to identify Influenza-A hosts & subtypes [10]. He used DTs to extract informative positions from the cDNA sequences then converted them into their corresponding protein sequences. The acquired protein sequences were then used as input to the HMM classifier. Both viral host & subtype classification were analyzed. The technique yielded an overall accuracy of 97%.

Although the research showed promising results, it still suffered from a few drawbacks: There was no unified method for comparing classifiers; classifier analysis was conducted on both DNA & protein data, so a standardized comparison of classification techniques could not be conducted. Feature selection was carried out using protein data but not fully explored with DNA data. The efficiency of the classification process was also unmeasured.

These problems were later addressed in [6,8,9]. Refs. [8, 9] studied the effect of using IG in terms of efficiency & speed on DNA classification of Influenza-A hosts. The classification performance improved despite IG being time consuming. Ref. [6] addressed some of the performance issues in [8,9] by analyzing the effect of using PCA compared to using IG on Adamantane-resistance determination. PCA proved to be more efficient than IG however its effect on detecting Influenza-A's resistance to Oseltamivir was not explored. The effect of using IG for preprocessing was also untested on smaller DNA datasets.

This paper seeks to improve detection of Oseltamivir-resistance by using PCA as a feature selection technique: As Oseltamivir has a biased dataset, its classification will be improved by: (a) reevaluating the performance of PCA features on two of the most common classification techniques: DTs & NNs; (b) comparing the results with those achieved with detecting Adamantane-resistance in [6]. The effect of using IG on detecting Oseltamivir-resistance will also be briefly analyzed.

ISBN: 978-988-14048-2-4
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)

III. RESEARCH AIM

The research seeks to improve the detection of Oseltamivir-resistance by creating features that will not cause overlearning. The system will be implemented by compressing the DNA features using PCA, then feeding the PCA features to two of the most commonly used classifier(s): DTs & NNs. This is done without protein conversion to enhance RNA/cDNA virus analysis & detect RNA mutations.

Using PCA is expected to reduce the features without deteriorating classifier performance, as shown with detecting Adamantine-resistance in [6]. However the optimal number of features for detecting Oseltamivir-resistance is not expected to be the same as Adamantane, due to its skewed dataset: only 10% of the dataset shows Oseltamivir-resistance. The PCA features required will be determined as follows:

--Running the experiment on the most important viral segment(s) belonging to the H1N1 strain & labeled with Oseltamivir-Resistance.

--Using PCA to extract the cDNA features showing the greatest variation.

--Feeding the extracted features directly to a NN & determining the least possible number of PCA features that will increase performance.

--Testing the performance of the optimal number of PCA features, from the previous step, on a DT.

--Comparing the classification performance of DTs & NNs in detecting Oseltamivir-resistance, when using PCA.

--Comparing the classification performance of Oseltamivir-resistance when using PCA features to that of Adamantane-resistance as in [6].

IV. APPLYING PCA TO DNA SEQUENCES

The PCA is a compression method that reduces high dimensionality data by projecting it unto vectors in the direction of highest variability of the data. It was selected for the following reasons: Firstly, previous research on detecting Adamantane-resistance showed that Influenza-A sequences can be reduced from around 1500 to 3 features while maintaining classification accuracy & speed [6]. Secondly, the DNA has only 4 differentiable units: "A", "G", "C", & "T". Since we are analyzing viruses of the same strain & species, the difference between the DNA sequences is not easily recognizable without preprocessing. Using PCA will help detect subtle differences among the DNA sequences while reducing the features significantly.

Third, PCA is predicted to prevent potential overlearning on the unbalanced Oseltamivir dataset. Without feature selection, the number of sequences per DNA far exceeds the number of training instances, and may result in poor classification. The repetition of the four nucleotide values, in addition to the biased dataset can cause classifier overfitting. This may also hold true even for feature selection techniques such as IG which simply selects the sequences' informative positions. Lastly, the feature reduction can also aid later in expanding the system to a multilabel classifier.

By using PCA features for classifying Oseltamivir-resistance we hope to:

--Find subtle differences in DNA that are otherwise hard to detect, since we are differentiating between viruses of the same strain & species.

--Find the optimal number of reduced features for detecting Oseltamivir-resistance that neither causes reduced performance nor overlearning.

--Measure which classifier achieves better performance with the optimal number of PCA features determined.

--Compare the classification performance of Oseltamivir-resistance to Adamantane-resistance when using PCA.

The derivation of the PCA algorithm is detailed in [13]. To apply the PCA algorithm to a dataset, the following steps must be performed:

1) After numerically encoding the DNA sequences, convert the dataset to mean centered points by subtracting the dataset values, x, from the mean, m. The mean attained is a row vector containing the mean per nucleotide position.

2) Calculate the scatter/covariance matrix, S, of the sequence using (1), where $(x\text{-}m)$ represents the mean centered sequences.

3) Determine the eigenvectors, e, arranged using highest eigenvalues, λ, of the resulting scatter matrix, S, as in (2).

4) After selecting best n-eigenvectors with highest eigenvalues, reduce the data to n-dimensions by multiplying the resulting eigenvectors, each individually represented by e, with the mean projected sequences as shown in (3). This is done once per sequence; each sequence will then be represented by n-features instead of roughly 1500 features.

$$S = \sum (x-m)'*(x-m) \quad (1)$$

$$Se = \lambda e \quad (2)$$

$$a_k = e^T (x_k - m) \quad (3)$$

V.METHODOLOGY

The following chapter explains the experimentation steps. The main steps are: data collection, sequence alignment, feature selection, classifier construction, classifier evaluation, & classifier comparison.

--*Data Collection*: Collect DNA data from online Influenza databases (http://www.fludb.org). Select Oseltamivir sequences structurally resembling the H1N1 2009 pandemic sequences. Select only sequences known to infect humans; sequences infecting Avian & Swine hosts are excluded due to insufficient data. Non-annotated sequences & duplicate sequences are deleted.

When analyzing Oseltamivir-resistance, the initial experiments were conducted on the *NA* segment as it is the target of Oseltamivir [14]. When comparing with studies on Adamantane-resistance in [6], the results of classifying the *M1/M2* segment are used; The *M1/M2* segment it is the target of Adamantane [15]. Additional experiments involving the use of the *NA* segment to detect Adamantine-resistance were conducted. This is to observe if the *NA* segment can be used in future work to construct a multilabel classifier for detecting both Oseltamivir & Adamantane resistance.

--*Data Alignment:* The PCA algorithm is only applicable to datasets containing features of equal lengths; DNA sequences are stored with variable lengths online. Align the

ISBN: 978-988-14048-2-4
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)

data with multiple sequence alignment (MSA) to unify the sequences' length. MSA was performed using *Mafft*, due to its ability to swiftly align high dimensionality data using fast Fourier transform. It is faster than classical MSA programs that take hours to align DNA sequences.

--*Data Preprocessing*:

(a) Separate the two datasets of DNA sequences using the annotation in the header files to two classes: Antiviral-Resistant & Antiviral-Susceptible. The Oseltamivir dataset is skewed, with the Oseltmivir-Susceptible data being much larger than Oseltamivir-Resistant data. To account for this, after sequence randomization, a smaller sample of cDNA sequences was selected from the Oseltamivir-Susceptible dataset, equivalent to the size of the Oseltamivir-Resistant dataset.

(b) Encode the viral sequences prior to classification. PCA is a statistical method, whereas the nucleotide values of DNA are nominal: "A","C", "G", & "T". The sequences are converted to numerical values using the encoding scheme in Table-I. The gaps, "-", introduced by MSA, are included to increase classification accuracy as described in [10]. The same encoding scheme was used for classifying Adamantane-resistance in [6] & showed promising results.

(c) Divide the Oseltamivir dataset into 70% training & 30% testing data.

TABLE I
DNA ENCODING SCHEME PRIOR TO USING PCA

Nucleotide Value	Numerical Encoding
"A"	-20
"G"	-10
"-"	0
"T"	10
"C"	20

--*Feature Selection*: Apply the PCA algorithm to each dataset as follows:

For the *training* dataset:

(a) Find the mean of the training dataset for both classes , across all nucleotide positions. The mean vector will have a rough size of 1500 post alignment. Obtain the mean-centered points by subtracting the mean from all the sequences.

(b) Apply the PCA algorithm to the dataset by finding the scatter matrix of the mean centered points. Then, obtaining the best n-eigenvectors corresponding to the highest n-eigenvalues from the scatter matrix. Finally, project the mean normalized data by multiplying it by each of the eigenvectors to get the new feature space.

For the *testing* dataset:

(a) For each class, obtain the mean-centered points by subtracting the mean vector of the training dataset from the encoded sequences of the testing dataset.

(b) Multiply the mean-centered points by the eigenvalues attained in the training step.

--*Classifiers Construction*:

For each antiviral under study: build, train & test the binary classifiers differentiating between antiviral-resistant strains & antiviral-susceptible strains. This is implemented using both NNs & DTs.

NNs were selected as they are robust to changes & missing data. Their ability to classify H1N1 based on antiviral resistance on Adamantane was already analyzed in previous experiments with PCA [6]. However their use on determining Oseltamivir-resistance was unexplored. The Oseltamivir dataset is also uneven with 90% of the viral sequences being susceptible to the antiviral. The effect of the uneven dataset might cause the NN to overlearn, so using PCA to counteract that will be analyzed.

DTs will also be analyzed as they produced promising results when detecting Adamantane-resistance [6]. Their performance with the skewed Oseltamivir dataset should be compared to NNs to select the better classifier for detecting Oseltamivir-resistance. PCA is a numerical method thus DTs that can process numerical attributes will be utilized.

For the Oseltamivir dataset, a similar NN structure to that in [6] was used. *The NN was constructed as follows*:

(a) A three layered feed forward NN was used.

(b) The network inputs are the n-compressed points attained by applying the PCA algorithm on the DNA sequences . In [6] using 3 PCA features for detecting Adamantane-resistance showed satisfactory results. However when analyzing Oseltamivir-resistance, the PCA features' dimensionality will be tested at 5 to 10 increment intervals, starting at 3 PCA features, so as to find the optimal number of features that enhance classification. More features might be needed to improve classification & decrease the data bias in the Oseltamivir dataset. The experiment will be stopped when accurate classification is achieved. A NN trained with DNA sequences sans feature selection will also evaluated on the *NA* segment to observe if overlearning will occur; It uses the encoding scheme in Table-I prior to classification. NNs using varying number of features generated by IG will additionally be tested.

(c) The default of ten hidden neurons was utilized.

(d) A binary target output of 0 was used for representing antiviral-resistance, whereas a value of 1 was used to represent antiviral-susceptibility.

(e) The nprtool in *Matlab* was used to build, train & test the NN. The scaled conjugate graduate was set as the default NN training algorithm. The mean squared error was used as the default NN evaluation algorithm.

(f) To prevent overfitting, the sequences in the training set were further divided into 70% training, 15% testing & 15% validation sets. The validation set will stop the training when its accuracy decreases below that of the training/ testing set.

The DT was constructed as follows:

(a) A Reduced Error Pruning (REP) DT is used to perform the classification as it can handle numerical values. The method uses reduced error pruning which increases the speed of building & training the classifier . IG is also used to split on the nodes or features, thus concentrating on the more important features.

(b) To prevent overfitting, the DT was trained using ten-fold validation .

(c) The *Weka* tool is used to train & test the REP-DT. The default *Weka* setting of the REP-DT are used.

DTs are built when the minimum number of PCA features producing satisfactory results is determined via NNs. This is done for both viral datasets: Oseltamivir & Adamantane. The target viral segment of each drug is used to build the DT.

ISBN: 978-988-14048-2-4
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)

--*Classifier Evaluation*: After each experiment, evaluate the results on the testing set using confusion matrices & ROC curves. Measure the time taken to train the classifier. Observe the effect of PCA on specificity or the true negative rate since Oseltamivir-resistance is the negative class.

--*Classifier Comparison:* Compare the performance of PCA features on NNs to DTs when detecting Oseltamivir-resistance. Compare the results to those attained when determining Adamantane-resistance in [6]. Compare the use of IG to PCA when determining Oseltamivir-resistance.

VI. EXPERIMENTAL RESULTS

Table-II summarizes the number of sequences/training instances used for training the classifiers. The aligned length of the sequence before feature selection is also shown. The number of nucleotides in the *NA* segment far exceeds the number of classifier training instances used for detecting Oseltamivir-resistance if PCA isn't applied.

TABLE II
NUMBER OF TRAINING INSTANCES VS. NUMBER OF INPUTS

Attribute of Interest	Training Instances	Nucleotides in (*M*) Segment	Nucleotides in (*NA*) Segment
Oseltamivir-resistance	**957**	1030	**1635**
Adamantan-resistance	3825	1030	1635

In [6], training NNs & DTs with PCA features to detect Adamantane-resistance, yielded the following results on the testing dataset summarized in Table-III. The experiments were conducted on the *M1/M2* (*M*) segment. The time in seconds represents the time used to build the classifier prior to testing.

TABLE III
THE EFFECT OF PCA ON DETECTING ADAMANTINE RESISTANCE .

Class-ifier	Seg-ment	No. of features	Time(s)	Acc-uracy	Sens-itivity	Spec-itivity	Prec-ision
DT	M	100 (IG)	0.3	98.2%	98%	98.6%	98.4%
DT	**M**	**3 (PCA)**	**0.06**	**98.5%**	**97.9%**	**99.3%**	**99.5%**
NN	M	3 (PCA)	5	96.5%	99.3%	94.4%	92.9%

The first entry shows the performance of information gain (IG).

When using PCA features to detect Adamantane-resistance, Table-III shows that DTs outperform NNs in both overall classification accuracy & efficiency. The accuracy, specificity & precision were increased by 2%, 5%, & 7% respectively. Only 3 PCA features were needed to achieve this. Ref. [6] shows that the performance of PCA features is comparable to using informative positions (IG) to detect Adamantane-resistance. PCA is also more time efficient.

Training the NNs & DTs to detect Oseltamivir-resistance, after feature compression with PCA, yielded the results summarized in Table-IV. The *NA* segment was used. The time in seconds represents the time used to build the classifier prior to testing. Unlike with Adamantane, using 3 PCA features did not yield optimal results. The sample size of the dataset used for determining Oseltamivir-resistance was smaller, leading to the deterioration of classification performance with smaller numbers of PCA features.

Additionally, Table-IV shows that training the NN with the raw DNA sequences, without applying PCA, results in an overfit classifier. The performance measure values were at a 100% indicating a classifier that cannot generalize. A NN was also trained by using information gain (IG) to select the most informative RNA/cDNA positions as described in [8]. Using IG as a preprocessing step similarly yielded an overfit classifier, even when varying the number of features from 3 to 100. The IG algorithm needs a large amount of data to build a classifier that generalizes. This is due to DNA sequences possessing only 4 possible values per nucleotide position. PCA solves this problem by numerically encoding the sequences & projecting the dataset in the direction of highest variability, thus creating far more than 4 possible values per nucleotide position.

Table-IV also shows that using 40 PCA features achieved satisfactory classification results when using NNs. Only the number of features generating optimum results on the NN are tested on the REP-DT, thus the DT performance using 40 PCA features was measured & recorded in Table-IV.

TABLE IV
THE EFFECT OF PCA ON DETECTING OSELTAMIVIR RESISTANCE

Class-ifier	No. of features	Time (s)	Acc-uracy	Sens-itivity	Spec-ificity	Prec-ision
	IG/raw	3	100%	100%	100%	100%
	3	2	87.5%	95.1%	79.9%	82.6%
	10	2	87.5%	94.1%	80.9%	83.1%
NNs	20	2	88.5%	92.6%	84.3%	85.5%
	30	2	96.1%	95.6%	96.6%	96.5%
	35	2	98%	99%	97.1%	97.1%
	40	**5**	**98.3%**	**98%**	**98.5%**	**98.5%**
DT	40	0.42	92.9%	94.6%	91.2%	91.5%

The first entry shows the results without PCA. IG and raw DNA data had the same results.

When using PCA features from the NA segment to detect Oseltamivir-resistance, the results show that NNs outperform DTs in overall classification accuracy. However this was done with 40 PCA features which is greater than the number of PCA features needed to detect Adamantane-resistance. The accuracy, specificity & precision were increased by 5.4%, 7.3%, & 7%. The results also show that training NNs with the right amount of PCA features, prevents NN overlearning from smaller viral datasets. The ROC curves in Fig. 2 summarize the performances of DT & NN when PCA features are used in detecting antiviral-resistance.

Additional tests were conducted to see if training a NN with more PCA features can improve the detection of Adamantane-resistance in both the *M1/M2*(*M)* & *NA* segments. The results are shown in Table-V. The performance on both segments improved. Compared to using 3 PCA features, using 40 PCA features significantly increased the accuracy of detecting Adamantane-resistance on the *NA* segment. The specificity increased by 7.8%.

TABLE V
THE EFFECT OF INCREASING PCA FEATURES ON ADAMANTINE RESISTANCE DETECTION WHEN USING NNS

Class-ifier	Seg-ment	No. of features	Time (s)	Acc-uracy	Sens-itivity	Spec-itivity	Prec-ision
NN	M	40	5	97.8%	98.7%	97.1%	96.2%
NN	**NA**	**40**	**41**	**96.9%**	**96.1%**	**97.4%**	**95.7%**
NN	NA	3	25	91.8%	95.3%	89.6%	84.9%

Using the above findings building a multilabel classifier that can classify both Oseltamivir & Adamantane-resistance may be possible with the use of the *NA* segment & enough PCA features as a future expansion to the project.

ISBN: 978-988-14048-2-4
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)

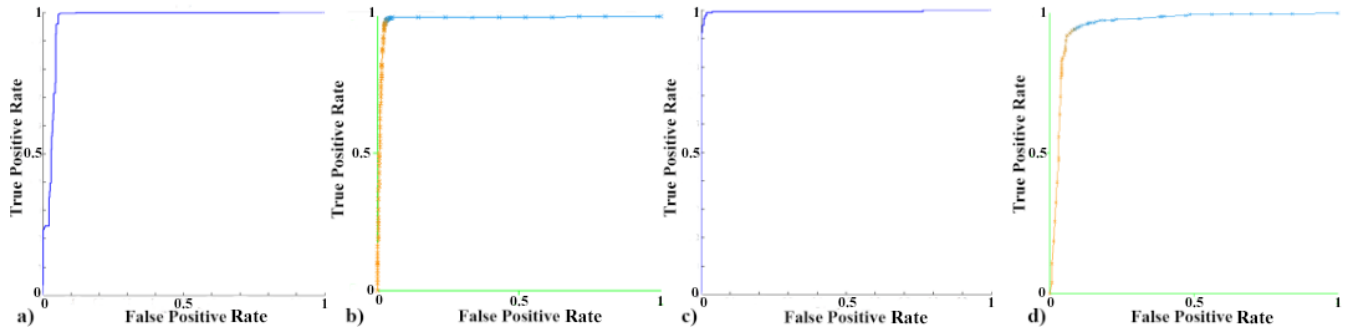

Fig. 2. shows the ROC curves generated with the testing set of the (a) NN. (b) DT, after using 3 PCA features from the *M1/M2* segment; Adamantine-resistance was being analyzed. It also shows the ROC curves generated with the testing set of the (c) NN. (d) DT, after using 40 PCA from the *NA* segment; Oseltamivir-resistance was being analyzed. The x-axis & y-axis are the false positive & true positive rates.

VII. CONCLUSION

The research shows the ability of PCA to prevent overfitting with biased data when classifying Oseltamivir-resistance using NNs. The bias is caused by the number of inputs exceeding the training instances. PCA decreases the complexity of the NN created, by decreasing the number of inputs & improving NN generalization. It also allows the sequences to have more than 4 possible values via projection while compressing the data. This renders detecting subtle differences in biased & smaller datasets possible.

DTs trained with PCA features performed more accurately when detecting Adamantane-resistance compared to NNs. Contrarily, NNs achieved better results when detecting Oseltamivir-resistance; the NNs were trained with a smaller, unbalanced dataset. Thus NNs can be used with PCA for classifying smaller or unbalanced viral datasets, & DTs can be used with PCA for larger & more balanced datasets, to achieve optimal results.

Determining the resistance of Influenza-A to antivirals using laboratory techniques can be time consuming. The techniques above can be used to predict the resistance or susceptibility of a virus to Adamantane or Oseltamivir. This can be used to predict whether the drugs should be utilized or avoided during outbreaks. By using PCA in combination with DTs or NNs, an educated guess can be taken swiftly during emergencies while Influenza-A antiviral resistance is determined manually in research labs.

Both Adamantane & Oseltamivir resistance were detected separately in the previous experiments. In some cases an antiviral strain can be resistant to both drugs [4]. In future work, a multilabel classifier that can detect both Adamantane & Oseltamivir resistance can be built to account for this. The binary classifiers for each antiviral drug, trained on their target viral segment, can be combined to achieve this as in [16]. Alternatively, the *NA* segment can be used to train a BPMLL, a multilabel backprogation NN as shown in [17] to achieve this. The multilabel classifier can be expanded further to a multi-output classifier that measures the viral strain's virulence. This way outbreaks with high morbidity & mortality rate can be detected while determining the antiviral drug to eliminate it if possible.

REFERENCES

[1] Dahran, Nila, Gubareva, Larisa, Meyer, John Meyer *et al*. "Infections With Oseltamivir-Resistant Influenza A(H1N1) Virus in the United States," in *JAMA*, vol. 301 no. 10, 2009, pp. 1034-1041.

[2] Bouvier, N. M., and Palese, P. (2008). "The biology of Influenza viruses" in *Vaccine*, 26, 2008, pp. D49-53.

[3] Ghedin, E., Sengamalay, N., Shumway, M., Zaborsky, J., Feldblyum T., *et al*. "Large-scale sequencing of Human Influenza reveals the dynamic nature of viral genome evolution," in *Nature*, 2005, 437, pp. 1162–6.

[4] "Sheu1, T. G., Fry A. M., Garten R. J., Deyde V. M., Shwe T. *et al*. "Dual Resistance to Adamantanes and Oseltamivir Among Seasonal Influenza A(H1N1) Viruses: 2008–2010," in *The Journal of Infectious Diseases*, vol. 203, no. 1, 2010, pp. 13-17.

[5] Saeys Y., In˜aki I. & Pedro L. "A review of feature selection techniques in Bioinformatics," in *Briefings in Bioinformatics*, vol. 23, no. 19, 2007, pp. 2507-2517.

[6] Shaltout, N. A., Mohamed Moustafa, ElHefnawi, M., Rafea, A., & Moustafa, A. "Comparing PCA to Information Gain as a Feature Selection Method for Influenza-A Classification," in *JBINS*, vol. 1, no. 1, 2015, pp. 48-54.

[7] Leung, K. S., Lee, KH. *et al*. "Data mining on DNA sequences of Hepatitis B virus," in *IEEE/ACM Trans Comput Biol Bioinform*, vol. 8, no. 2, 2011, pp. 428-40.

[8] Shaltout, N. A., ElHefnawi, M., Rafea, A., and Moustafa, A. "Using Information Gain to Compare the Efficiency of Machine Learning Techniques when Classifying Influenza Based on Viral Hosts, *Transactions on Engineering Technologies*, 2015, pp. 707-722.

[9] Shaltout, N. A. *Improving machine learning techniques for Influenza-A classification*, M.S. thesis, Dept. of CSCE, AUC, Cairo, Egypt, 2014.

[10] Attaluri, P. K. "Classifying Influenza Subtypes and Hosts using Machine Learning Techniques," in *ProQuest, UMI Dissertation Publishing*, 2012.

[11] ElHefnawi M., Kadah, Y.M. and Sherif, F. "Influenza A subtyping and host origin classification using Profile Hidden Markov Models," in *Journal of Mechanics in Medicine and Biology*, vol. 12, no. 2, 2012.

[12] ElHefnawi M., Kadah, Y.M., and Sherif, F. "Accurate classification and Hemagglutinin amino acid signatures for Influenza A virus host origin association and subtyping. ," in *Virology*, vol. 449, 2014.

[13] Abdi N. H., and Williams, L. J. (2010). "Principal component analysis," in *Wiley Interdisciplinary Reviews: Computational Statistics*, vol. 2, no. 4, 2010, pp. 433–459.

[14] Thorlund, K., Awad, T., Boivin, G., Thabane, L. (2011). "Systematic review of influenza resistance to the neuraminidase inhibitors," in *BMC Infectious Diseases*, vo1. 11, no. 1, 2011.

[15] Jing, X., Ma, C., Ohigashi, Y. *et al*. "Functional studies indicate Amantadine binds to the pore of the influenza-A virus M2 proton-selective ion channel," in *Proc. Natl. Acad. Sci. U.S.A*, vol. 10, no. 301, 2008, pp. 10967–10972.

[16] A. C. P. L. F. De Carvalho and A. A. Freitas, "A Tutorial on Multi-label Classification Techniques," Studies in *Foundations of Computational Intelligence*," vol. 5, 2009, pp. 177–195.

[17] M.-L. Zhang and Z.-H. Zhou. Multi-label neural networks with applications to functional genomics and text categorization. " in *IEEE Transactions on Knowledge and Data Engineering*, vol. 18, no. 10, 2006, pp. 1338-1351.

ISBN: 978-988-14048-2-4
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)