Improving Prediction Accuracy of Influenza-like Illnesses in Hospital Emergency Departments

Jiefu Pei, Bo Ling and Simon Liao*
Department of Applied Computer Science
University of Winnipeg
Winnipeg, Canada

Baiyan Liu and Jimmy Xiangji Huang*
School of Information Technology
York University
Toronto, Canada

Trevor Strome, R. Lobato de Faria and Michael G Zhang*
Seven Oaks General Hospital
Winnipeg Regional Health Authority
Winnipeg, Canada

*Contact authors: Michael Zhang, Jimmy Huang, Simon Liao. Email: gzhang@wrha.mb.ca, jhuang@yorku.ca, s.liao@uwinnipeg.ca

***Abstract*—Influenza poses a significant risk to public health, as evident by the 2009 H1N1 pandemic. Hospital emergency departments monitor infectious diseases such as influenza with surveillance systems based on arriving chief complaints. However, existing systems are too reliant on the completeness of data and are not acceptably accurate in a practical setting. To improve prediction accuracy, we propose a data cleaning process for data collected in hospital settings. Besides, we also propose a novel feature selection method called the Importance Contribution Index (ICI) which is based on orthogonal transformation. Various feature selection and pattern classification approaches are analyzed. The ICI and C4.5 decision tree are eventually adopted in the new surveillance system. Validation results have shown that the total accuracy has been improved by 7.1% in the enhanced system.**

Keywords—flu prediction; pattern classification; AI; C4.5 Decision Tree; ICI

I. INTRODUCTION

Influenza is an important public health issue. An emergency department (ED) based influenza surveillance system, Emergency Department Information System (EDIS), is in service in six hospitals in Manitoba. EDIS monitors the spread of the influenza virus in order to steer resource allocation, clinical development, and disease control planning. Reports based on data gathered from those systems estimate that 23% and 27% of all emergency department visits in June and November of 2009 were due to Influenza-like Illnesses (ILI). However, the Manitoba provincial laboratory later found that the number of influenza cases were fewer than what was identified by those systems [1], meaning some Non-ILI patients were misidentified as ILI patients. Main factors reducing the performance of surveillance systems include data quality and inadequately selected feature sets heavily reliant on patient arriving chief complaints.

To increase prediction accuracy, a new flu alert system is proposed in this research. Three feature selection algorithms and three pattern classification methods are compared and analyzed following raw data cleaning. The combination that resulted in optimal accuracy and performance is implemented into the new system. Fig.1 shows the system structure.

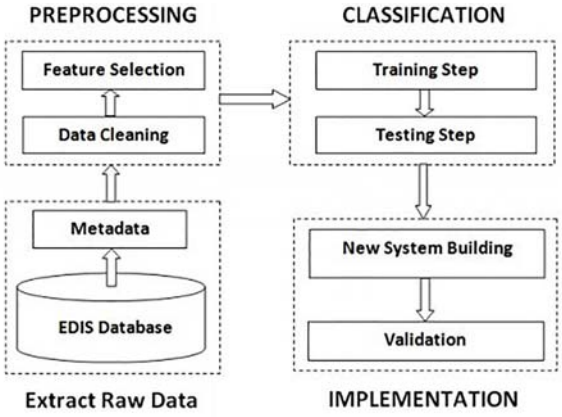

Fig. 1. Pattern Classification System Structure

Our major contributions of this paper are as follows. First, a raw data cleaning process is devised in this research to reduce data noise and improve data quality in data from EDIS. Second, a novel feature selection method called ICI, which is based on orthogonal, is proposed.

II. RELATED WORK

Pattern classification is taking raw data and making a decision action based on categories of the pattern [2]. Recently, popular classification models have been applied in influenza identification and other medical applications.

Heil et al applied neural networks to distinguish seasonal human H1N1, North American swine H1N1 and the 2009 H1N1 pandemic[3]. C.K See et al demonstrated that artificial neural network (ANN) are more efficient than the Gaussian mixture model (GMM) in identifying diabetes type-2 subjects [4]. Yao et al used support vector machines (SVM) in the pulmonary infection detection to identify early stage infections

such as novel H1N1 influenza [5]. Sandra et al presented that basic care sensitivity analysis from a decision tree model found treatment of ILI with oseltamivir more cost-effectiveness than usual care [6]. Paola et al used dynamic Bayesian networks and discovered the interplay of the data source monitored by influenza surveillance during the Severe Acute Respiratory Syndrome (SARS) epidemic period [7]. Wong et al applied Bayesian networks in a disease outbreak detection system and demonstrated that it was able to detect outbreaks in simulated data in early stages while keeping a low false positive count[8].

III. OUR APPROACH

Raw data of more than 20000 patient samples are extracted from the EDIS database. Each sample includes features such as length of stay, chief complaint, age, and gender. In this research, preprocessing consists of data cleaning and relevant feature selection.

A. Raw Data Cleaning

In hospital emergency department settings, patient care is prioritized over data accuracy, often leading to incorrect or missing information. Additionally, other issues such as hard to describe or vaguely described symptoms and inconsistent usage of shorthand further degrade data quality. Data noise such as these must be reduced.

First, patient samples from Children's Hospital are ruled out since the emergency department patient triage system in this hospital differs from the others. From the remaining instances, those meeting any of the following three conditions are selected: "triage chief complaint" matches one of the WHO flu patient criteria; "patient visit reason" contains flu-related keywords such as "FLU", "influenza", or "H1N1"; ILI clinical assessments are available. In terms of noise, there are three types basically. Illegal symbol such as "?", "/", "()" and "≥" are moved. Unit such as "degree Celsius" and "beats/min" are ruled out leaving only numeric part. Errors in patient temperature are revised. For values beyond the normal range (35~40), if they are between 350 and 400, then recalculate the values as one tenth of the original values. Otherwise remove those values. For missing data, features not present in over 50% of the remaining instances are removed, otherwise default values (mean value for numeric and "unknown" for non-numeric) are substituted for missing values. The original chief complaint is expanded into the nine most frequent complaints, cough, weakness, fever, headache, minor complaint unspecified, nausea, shortness of breath, sore throat, URTI, and "other" for low frequency complaints. Each sample is assigned a 1 or 0 for each complaint based on presented complaints. Dduplicated rows and rows where the data contradict each other are removed. Lastly, the data is normalized using (1):

$$T_{norm} = \frac{T_{original} - T_{min}}{T_{max} - T_{min}} \quad (1)$$

Fig.2 illustrates the statistical properties of the treated data set prepared for feature selection and classification.

Features	Maximum	Minimum	Means	StdDev
Length of Stay	1	0	0.247	0.332
Weakness	1	0	0.258	0.438
Cough	1	0	0.155	0.362
Other Complaints	1	0	0.020	0.142
Fever	1	0	0.065	0.247
Headache	1	0	0.012	0.107
Minor Complaint	1	0	0.013	0.113
Nausea	1	0	0.010	0.098
Shortness of Breath	1	0	0.344	0.475
Sore Throat	1	0	0.115	0.319
URTI	1	0	0.007	0.085
Airway	1	0	0.771	0.420
Triage CC	1	0	0.960	0.196
Allergy	1	0	0.297	0.449
Age	1	0	0.570	0.371
Gender	1	0	0.449	0.497
Heart Rate	1	0	0.705	0.239
CTAS	1	0	0.595	0.225
Breathing	1	0	0.715	0.451
Temperature	1	0	0.116	0.228
Respiratory Rate	1	0	0.422	0.287
MRSA/VRE Screening	1	0	0.148	0.355
Circulation	1	0	0.655	0.475
CD Exposure	1	0	0.021	0.143

[a] Mean: average values; StdDev: standard deviation of every feature.

Fig. 2. Statistical List of Preprocessed Data

B. Our Proposed Feature Selection Method

Features do not contribute equally to the prediction of influenza. The most contributing features will be chosen to assist the system formulating a robust prediction rule. In this research, importance contribution index (ICI), information gain and wrappers feature selection are used. The first two methods are independent of the classification models [9], while the wrapper feature selection, combines both steps together.

1) Importance Contribution Index

Importance Contribution Index (ICI) is a novel feature selection approach proposed in this research. The main idea is to transform the original coordinates to a new orthogonal space, map the converted vectors onto the new coordinate axis, and calculate the contributing values in the new space. Finally, process the reversed transform to obtain the values in the original space. Instead of the linear combination of the original vectors, ICI is the contributing value of each individual feature.

In a large data set, several features may move together and be driven by a specific driving force. Technically, there are only a few such driving forces that can manipulate data distribution. Principal component analysis (PCA) generates a new set of factors and every principal component is the linear combination of original data vectors [10].

For space transformation, mean subtraction of the original data is applied, then the covariance between every two features is calculated.The covariance matrix is expressed as

$$\begin{bmatrix} cov(x_1,x_1) & \cdots & cov(x_1,x_n) \\ \vdots & \ddots & \vdots \\ cov(x_n,x_1) & \cdots & cov(x_n,x_n) \end{bmatrix}$$

The general form of PCA is

$$\begin{bmatrix} cov(x_1,x_1) & \cdots & cov(x_1,x_n) \\ \vdots & \ddots & \vdots \\ cov(x_n,x_1) & \cdots & cov(x_n,x_n) \end{bmatrix} \begin{bmatrix} e_{11} & \cdots & e_{1n} \\ \vdots & \ddots & \vdots \\ e_{n1} & \cdots & e_{nn} \end{bmatrix}$$

$$= \left(\begin{bmatrix} \mu_1 & & \\ & \ddots & \\ & & \mu_n \end{bmatrix} \begin{bmatrix} e_{11} & \cdots & e_{1n} \\ \vdots & \ddots & \vdots \\ e_{n1} & \cdots & e_{nn} \end{bmatrix} \right)^T = A^T \quad (2)$$

where μ_i is eigenvalue of the covariance matrix, and $[e_{1i} \quad \cdots \quad e_{ni}]^T$ is the corresponding eigenvector.

Then the transformed features is a multiplication of the original features and a loading matrix as follows:

$$\begin{bmatrix} T_1 \\ T_2 \\ \vdots \\ T_n \end{bmatrix} = \begin{bmatrix} e_{11} & \cdots & e_{1n} \\ \vdots & \ddots & \vdots \\ e_{n1} & \cdots & e_{nn} \end{bmatrix} \begin{bmatrix} O_1 \\ O_2 \\ \vdots \\ O_n \end{bmatrix} \quad (3)$$

where

$$\begin{bmatrix} O_1 \\ O_2 \\ \vdots \\ O_n \end{bmatrix} = \begin{bmatrix} o_{11} & \cdots & o_{1k} \\ \vdots & \ddots & \vdots \\ o_{n1} & \cdots & o_{nk} \end{bmatrix}$$

represents the original features,

$$\begin{bmatrix} T_1 \\ T_2 \\ \vdots \\ T_n \end{bmatrix} = \begin{bmatrix} t_{11} & \cdots & t_{1k} \\ \vdots & \ddots & \vdots \\ t_{n1} & \cdots & t_{nk} \end{bmatrix}$$

corresponds to the converted features, and

$$\begin{bmatrix} e_{11} & \cdots & e_{1n} \\ \vdots & \ddots & \vdots \\ e_{n1} & \cdots & e_{nn} \end{bmatrix}$$

is the loading matrix which is composed of eigenvectors.

Therefore, t_{ij} can be obtained with the following equation:

$$t_{ij} = \sum_{\alpha=1}^{n} e_{i\alpha} \cdot o_{\alpha j} \qquad i = 1,2,3 \dots n; j = 1,2,3 \dots k \quad (4)$$

where n is the number of features (clinical indicators), k is the number of instances (patient samples), and t_{ij} is the projected value on component i of instance j .

Then the relative contribution of the original instances to the components is calculated by

$$p_i = \frac{(t_{i1} + t_{i2} + \cdots + t_{ik})^{\frac{c \cdot \mu_i}{\mu_1 + \mu_2 + \cdots + \mu_n}}}{k} \quad (5)$$

where c is a constant and μ_i is the corresponding eigenvalues for each eigenvectors.

The contributing values are obtained in the new space, and ICI values are gained by applying the reversed space transformation

$$\begin{bmatrix} ICI(O_1) \\ ICI(O_2) \\ \vdots \\ ICI(O_n) \end{bmatrix} = \begin{bmatrix} b_{11} & \cdots & b_{1n} \\ \vdots & \ddots & \vdots \\ b_{n1} & \cdots & b_{nn} \end{bmatrix} \begin{bmatrix} p_1 \\ p_2 \\ \vdots \\ p_n \end{bmatrix} \quad (6)$$

where

$$\begin{bmatrix} b_{11} & \cdots & b_{1n} \\ \vdots & \ddots & \vdots \\ b_{n1} & \cdots & b_{nn} \end{bmatrix}$$

is an inverse matrix of the loading matrix. With the principle above, 24 features and 3717 samples are applied in Importance Contribution Index algorithm. Fig.3 lists the ranking based on adjusted ICI values.

2) Information Gain Based Feature Selection

Entropy and information gain (IG) are commonly used in medical data feature selection. MacKay defined entropy as a measure of the randomness associated with a random variable [11]. Suppose X is a random variable which has the possible values of $\{x_1, x_2, x_3 \dots x_n\}$, and each value has its corresponding probability $\{p_1, p_2, p_3 \dots p_n\}$. Then the entropy is defined as:

$$H(X) = -\sum_{i=1}^{n} p_i \cdot log_2 p_i \quad (7)$$

Suppose we have another sequence T with the possible values of $\{t_1, t_2, t_3 \dots t_m\}$, and the conditional entropy [12] is defined as

Features	ICI Values	Features	ICI Values
Minor Complaint	5.46	Temperature	3.05
Nausea	3.04	Fever	2.97
Cough	2.87	Weakness	2.84
URTI	2.82	Headache	2.79
CTAS	2.70	Age	2.45
Circulation	2.43	CD Exposure	2.42
Shortness of Breath	2.38	Triage	2.38
Resp Rate	2.36	Breathing	2.23
Allergy	2.21	Heart Rate	2.15
Airway	2.01	MRSA/VRE Screening	1.99
Sore Throat	1.98	Gender	1.75
Other Complaints	1.60	Length of Stay	0

Fig. 3. Features Ranking Based on ICI Values

Features	IG Values	Features	IG Values
Cough	0.208	Age	0.180
Weakness	0.151	Length of Stay	0.143
Triage CC	0.060	Temperature	0.054
CTAS	0.051	MRSA/VRE Screening	0.036
Shortness of Breath	0.031	Fever	0.026
CD Exposure	0.022	Minor Complaint	0.019
Heart Rate	0.018	Headache	0.017
Nausea	0.014	Other Complaints	0.012
Resp Rate	0.010	Sore Throat	0.004
Allergy	0.003	Breathing	0.003
Circulation	0.003	Gender	0.002
Airway	0	URTI	0

Fig. 4. Information Gain Values of All Features

$$H(X|T) = \sum_{i=1}^{m} p(T = t_i) \cdot H(X|T = t_i) \tag{8}$$

Information gain (IG) is the difference between entropy and conditional entropy

$$IG(X|T) = H(X) - H(X|T) \tag{9}$$

3717 samples are applied in the entropy and information gain theory. IG values of all features were obtained as Fig.4 illustrates.

3) Wrappers Features Selection

Wrappers feature selection combines feature selection and classification together. A special search algorithm is "wrapped" around the classification model to search the space of possible feature subsets [13]. In this research, greedy-stepwise and best-first feature selection are applied with various strategies such as forward, backward and bi-directional search.

Greedy-stepwise algorithm, in each iteration, searches all remaining individual features to select or remove the unique one to achieve an optimal result. The process is repeated until system performance cannot be improved by increasing or decreasing features. Unlike greedy-stepwise, best-first feature selection provides backtrack when the search process encounters a bottleneck. In this research, the number of backtracks is limited to three in order to avoid the full space search.

IV. CLASSIFIERS

Three different classifiers, artificial neural network (ANN), optimization parameters based support vector machine (OPSVM), and C4.5 decision tree, are used to detect patterns of given subsets and assign each sample with a class. The classifiers evaluate the effectiveness of feature selection algorithm. Within each classifier, features selected by all three feature selection methods are utilized for training and testing.

A. Artificial Neural Network

Artificial neural networks rely on propagating input through hidden layers of nodes for classification and have been used in cancer diagnosis systems. A two hidden-layer Artificial Neural Network is applied to enhance the disturbance resistance of the system.

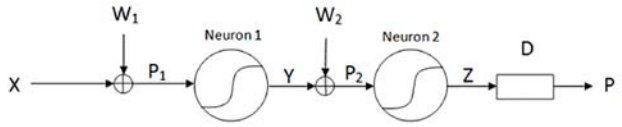

Fig. 5. Working Principle

Gradient descent is applied to optimize the transfer parameter matrix [2]. Suppose an input X and an output Z are connected by two single neurons, with performance P, as Fig.5 shows. Let $P = (Z - D)^2$, where D is desired output, or training label. The gradients are defined as $\frac{\partial P}{\partial w_2}$ and $\frac{\partial P}{\partial w_1}$.

The discriminant function in gradient descent is a sigmoid function

$$f = \frac{1}{1 + e^{-x}} \tag{10}$$

The Gradient Descent used for weight adjustment is obtained as shown in (11) and (12)

$$\frac{\partial \mathbf{P}}{\partial \mathbf{W}_2} = 2(\mathbf{Z} - \mathbf{D})f(1 - f)\mathbf{Y} \tag{11}$$

$$\frac{\partial \mathbf{P}}{\partial \mathbf{W}_1} = 2(\mathbf{Z} - \mathbf{D})f(1 - f)\mathbf{W}_2F(1 - F)\mathbf{X} \tag{12}$$

where each bold letter represents a parameter vector. $\mathbf{W_1}$ and $\mathbf{W_2}$ represent weight matrices. F is the discriminant function of neuron 1 and f is the discriminant function of neuron 2. The number of neurons for each hidden layers is 12 and 5, which is based on training and testing results.

Suppose m is the current training and (m-1) is the last training, then the weight adjust algorithm is

$$\frac{\partial \mathbf{P}(\mathbf{m})}{\partial \mathbf{W}_2(\mathbf{m})} = -2\eta((\mathbf{Z}(\mathbf{m}) - \mathbf{D}(\mathbf{m}))f(1 - f)\mathbf{Y}(\mathbf{m})) + \frac{\partial \mathbf{P}(\mathbf{m}\text{-}1)}{\partial \mathbf{W}_2(\mathbf{m}\text{-}1)} \tag{13}$$

$$\frac{\partial \mathbf{P}(\mathbf{m})}{\partial \mathbf{W}_1(\mathbf{m})} = -2\eta((\mathbf{Z}(\mathbf{m}) - \mathbf{D}(\mathbf{m}))f(1 - f)\mathbf{W}_2(\mathbf{m})F(1 - F)\mathbf{X}(\mathbf{m})) + \frac{\partial \mathbf{P}(\mathbf{m}\text{-}1)}{\partial \mathbf{W}_1(\mathbf{m}\text{-}1)} \tag{14}$$

where α is the momentum coefficient and η is the learning rate.

Momentum coefficient is used to increase the speed of convergence and different learning rates may result in different system performance on a specific data set. The optimal learning rate leads to the local error minimum in the current learning step [2].

B. Parameters Optimization Based Support Vector Machine

The basic theory of support vector machines (SVM) is to maximize the classification margin to determine the optimal hyper-plane. Intuitively, such classifications will be valid for those testing data that are close, but not necessarily identical to the training data [14].

The margin of a linear classifier is the width that the boundary could be increased by before hitting a data point [15]. As Fig.6 shows, maximum margin determines the optimal classifier.

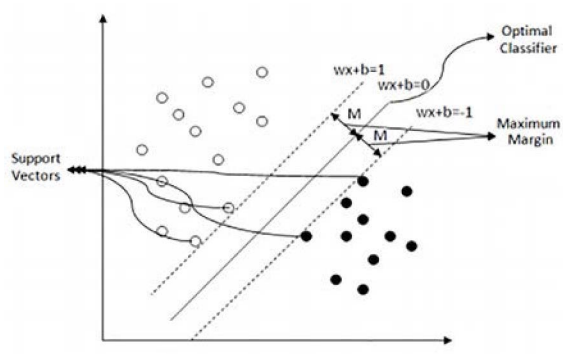

Fig. 6. Optimal Classifier with Maximum Margin

Suppose x+ and x- are two points on hyper-plane wx + b = 1 and wx + b = −1 respectively, and β is a parameter that satisfies

$$x^+ = x^- + \beta\omega \tag{15}$$

Then the margin M can be expressed as

$$M = \frac{1}{2}(x^+ - x^-) = \frac{\beta|\omega|}{2} = \frac{1}{|\omega|} \tag{16}$$

However, in practical classification problems with noise, maximizing margin to seek for the safest classifier and minimizing the classification errors, need to be considered simultaneously. The solution is to minimize

$$T = |\omega| + \sum_{k=1}^{R} CE_k \tag{17}$$

where C is a parameter about the penalty and R is the number of inseparable points.

Mapping the data to a higher dimension alleviates the difficulty and can be done with a kernel function:

$$K(x_j, x_k) = \left(\varphi(x_j)\right)^T \varphi(x_k) \tag{18}$$

Widely-used kernel functions include the radial basis function (RBF), which is defined as [15]

$$K(x_j, x_k) = e^{\frac{-|x_j - x_k|^2}{2p^2}} \tag{19}$$

and the polynomial function. which is defined as [16]

$$K(x_j, x_k) = (x_j \cdot x_k + 1)^d \tag{20}$$

where *d* is the degree of the polynomial function.

In this research, the optimal parameter for RBF kernel is 0.05 and 2.00 for polynomial kernel.

C. C4.5 Decision Tree

Decision trees is a special kind of classification method that is able to find subtle differences in real world data [17]. Within decision trees, training involves building an integrated tree starting from the root, and pruning based on calculating the estimated error for new samples.

In C4.5 decision tree [18] training, node split and stop split are two key steps. A sub-tree down to node N with entropy impurity I(N)

$$I(N) = -\sum_j p(\omega_j)\log_2 p(\omega_j) \tag{21}$$

where $p(\omega_j)$ is the proportion of patterns at node N in category ω_j. Then a heuristic way to split is to consider the decreases of the entropy impurity

$$\Delta I(N) = I(N) - P_L I(N_L) - (1 - P_L) I(N_R) \tag{22}$$

where N_L and N_R are left and right sub-node respectively, $I(N_L)$ and $I(N_R)$ are their entropy impurity.

The optimal query is determined by the largest $I(N)$, whichmeans the split provides most entropy drop [2]. The stopping condition of the splitting process is when the sub-tree consists of a single leaf node or if there was no change in $\Delta I(N)$.

The degree of prune is determined by the value of the confidence factor CF

$$CF = \sum_{i=0}^{E} p^i (1-p)^{N-i} \tag{23}$$

where CF is confidence factor, N is number of cases and E is number of errors. A larger CF should result in less pruning. In this research, 0.25 is the optimal confidence factor.

V. Experiment Settings and Results

A. Experimental Settings

In this experiment, a computer with processor of Intel(R) Core(TM) i3 CPU 2.40GHz and memory of 4.00 GB is utilized for testing. Feature selection methods with each of the three classifiers are applied to 3717 data samples. For IG and ICI feature selection methods, sensitivity, specificity, and classification accuracy are compared by adding features to the selected subset one at a time in descending order of relevance. Specifically, sensitivity is the percentage of identified ILI patients within all diagnosed influenza patients. Specificity is the percentage of identified Non-ILI patients within all diagnosed Non-ILI patients. Accuracy is percentage of correctly detected patients in all ED visits. For best-first, forward, backward, and bi-directional search are used. Forward and backward search are applied to greedy-stepwise. For ANNs, learning rates of 0.01, 0.05 and 0.1 to 1.0 in increments of 0.1 are tested, and it is found that the optimal learning rate for the given data sets is 0.1. In order to evaluate the overall performance of a statistical classifier, k-fold cross-validation is used [19].

TABLE I. RESULTS COMPARISON

Classifier	Feature selection	Number of features	Sensitivity (%)	Specificity (%)	Accuracy (%)	Time cost (sec)
ANN	IG	20	74.1	96.4	88.2	754
	ICI	17	75	95.7	88	738
	Wrappers	19	76	96.3	89	>172800
SVM-poly	IG	20	75	97.4	89	8378
	ICI	19	76.6	96.8	89.3	8378
	Wrappers	19	76.8	97.3	89.7	>345600
SVM-RBF	IG	15	67	96	85.3	4333
	ICI	14	67	96	85.3	4214
	Wrappers	19	67	96	85.3	>172800
C4.5 Decision tree	IG	20	76.5%	96.4	89	74
	ICI	17	77.5	95.3	89	63
	Wrappers	14	77	97.4	90	>21600

B. Results

The results based on optimal combination of sensitivity, specificity, accuracy, selected number of features and time consumption for the three different classifiers and four feature space searches are listed in Table I.

VI. CONCLUSIONS AND FUTURE WORK

As shown in Table II, the C4.5 decision tree achieves acceptable accuracy in the least time. The accuracy from IG and ICI selected feature sets are similar, ICI performs better with respect to the number of selected features and running time. Therefore, C4.5 decision tree and ICI selected optimal subset are chosen for the influenza alert system.

In addition, two different levels of realistic validation are performed. For high level validation, the prediction algorithm is applied to more than 100000 emergency visits in Seven Oaks General Hospital, St Boniface General Hospital, Health Science Center Adult Department, Concordia Hospital, Grace Hospital and Victoria General Hospital in the fiscal year of 2009-2010. The new system is able to detect the H1N1 pandemic peaks which that are identified by Ginsberg et al [20]. Moreover, the number of cases identified as ILI by the new system is less than the original system, which is more in-line with provincial laboratory-confirmed influenza cases. For low level validation, 476 individual cases from Seven Oaks General Hospital are reviewed. Validation comparison shows the new system correctly identifies 31 Non-ILI cases more than the original one, and total accuracy is improved by 7.1%.

The scope of this experiment is limited since the data available is limited. Therefore, there is potential for improvements to prediction sensitivity. To deploy this system in emergency rooms and monitor patients in real-time is another goal, but requires real-time data, which is not available.

Influenza alert system enhanced with ICI and decision tree classification algorithms are deployed by the WRHA emergency program in the summer of 2011.

ACKNOWLEDGMENT

This research is supported by the Seven Oaks General Hospital and in part by a research grant from Natural Sciences and Engineering Research Council of Canada (NSERC). We also thank reviewers for their valuable comments on this paper.

REFERENCES

[1] Provincial Laboratory, Influenza in Manitoba, 2009/2010 Season, July 1, 2010

[2] R. O. Duda, P. E. Hart, D. G. Stork, Pattern Classification, Second Edition, 2001

[3] G. L. Heil et al, "MChip, a low density microarray, differentiates among seasonal human H1N1, North American swine H1N1,and the 2009 pandemic H1N1", Influenza and Other Respiratory Viruses 4(6). pp, 411-416, 2010

[4] C. K. See et al, "Automated identification of diabetes type-2 subjects with and without neuropathy using eigenvalues", Proc. IMechE Vol. 224 Part H: Journal Engineering in Medicine, pp. 1-10, 2009

[5] J. Yao and A. Dwyer, "Computer-aided diagnosis of pulmonary infections using texture analysis and support vector machine classification", Academic Radiology, Vol 18, pp. 306-314, 2011

[6] S. E. Talbird, A. J. Brogan, A. P. Winiarski and B. Sander, "Cost-effectiveness of treating influenzalike illness with oseltamivir in the United States", American Journal of Health-System Pharmacy 66. pp. 469-480, 2009

[7] P. Sebastiani, K. D. Mandl, P. Szolovits, I. S. Kohane and M. F. Ramoni, "A Bayesian dynamic model for influenza surveillance", Statistics in Medicine 25(11), pp. 1803-1816, 2006

[8] W-K Wong, A. Moore, G. Cooper and M. Wagner, "Bayesian network anomaly pattern detection for disease outbreaks", In Proceedings of the Twentieth International Conference on Machine Learning, 2003

[9] A. G. K. Janecek, W. N. Gansterer, M. A. Demel and G. F. Ecker, "On the relationship between feature selection and classification accuracy", Journal of Machine Learning and Research: Workshop and Conference Proceedings 4: pp. 90-105, 2008

[10] I. Jolliffe, Principal Component Analysis, Springer-Verlag, New York, 1986

[11] D. J. C. MacKay, "Information theory, inference and learning algorithms", Cambridge University Press, Cambridge. 2003

[12] A. W. Moore, "Information gain", online tutorial: www.autonlab.org/tutorials/infogain11.pdf, 2001

[13] Y. Saeys, "Feature selection for classification of nucleic acid sequences", PhD thesis, Ghent University, Belgium, 2004.

[14] Y. Liu, X. Yu, J. X. Huang and A. An, "Combining integrated sampling with SVM ensembles for learning from imbalanced datasets", Information Processing & Management 47(4) (2011) 617–631

[15] C. J. C. Burges, "A tutorial on support vector machine for pattern recognition", Data Mining and Knowledge Discovery, 2, 121-167, 1998.

[16] C. Cortes and V. Vapnik, "Support vector networks", Machine Learning, vol. 20, pp. 273-297, 1995.

[17] P. K. Attaluri and X. Zheng, "Applying machine learning techniques to classify H1N1 viral strains occurring in 2009 flu pandemic", 6th Annual Biotechnology and Bioinformatics Symposium, 2009

[18] J. R. Quinlan, "C4.5: programs for machine learning", Morgan Kaufmann, 1993

[19] M. Stone, "Cross-validatory choice and assessment of statistical predictions", Journal of the Royal Statistical Society. B 36 (1), pp. 111-147, 1974.

[20] J. Ginsberg et al, "Detecting influenza epidemics using search engine query data", Nature 457