Cost-Sensitive Risk Stratification in the Diagnosis of Heart Disease

Selen Uguroglu and **Jaime Carbonell**
Language Technologies Institute
Carnegie Mellon University
5000 Forbes Avenue
Pittsburgh, PA 15232

Mark Doyle and **Robert Biederman**
Division of Cardiology
Allegheny General Hospital
320 E North Avenue
Pittsburgh, PA 15212

Abstract

We investigate machine learning methods for diagnostic screening of heart disease. Coronary heart disease is the leading cause of death in the US, causing more deaths than all types of cancers combined. Early diagnosis of heart disease in women is harder than it is in men and typically requires the administration of several clinical tests on the patient. Most risk stratification methods aggregate the results of such tests, including the risky, invasive procedures that cannot be administered on all patients. In this paper, our goal is to identify patients who are under high-risk of having heart disease and related adverse events, using a minimal number of diagnostic tests, especially less invasive ones. The low frequency of patients with severe heart disease in the dataset is challenging for most conventional machine learning methods. To overcome this problem, we develop and apply a cost-sensitive k nearest neighbor algorithm. Our contributions are two fold: First, we compare the predictive value of several diagnostic procedures for heart disease, including electrocardiography, angiography, radionuclide perfusion and conclude that in womens heart disease, certain combinations of non-invasive techniques are more predictive than some of the widely used invasive procedures. Then, we evaluate held out data and achieve an AUROC over 0.70, signifying valuable clinical utility, using only the least costly and least invasive tests.

Introduction

According to the heart disease and stroke statistics in the US, annually one in every six deaths is caused by coronary heart disease (CHD) (Lloyd-Jones et al. 2010). For women, the mortality rate is even higher, approximately one in every four women dies of the complications caused by coronary heart disease (Stangl et al. 2008). Several diagnostic tests exist to detect the disease, but yet, 64% of women die suddenly from an adverse event related to CHD, without showing any symptoms of disease prior to the event (Lloyd-Jones et al. 2010). Women who have symptoms such as chest pain and angina, are referred to coronary angiography, an invasive procedure that involves threading a catheter into heart, costing thousands of dollars per patient, yet half of them don't have heart disease (Davis et al. 1995).

In order to enhance the understanding of clinical presentations of ischemic heart disease in women, WISE (Womens Ischemic Syndrome Evaluation) was initiated in 1996 as a part of a National Heart, Lung and Blood Institute sponsored clinical trial (Bairey Merz et al. 1999). During this study, 936 women with suspected ischemic heart disease underwent various diagnostic tests and were followed-up at six weeks and then annually to assess symptoms, hormonal status and prognostic indicators of adverse events (Bairey Merz et al. 1999).

The process of diagnosing heart disease is typically as follows: the simplest and least costly measurements are collected from the patient with the suspected heart disease, along with her treatment history and a questionnaire regarding of her illness. The simplest baseline measurements can be blood pressure taken from an arm cuff, weight, height etc. Patients who are suspected to have the disease are administered additional, more specialized tests, which are more costly and possibly more invasive. The decision regarding the next stage diagnostic testing is left to the patients health care specialists judgment.

In this paper, we investigate the predictive value of diagnostic tests, focusing on the least costly and least invasive ones, including different combination thereof in assessing risk of the patients, in order to determine which sets of diagnostic tools are the most beneficial for female heart patients. We show that conventional invasive and costly diagnostic tests that have been useful in assessing risk in heart disease in men, are not as effective in women, and the combination of certain non-invasive tests can achieve a higher predictive score. Our method is extremely successful in classifying patients under high risk of having an adverse event with the least invasive and the least costly features when tested on held-out data. With the addition of more costly, but still not invasive diagnostic tests, we achieve an even higher AUROC score, signifying valuable clinical utility. We compare our method to the established risk stratification techniques in the literature and show that our method is significantly more effective. Our results can have a large effect on health care, reducing costs and limiting diagnostic hurdles that the patients may be reluctant to undertake.

Rest of the paper is organized as follows: First, we describe related work, and then, we formalize the problem from the machine learning perspective. After we describe our approach, we present the results of our experiments, comparing our method to the previous approaches. Lastly, we conclude with possible future directions.

Related Work

Risk assessment for heart disease has been an ongoing research topic for several years. The Framingham risk score (FRS) is one of the most commonly used risk assessment technique. As an input, FRS takes the simplest traditional risk factors, such as age, smoking, and blood pressure and outputs whether the patient is under low, intermediate or high risk for up to 10 years, by scoring each risk factor (Mieres et al. 2005). However, this metric is not sufficient for an accurate risk stratification: it classifies more than 90% of women as low risk, and among the remaining 10%, very few patients under 70 are classified as high risk (Shaw, Bugiardini, and Merz 2009). Thus, American Heart Association (AHA) recently defined a new concept of ideal cardiovascular health based on good medical history, the absence of major CVD risk factors, the presence of ideal physical examination results, low 10-year risk scores and healthy lifestyle habits (Mosca et al. 2011). Their risk classification algorithm stratifies women into three risk groups: high-risk, at-risk and optimal-risk. The details of this approach can be found in (Mosca et al. 2011).

Recently machine-learning approaches have been applied to clinical data to diagnose patients and classify them into different risk groups. (Syed and Rubinfeld 2010) formalized clinical risk stratification as anomaly detection problem and applied Minimum Enclosing Ball (MEB) to identify patients who are at an increased risk of adverse events. They applied their method to the MERLIN trial data and to the National Surgical Quality Improvements Program (NSQIP) data and they were able to achieve an AUROC as high as 0.86 for mortality (Syed and Rubinfeld 2010). However, this method ignores the available label information, and it may not be applicable to noisy datasets.

On the clinical datasets, traditional supervised classification methods perform poorly due to class skew: high-risk patients (minority class) are generally much less frequent than the low-risk patients (majority class). These algorithms fail to classify instances belonging to minority class to their correct risk groups, since assigning every instance to the majority class is more favorable when minimizing the loss function.

The imbalanced dataset problem has been tackled via data renormalization and classifier modification. As an example of the latter approach, cost-sensitive SVMs can be given, where SVMs are modified by incorporating a cost-matrix to the soft margin optimization problem to handle class imbalance. (Lessmann 2004), (Akbani, Kwek, and Japkowicz 2004), (Visa 2005). Cost-sensitive learning acknowledges that misclassification costs or feature acquisition costs are not uniform. In the case of medical diagnosis, both types of costs should be modeled: misclassifying disease as healthy can be lethal, whereas a false positive prediction has less severe consequences. Similarly using too many unnecessary tests to diagnose a patient should be penalized as well, since diagnostic tests can be costly, invasive and risky. However, most prior work in cost-sensitive learning addresses misclassification without considering attribute cost (Elkan 2001), or addresses attribute costs without taking into account misclassification costs (Melville et al. 2005). Especially in medical diagnosis two way cost-sensitive learning is fairly unexplored.

Different from the previous approaches, we are interested in modeling the predictive value of the diagnostic tests, particularly in combination, while applying different penalties to two different types of misclassifications to handle imbalance. We apply our technique to the WISE dataset. This is a novel class of applications of our method, which has important value for systemizing the diagnostic heart disease tests for women.

Problem Formulation

We formulate the problem as a binary classification problem. Given labeled training dataset D of n tuples, D = $\{(x_1, y_1), (x_2, y_2), (x_3, y_3), ..., (x_n, y_n)\}$ where $x_i \in \mathcal{R}^p$ are the feature vectors, our goal is to infer binary class labels $y_i \in t_1, t_2$.

K nearest neighbors (KNN) is the supervised learning algorithm which classifies an instance based on the labels of its k closest neighbors in the feature space. Neighbors are found using a distance function, which is chosen based on attribute types in the dataset. In this paper, we used a distance measure appropriate for mixed typed attributes. Given Q as the set of quantitative features and C as the set of categorical features, let L_c be an $M \times M$ symmetric matrix describing the distance between two categorical variables. Using squared distance for quantitative variables, we can calculate the distance between two feature vectors with the following equation:

$$d(x_i, x_j) = \sum_{q \in Q}(x_{iq} - x_{jq})^2 + \sum_{c \in C} L_c(x_{ic}, x_{jc}) \qquad (1)$$

Let N_i be the set of K nearest neighbors for a test instance x_i based on the distance measure d. Using a majority voting scheme, let $V_i(t)$ be the total votes of neighbors of x_i with the label t. More formally,

$$V_i(t) = \sum k \in N_i(I(t, y_k)) \qquad (2)$$

where I is an indicator function, that is I(t, y_k) = 1 if t = y_k, 0 otherwise. A refinement is to perform similarity-weighted voting. Let T the target space, that is T = $\{t_1, t_2\}$, then predicted target variable of x_i is: $\hat{y}_i = \text{argmax}_{t \in T} V_i(t)$

In the case of imbalanced datasets, instances of the majority class dominate the neighborhood; hence the majority vote tends to be the majority class label in most cases. In the cost-sensitive KNN (C-KNN), the class imbalance problem can be addressed with class-based weighting of the votes. Let be $w = \{w_{t_1}, w_{t_2}\}$ be the weight vector corresponding to the class labels t_1 , t_2. The new weighted majority-voting scheme is:

$$\hat{y}_i = \text{argmax}_{t \in T} w_t V_i(t) \qquad (3)$$

With this new weighted majority voting scheme, majority class votes can be penalized by applying lower weights than the minority class votes. Hence, even when the minority class instances are a minority in the k-neighborhood, their presence are amplified by using higher weights.

Data

In the WISE study, female participants who were undergoing coronary angiogram for chest pain or suspected myocardial ischemia, had been assigned further diagnostic testing to understand the clinical representation of coronary artery disease (Bairey Merz et al. 1999). Diagnostic tests can be divided into two groups: invasive and non-invasive tests. Procedures performed without the insertion of needle, instruments or fluids into the body can be considered non-invasive. Invasive procedures range from blood tests (as it involves needles) to surgeries. At the start of the clinical trial, prior to diagnostic testing, baseline evaluation data is collected from the patients. This data included demographic, clinical, angiographic information and Duke Activity Status Index questionnaire inquiring about patients activity levels (Bairey Merz et al. 1999). Their physical symptoms such as the location and severity of pain, was also included in the baseline evaluation (Bairey Merz et al. 1999). Patients were contacted at six weeks periods, and then annually to collect further information regarding their symptom status, hormonal status and adverse event encounter (Bairey Merz et al. 1999). After baseline evaluation, patients underwent several diagnostic tests such as electrocardiogram (ECG), Dobutamine stress tests (DS), pharmacologic stress tests without Dobutamine (PS), angiogram (AN), exercise stress test (EX), radionuclide perfusion (PERF), brachial artery ultrasound (MD). Among these tests, ECG, EX, and MD are non-invasive, and DS, PS, PERF and AN are invasive with ECG being the least invasive and the least expensive test. However, even though it is a costly and invasive procedure, angiogram is seen as the gold standard test for the diagnosis of heart disease.

Preprocessing

Missing value imputation Missing values in clinical trials is a serious problem for analysis and interpretation of data. In this paper, missing values are first imputed by mean value imputation, and then their final values are found by training a linear regression estimator with ridge penalty on the full attribute-patient value matrix.

Feature Selection Feature selection is performed firstly on the initial dataset and then after each additional diagnostic test or tests. The chi-square test for independence is chosen as the feature selection method. Prior to chi-square testing, quantitative features are categorized using level binning. Features that have a p-value of less than 0.05 ($p \leq 0.05$) were kept in the dataset for classification.

Class Labels Patients are labeled based on the events they had through out the study. We considered 4 different types of adverse events: death, congestive heart failure (CHF), stroke, and myocardial infarction (MI).

Test Ordering

We evaluated the different combination of diagnostic test in five stages. Stage 1 uses baseline evaluation features which are augmented with follow up information. In stage 2, features from electrocardiogram (ECG) results are combined with the baseline evaluation features. In stage 3, the three kinds of stress tests, exercise stress (EX), Dobutamine stress (DS), and pharmacologic stress (PS) are evaluated separately by combining the results of each test to the feature set from stage 2. Feature set combinations at this stage are therefore: Augmented baseline evaluation (BE), ECG, DS; BE, ECG, PS; BE, ECG, EX. In stage 4, we separately add angiogram (AN) and perfusion imaging (PERF) results to each feature set from stage 3 obtaining following feature sets: BE, ECG, DS, AN; BE, ECG, PS, AN; BE, ECG, EX, AN; BE, ECG, DS, PERF; BE, ECG, DS, AN; BE, ECG, EX, PERF. Perfusion imaging or angiogram is usually next stage diagnostic tests after stress testing so we evaluated both of them after stage 3. Finally, in stage 5, we added the brachial artery testing results to the previous feature set. Conventionally, brachial artery ultrasound is usually performed after or during angiogram, which is why it is the latest diagnostic test in the process.

Experiments

Among 936 patients, we removed the patients who passed away or dropped out from the trial within the first 6 months of their admission. Follow-up information of the remaining 638 patients is added to their baseline evaluation data, and used as initial feature set. After the inclusion of each diagnostic test, we performed feature selection using chi-square independence testing and standard normalization.

Metrics

Specificity measures the proportion of true negatives in the dataset, whereas, sensitivity, also known as the recall rate, measures the proportion of the true positives. Ideally a diagnostic test has sensitivity and specificity both close to 100%. It can also be the case that a test has a low specificity and high sensitivity or vice versa. In such cases, the choice of the diagnostic test depends on other variables such as the cost of the test or the seriousness of the condition (Pepe 2003). The ROC curve is a plot of sensitivity against specificity of a classifier, as its discrimination threshold is varied. In clinical research, ROC curves are extensively used to evaluate statistical models. Area under ROC curve provides a quantitative measure for the performance of a diagnostic test (Pepe 2003). AUROC score of 0.5 denotes the random classifier, and anything above 0.5 has a predictive value. In this paper, we compared the discriminative value of the diagnostic tests in terms of their specificity and sensitivity. We also provided the AUROC for our approach and other benchmark approaches for various diagnostic tests.

KNN Parameters

For kNN, the value of k and class voting weights are selected empirically on the training set. We used the following

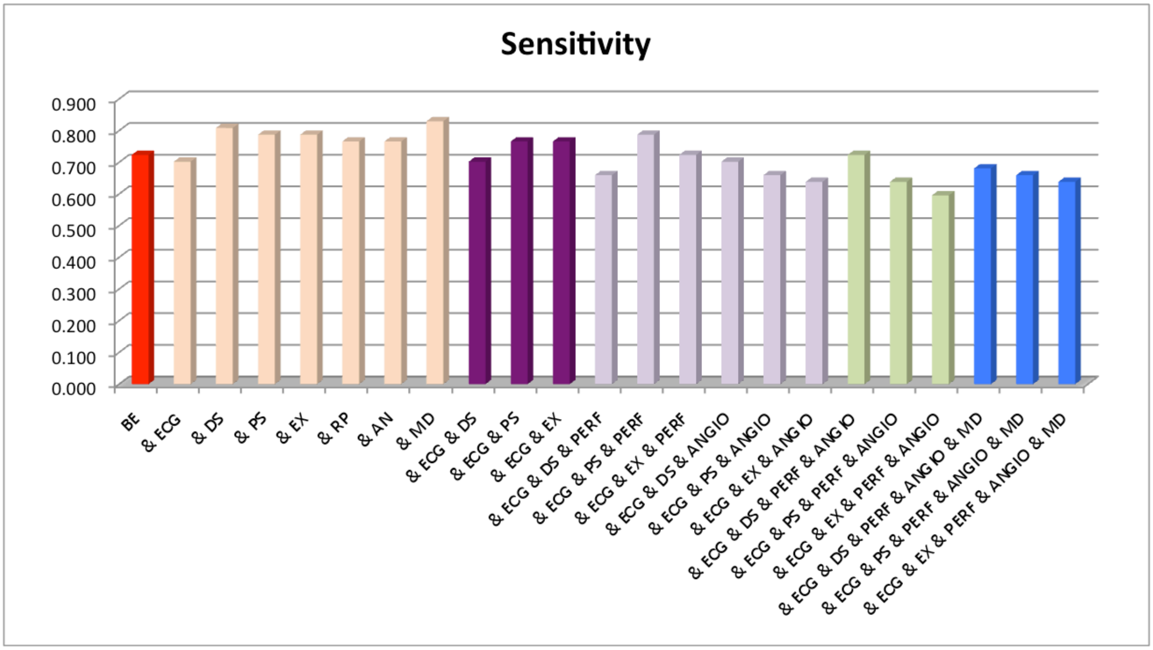

Figure 1: Sensitivity comparison of diagnostic tests and their combinations for mortality prediction using C-KNN. Procedures are ordered with respect to their stages, different colors represent different stages

formulation for class weights:

$$w_0 = 1, w_1 = \left\lceil \frac{K}{2} \right\rceil + 1 \qquad (4)$$

where w_0 is majority class weights (patients who didnt have an event), and w_1 is minority class weights (patients who had an event). $\frac{w_0}{w_1}$ are estimated based on the $\frac{n_0}{n_1}$ where n_0 and n_1 are the number of negative and positive instances in the training set respectively.

Results

Sensitivity of diagnostic procedures Using cost sensitive KNN as the classifier we predicted mortality with the combination of feature sets obtained from each stage. The sensitivity for each combination of diagnostic tests is reported in Figure 1. Different stages are denoted with different colors. As can be seen from Figure 1, baseline evaluation with brachial artery ultrasound performs the best in sensitivity; demonstrating that rather than a combination of several diagnostic procedures, brachial artery ultrasound may be the best diagnostic procedure to administer for patients with high risk factors to identify whether they indeed have an heart disease. A less costly option, dobutamine stress, is the next best alternative, which is rightly used as initial diagnostic procedure. For the general population performing more operations does not significantly improve sensitivity, as different tests may give contradictory results.

Performance of C-KNN AUROC of C-KNN for diagnosing 4 different types of events, death, CHF, MI and stroke is shown in Figure 2. For each of the cases, C-KNN performs extremely well, having an AUROC above 0.65. Especially for mortality and congestive heart failure (CHF) prediction, AUROC is above 0.7, which signifies a valuable clinical diagnostic utility.

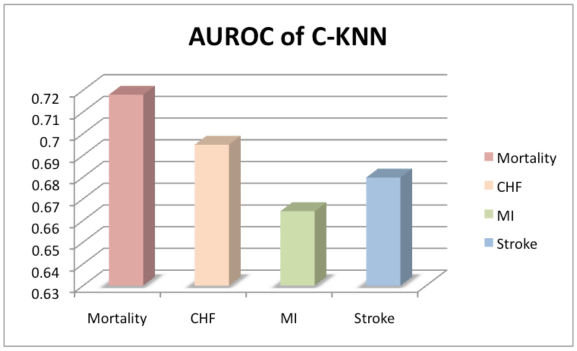

Figure 2: AUROC of C-KNN

Comparison with the baseline method We compared our approach with the state-of-the-art risk classification algorithm by the American Heart Association (AHA), to assess how effective our approach in classifying women into correct risk groups. To perform risk stratification with the AHA's algorithm, we first identified the variables that are used by the algorithm in the WISE data. All of the variables except for age, and whether the patient follows a healthy diet, was present in the WISE data and are incorporated in the risk calculation as described in (Mosca et al. 2011). Following (Mosca et al. 2011), Framingham score is calculated based on the work by (D'Agostino et al. 2008). The definition from (Grundy et al. 2004) is used to calculate the pres-

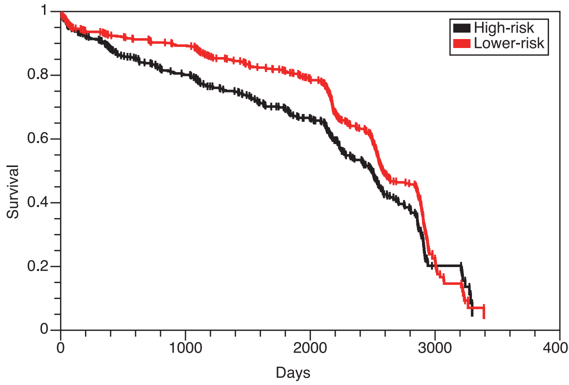

Figure 3: Survival curves after classification using AHA guidelines

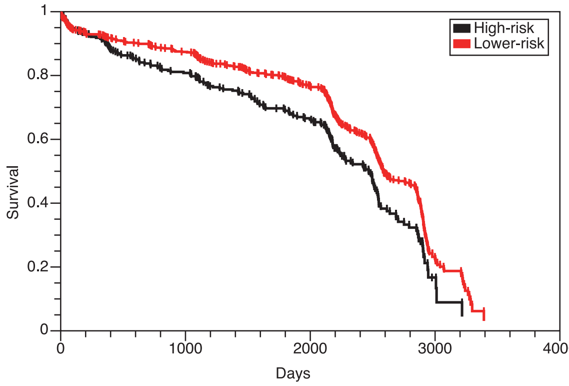

Figure 4: Survival curves after classification with C-KNN

ence of metabolic syndrome. Next, we applied C-KNN on the baseline evaluation data. Using C-KNN we partitioned patients into high and low risk groups, based on whether they had any CHD related adverse events. AHA's risk calculation algorithm identified only one optimal risk patient among all WISE patients, hence all at-risk patients are combined with optimal-risk patients to compare against high-risk patients.

After classification with each method, to compare the rates of having a cardiovascular related adverse event (death, stroke, MI or CHF), we employed Kaplan-Meier survival analysis (Kaplan and Meier 1958). Since Kaplan-Meier is suitable for patient censoring, we used all available patient data, rather than removing patients who left or passed away before the end of the clinical trial. We also calculated the hazard ratios (HR) and p-values for each method. For Kaplan-Meier survival analysis, HR and p-value estimation, we used the MStat package (Drinkwater 2010).

The survival curves for C-KNN and AHA can be seen in Figure 3 and Figure 4 respectively. Patients who left the trial prior to the completion of the trial (censored patients) are represented by ticks. Top (red) line corresponds to the predicted lower-risk patients.

Method	HR	P Value
AHA	1.284	0.0026
C-KNN	1.4	0.0002

Table 1: Hazard ratios and p-values for the predictions by AHA and C-KNN respectively.

The hazard ratios and p-values for each method is shown in Table 1. It can be seen from Table 1 that the classifications of both methods are correlated with the adverse event rates, with a high statistical significance $(p \leq 0.05)$. Yet, C-KNN clearly outperforms AHA: it has a lower p-value than AHA and more importantly, it has a higher hazard ratio than AHA. Additionally, Figure 3 and 4 show that towards the end of the trial, low risk patients predicted by AHA have actually lower survival rate than those under high risk. The patients who are identified as low risk by C-KNN have consistently higher survival rate than high risk patients. This suggests that the current heart disease risk guidelines may not be sufficient in the long run in classifying women into the correct risk groups, whereas a machine learning approach may be more reliable.

Comparison with MEB, SVC and LR We compared our method with MEB, cost sensitive SVM (C-SVM) and logistic regression (LR) and presented the best results for all 4 approaches in Figure 5. To learn MEB, we used the implementation provided by Kumar et al with epsilon value of 0.5 (Kumar, Mitchell, and Yildirim 2003), following their exact methodology.

In our experiments, we used the ratio of positive examples (patients who had an event) to the negative examples (patients who did not have an event), for the costs of positive class and negative class,. The events can be death, CHF, MI or stroke. For C-SVM, we used the LibSVM implementation (Chang and Lin 2011), with linear kernel (since it outperformed RBF and polynomial kernels). The reported AUROC scores are averages for all test combinations. As it can be seen from Figure 5, for all four types of events, our approach significantly performs better than the previous approaches and benchmark classifiers, logistic regression and SVM. We predict mortality with an impressive AUROC of 0.72.

Discussion and Conclusion

In this paper, we provided an effective approach for risk stratification in heart disease and we investigated predictive capabilities of the common diagnostic procedures. As a risk stratification method, we proposed cost-sensitive KNN, and applied it to a clinical trial dataset on womens heart disease, a dataset that has previously not been analyzed using machine-learning methods. Our method outperforms previous comparable risk stratification methods for mortality, CHF, Stroke and MI prediction. In this study, we have achieved significant results: 1. We obtained over 0.72 AUROC score, which is a very significant achievement for diagnosing heart disease in women. To our knowledge,

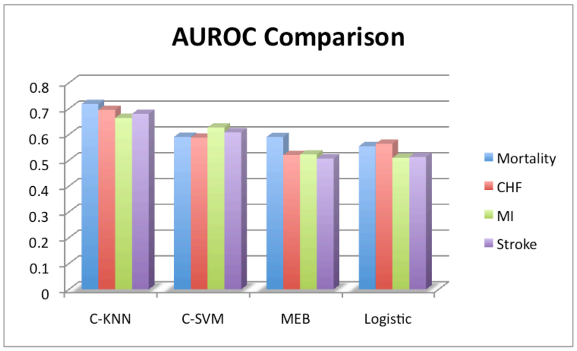

Figure 5: AUROC comparison of C-KNN, C-SVM, MEB and LR for classification in 4 different types of events

no other machine learning approach is as accurate for diagnosing womens heart disease. 2. We showed that our approach outperforms state-of-the-art, conventional risk guidelines for CHD. Our results are statistically significant ($p < 0.0002$) and as evident by the consistent adverse event rates, our approach separates women into correct risk groups. This suggests that our approach can apply to a larger population as well. 3. We achieved high AUROC scores using the least invasive, least costly and least risky tests. For CHF, MI prediction, ECG and stress tests yield the highest AUROC, and for stroke and mortality prediction, same combination along with perfusion results, give the highest AUROC. It is important to point out that none of these combinations involve angiogram, an invasive, costly procedure. 4. We challenged conventional procedures for diagnosing heart disease, and showed that in terms of predictive accuracy such procedures are unnecessary, and it is possible to achieve comparable, even better results with much less invasive diagnostic tests in the early stages. We propose that for patients with high risk factors, it is more rational to administer brachial artery ultrasound: It is not invasive, or risky, yet it alone provides highest specificity. Moreover, although ECG, exercise stress test and angiogram combination define routine practice for screening heart disease, they are not as reliable for classifying the patients as high risk, but are reliable for identifying low-risk patients. In the future, we plan to investigate which tests are the best for a specific patient.

References

Akbani, R.; Kwek, S.; and Japkowicz, N. 2004. Applying support vector machines to imbalanced datasets. In *In Proceedings of the 15th European Conference on Machine Learning (ECML*, 39–50.

Bairey Merz, C. N.; Kelsey, S. F.; Pepine, C. J.; Reichek, N.; Reis, S. E.; Rogers, W. J.; Sharaf, B. L.; Sopko, G.; and for the WISE Study Group. 1999. The women's ischemia syndrome evaluation (wise) study: protocol design, methodology and feasibility report. *J Am Coll Cardiol* 33(6):1453–1461.

Chang, C.-C., and Lin, C.-J. 2011. Libsvm: A library for support vector machines. *ACM Trans. Intell. Syst. Technol.* 2:27:1–27:27.

D'Agostino, R. B.; Vasan, R. S.; Pencina, M. J.; Wolf, P. A.; Cobain, M.; Massaro, J. M.; and Kannel, W. B. 2008. General cardiovascular risk profile for use in primary care. the framingham heart study. *Circulation*.

Davis, K. B.; Chaitman, B.; Ryan, T.; Bittner, V.; and Kennedy, J. W. 1995. Comparison of 15-year survival for men and women after initial medical or surgical treatment for coronary artery disease: A cass registry study. *Journal of the American College of Cardiology* 25(5):1000 – 1009.

Drinkwater, N. 2010. *MStat*. http://www.mcardle.wisc.edu/mstat/download/download.html.

Elkan, C. 2001. The foundations of cost-sensitive learning. In *In Proceedings of the Seventeenth International Joint Conference on Artificial Intelligence*, 973–978.

Grundy, S. M.; Brewer, H. B.; Cleeman, J. I.; Smith, S. C.; Lenfant, C.; and for the Conference Participants. 2004. Definition of metabolic syndrome. *Circulation* 109(3):433–438.

Kaplan, E. L., and Meier, P. 1958. Nonparametric estimation from incomplete observations. *Journal of the American Statistical Association* 53(282):pp. 457–481.

Kumar, P.; Mitchell, J. S. B.; and Yildirim, E. A. 2003. Approximate minimum enclosing balls in high dimensions using core-sets. *J. Exp. Algorithmics* 8.

Lessmann, S. 2004. Solving imbalanced classification problems with support vector machines. In *IC-AI*, 214–220.

Lloyd-Jones, D.; Adams, R. J.; T, B.; and et al. 2010. Heart disease and stroke statistics 2010 update a report from the american heart association. *Circulation* 121(7):e46–e215.

Melville, P.; Saar-Tsechansky, M.; Provost, F.; and Mooney, R. J. 2005. Economical active feature-value acquisition through expected utility estimation. In *Proceedings of the KDD-05 Workshop on Utility-Based Data Mining*, 10–16.

Mieres, J. H.; Shaw, L. J.; Arai, A.; Budoff, M. J.; Flamm, S. D.; Hundley, W. G.; Marwick, T. H.; Mosca, L.; Patel, A. R.; Quinones, M. A.; Redberg, R. F.; Taubert, K. A.; Taylor, A. J.; Thomas, G. S.; and Wenger, N. K. 2005. Role of noninvasive testing in the clinical evaluation of women with suspected coronary artery disease. *Circulation* 111(5):682–696.

Mosca, L.; Benjamin, E. J.; Berra, K.; and et al. 2011. Effectiveness-based guidelines for the prevention of cardiovascular disease in women–2011 update: a guideline from the American Heart Association. *Journal of the American College of Cardiology* 57(12):1404–1423.

Pepe, M. S. 2003. *The Statistical Evaluation of Medical Tests for Classification and Prediction*. New York: Oxford.

Shaw, L. J.; Bugiardini, R.; and Merz, C. N. B. 2009. Women and ischemic heart disease: Evolving knowledge. *J Am Coll Cardiol* 54(17):1561–1575.

Stangl, V.; Witzel, V.; Stangl, G.; and Stangl, K. 2008. Current diagnostic concepts to detect coronary artery disease in women. *European Heart Journal* 29(6):707–717.

Syed, Z., and Rubinfeld, I. 2010. Unsupervised risk stratification in clinical datasets: Identifying patients at risk of rare outcomes.

Visa, S. 2005. Issues in mining imbalanced data sets - a review paper. In *in Proceedings of the Sixteen Midwest Artificial Intelligence and Cognitive Science Conference, 2005*, 67–73.