*Journal of the American Medical Informatics Association*, 25(11), 2018, 1481–1487
doi: 10.1093/jamia/ocy109
Research and Applications

Research and Applications

# Machine learning for psychiatric patient triaging: an investigation of cascading classifiers

**Vivek Kumar Singh,[1] Utkarsh Shrivastava,[2] Lina Bouayad,[3,4] Balaji Padmanabhan,[1] Anna Ialynytchev,[4] and Susan K Schultz[5,6]**

[1]Information Systems and Decision Sciences, MUMA College of Business, University of South Florida, Tampa, Florida, USA, [2]Haworth College of Business, Department of Business Information Systems, Western Michigan University, Kalamazoo, Michigan, USA, [3]College of Business, Information Systems and Business Analytics, Florida International University, Miami, Florida, USA, [4]HSR&D Center of Innovation on Disability and Rehabilitation Research (CINDRR), James A. Haley Veterans Hospital, Tampa, Florida, USA, [5]James A. Haley Veterans Hospital, Geriatric Psychiatry, Tampa, Florida, USA, and [6]Department of Psychiatry and Behavioral Neurosciences, Morsani College of Medicine, University of South Florida, Tampa, Florida, USA

Corresponding Author: Vivek Kumar Singh, Doctoral Candidate, Information Systems and Decision Sciences, MUMA College of Business, University of South Florida, 4202 E Fowler Ave, Tampa, FL 33620, USA (vivek4@mail.usf.edu)

Received 22 March 2018; Revised 29 June 2018; Editorial Decision 23 July 2018; Accepted 26 July 2018

**ABSTRACT**

**Objective:** Develop an approach, One-class-at-a-time, for triaging psychiatric patients using machine learning on textual patient records. Our approach aims to automate the triaging process and reduce expert effort while providing high classification reliability.

**Materials and Methods:** The One-class-at-a-time approach is a multistage cascading classification technique that achieves higher triage classification accuracy compared to traditional multiclass classifiers through 1) classifying one class at a time (or stage), and 2) identification and application of the highest accuracy classifier at each stage. The approach was evaluated using a unique dataset of 433 psychiatric patient records with a triage class label provided by "I2B2 challenge," a recent competition in the medical informatics community.

**Results:** The One-class-at-a-time cascading classifier outperformed state-of-the-art classification techniques with overall classification accuracy of 77% among 4 classes, exceeding accuracies of existing multiclass classifiers. The approach also enabled highly accurate classification of individual classes—the severe and mild with 85% accuracy, moderate with 64% accuracy, and absent with 60% accuracy.

**Discussion:** The triaging of psychiatric cases is a challenging problem due to the lack of clear guidelines and protocols. Our work presents a machine learning approach using psychiatric records for triaging patients based on their severity condition.

**Conclusion:** The One-class-at-a-time cascading classifier can be used as a decision aid to reduce triaging effort of physicians and nurses, while providing a unique opportunity to involve experts at each stage to reduce false positive and further improve the system's accuracy.

**Key words**: triage, cascading classification, decision aid, ensemble algorithms, I2B2 challenge, text mining

## BACKGROUND AND SIGNIFICANCE

Triaging, a process of classifying patients based on the severity of their condition, is an important part of patient treatment workflow. Medical personnel including physicians, nurses, and paramedics must frequently make triaging decisions. In high-risk and costly settings, patient triaging is commonly used to prioritize patients and optimize the utilization of limited healthcare resources and staff.[1] The triaging staffs evaluate incoming patients before routing them to the most appropriate treatment pathway.

There are 2 key challenges in triaging incoming patients: under-triage and over-triage. Under-triage represents misclassification of a patient with a severe medical condition to a non-severe category. Healthcare experts agree on the importance of keeping under-triage as low as possible—to about less than 5%.[2] Over-triage, on the other hand, represents misclassification of a patient with a non-severe medical condition to a severe category. While over-triage is still a concern among healthcare experts, with an acceptable range of 25% to 30%,[2] it is less likely to impact patient health outcomes.[3]

Because of the nature of psychiatric assessment and variability in triage standards across providers, prioritizing care for patients with mental illness through an efficient triage mechanism is a complex task. Patients' psychiatric assessments are based on both the level of danger posed to themselves and others, as well as the extent of impairment in social functioning. Triaging of medical emergencies, on the other hand, solely focuses on identification of patients with life-threatening physiological conditions or diseases. In some patient cases, medical and psychiatric illnesses may coexist, and one may be the cause or contributing factor of the other.

The preliminary psychiatric screening information on the patient is organized as an initial psychiatric evaluation (IPE) record by the attending nurse or psychiatrist. The IPE records include a variety of psychiatric assessments on different dimensions such as lifestyle, social behavior, addictions, and relationships of the patient. These records also include healthcare providers' or clinicians' interpretations of the patient's condition derived from clinical observations and conversations with the patient, which also include information such as financial stress or frequent alcohol intake.

While the reliability of triage decisions based on information in the IPE records has generally improved by the introduction of standardized scales, a recent study found that there is insufficient scientific evidence on the inter-rater reliability among these scales.[4] Moreover, the experience and confidence of triaging physicians in evaluating initial psychiatric interview information contribute to the variation in assessment of mental health severity.[5]

The digitalization of patient records has provided new avenues for improving the accuracy and efficiency of the initial patient categorization procedures. Some recent studies have found that automating the patient triaging procedure through machine learning algorithms using the information extracted from electronic medical records (EMRs) can be effective in improving predictability of medical interventions.[6] However, unlike standardized EMRs, the psychiatric evaluation records are highly unstructured containing textual elaboration of patients' experiences and psychiatrists' interpretation of the patient's response. The unavailability of IPE records due to privacy concerns implies that very few studies in the past have investigated challenges in extracting information from psychiatry records for addressing patient triaging challenges. To address this knowledge gap, our paper focuses on a patient-diagnosis scenario in which IPE records are classified into severity categories that will determine how those psychiatric cases will be handled.

Building classification models from IPE records is complex because of 1) the difficulty of the underlying information extraction and feature selection tasks, and 2) the cost of severity misclassification in terms of health outcomes, which in certain cases, may exceed the benefits from automation. These complexities lead us to build a hybrid system that is capable of addressing the challenges of information extraction from psychiatric records, over-triage, and under-triage, but is flexible enough to include domain experts in critical scenarios. For example, an expert may review the outcome of our proposed model at each stage and may reject the outcome if he or she believes the patient has been misclassified. In this case, the record will be considered for triage in subsequent stages or severity levels.

Classification models that mirror the triaging mechanism fit well within this framework. Examples of this in the machine learning literature are *cascading classifiers*,[7–9] a special type of ensemble method in which sequential or multi-stage classifiers are used. The output of a prior classification model informs the next classifier applied in sequence. A popular technique is *AdaBoost*[7,10], which sequentially builds classifiers based on instance weighting adjustments to increase the likelihood that the next classifier will minimize error made by the previous classifiers. However, this is a method that goes after "one classifier at a time" as opposed to one "class" at a time, as each application of the classifier still predicts *all* classes. Motivated by the need to support triaging scenarios, in this work, we explore cascading classifiers that target "one class at a time" in multi-class classification problems.[8] Specifically, we work with 4 severity categories, namely severe, moderate, mild, and absent. The classifier in each stage is a binary classifier built to distinguish between the target class and the rest of the classes.

In this domain, because of the urgency treatment required for a severe patient, focusing on accurately classifying severe cases (or a particular class) may typically be more important than overall classifier accuracy. Cascading classifiers are a natural fit for these circumstances, as each classifier in the sequence can predict a specific class in importance or severity. If the goal of classification is primarily overall cost minimization, a wide range of cost-sensitive modeling approaches is available.[11] However, if we also consider the need to support an existing triaging process, cascading classifiers that focus on one class at a time provide a natural framework. In this paper, we propose and develop a novel One-class-at-a-time cascading classifier with overall classification accuracy of 77%.

## MATERIALS AND METHODS

### Data

A set of 433 fully de-identified IPE records describing patients' mental state as recorded by psychiatrists was sourced from Partners Healthcare and the Neuropsychiatric Genome-Scale and Research Domain Criteria (RDoC) Individualized Domains (N-GRID) project of Harvard Medical School.[12] There are many dimensions of human behavior that can be extracted from these records, but we focused on the specific domain of "positive valence" as specified by the I2B2 challenge. Positive valence is one of the 5 dimensions of the RDoC framework. It is defined as "events, objects, or situations that signal mental disorders but are attractive to the patients, to the point that they actively engage in them" and includes behavioral issues such as alcohol and drug abuse, gambling, repetitive and/or compulsive behavior, craving, and counting.[12] This focus on positive valence is due to its relationship to harmful outcomes and direct relation to a

higher need for care. While positive valence symptoms do tend to aggregate within some diagnoses such as substance use, this relates to the more prominent role of these conditions in the safe triage of patients and explains their selection for this study.

These records, representing our "gold standard", were annotated by 3 expert psychiatrics from Massachusetts General Hospital (MGH) and the Harvard Medical School, with several years of experience and were first shared with the research community through the 2016 CEGS N-GRID challenge. In the dataset, 325 records were annotated by 2 annotators, while a third annotator was asked for a majority opinion for 108 records. The ratings ranged from absent, mild, moderate, and severe in reference to having positive valence symptoms that required hospitalization or emergency room visit or otherwise implicated a potentially major safety concern. Four labels were used for annotations: *absent*—describing patients with no positive valence symptoms, *mild*—have some symptoms but not the focus of treatment, *moderate*—having symptoms that are a focus of treatment but do not require hospitalization, and *severe*—having symptoms that require hospitalization, ED visit, or otherwise having a major consequence. It should be noted that identifying the severity of symptoms is key to an effective treatment path. In the dataset provided, 22.17% of cases were severe, 25.4% moderate, 38.3% mild, and 14.08% absent.

The IPE records include a range of information related to the patient's actions (eg, suicidal behavior), experiences (eg, military services), and medication history. There are some standard yes/no questions (eg, psychiatric history of inpatient treatment), as well as open-ended questions (eg, detailed description of how patient tried to commit suicide). This patient information was first extracted from IPE records, and then used as features in our severity classification model.

The data used in this paper were fully de-identified and acquired through the I2B2 challenge upon signature of a data use agreements by all of the team members. For additional diligence, approval was sought from IRB at our institution and approval was received.

## Feature extraction and selection

To classify IPE records based on symptom severity, we developed a feature extraction framework to recognize the factors of interest for the classification problem as shown in Figure 1. Consulting with domain experts, our feature set was divided into different categories:

- General Information: included features such as appearance, clothing, gender, and age.
- Disorders: included positive valence-related disorders such as substance disorders.
- Symptoms: included use of substances such as cocaine, alcohol, marijuana, and hallucinogens.
- History: included historical inpatient and outpatient treatments, as well as past suicide attempts, violence, and drinking behaviors.
- Treatments: included common medications and therapies used for treating positive valence disorders.
- Severe consequences: included hospitalizations and other serious events such as arrest or use of firearms.
- Other: included all other patient attributes such as employment and marital status.
- The Brief Psychiatric Rating Scale (BPRS)[13], which is one of the widely used methods for measuring mental illness severity, was used for coding as per the scale.

## One-class-at-a-time cascading model

Cascading classifiers have been used in image pattern recognition literature.[8] In pattern recognition and image classification, cascading is used as multi-stage classification wherein images are classified in different stages. In the initial stage (s), using a small subset of image features, the majority of images that do not contain a required pattern, such as a human face in a face detection problem, are removed. In a later stage (s), the majority of features are used with more complex classification algorithms to detect and identify the content in the image, such as identifying the person in the image.

Prior research has shown that multi-stage classifiers are more powerful than ensemble methods, such as voting-based selection of best-performing classifiers.[2] AdaBoost or adaptive boosting algorithms have been used in conjunction with cascading classifiers to choose the appropriate feature set from the large set of potential features, and has shown to be an effective algorithm for machine learning with strong bounds on the generalization performance.[3] We therefore elected to build a cascading classification model to classify psychiatric cases based on severity. This approach has potential to include human decision making at every step of the classification process, as shown in Figure 2.

The performance of our One-class-at-a-time classifier depends on the order in which the classes are considered. As suggested by our domain application, one option to consider is classification order based on class importance (ie, severe cases first, followed by moderate second, etc.). In this domain, the order of class importance maps well with ease of predictability. Severe cases are often easier to predict, as the symptoms are more pronounced in the data. This process of classifying easy classes first is also consistent in principle with AdaBoost algorithms. However, if the goal is to improve the overall classification accuracy, we might also want to focus on frequent classes first, particularly if the frequent classes are also easier to predict. This provides 2 dimensions to consider—frequency and predictability. In this paper, we use an exhaustive search to determine the order of prediction and selection of classifier at each stage, which provides maximum overall classification accuracy across the 4 categories for our evaluation.

The One-class-at-a-time cascading algorithm is outlined in Figure 3. In our One-class-at-a-time cascading algorithm, $T$ represents the set of patient records to be used for training the cascade of classifiers, and $T$' represents the set of records in the holdout sample for testing the classification accuracy. In the first step, the set of classes/categories C = {severe, mild, moderate, absent} is ordered based on the preference of classification guided by a heuristic, such as frequency or predictability. C' denotes the ordered classes (ie, severe, moderate, mild, and absent based on severity). In this study, we explore the possibilities of determining a cascading sequence based on frequency distribution (high/low) of the classes in the training set or the severity of classes (high/low).

Next, we select the best classifier or algorithm from set A (total of 9 in this study) of the possible algorithms to classify C' in sequence at each stage of the cascade. The idea is to use the best estimator for a class $c'_i$ at the level/stage "I" of the cascade. This step is critical since the classification accuracy at each level depends on the feature set of the training set. As an example, the features used to differentiate between severe and absent cases may be different from the set of features needed to differentiate between severe and moderate cases. At each stage, "s" of the cascade, we train a binary classifier for classifying the class $c'_s$ against the remaining classes $c'_i$ ($i > s$) using the feature set $F_i$. In the subsequent stage, "s + 1", a

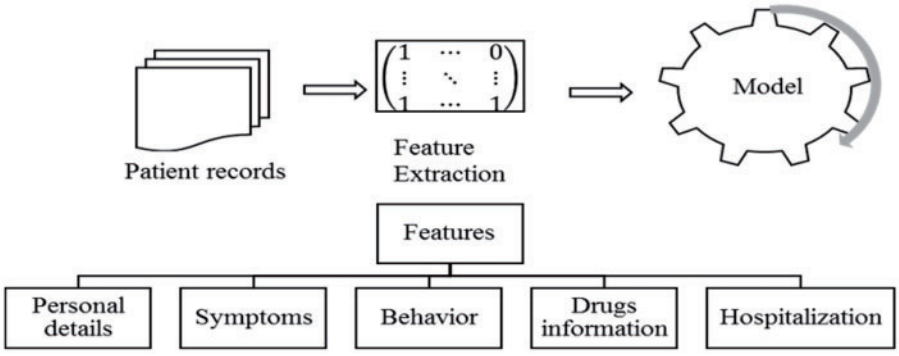


**Figure 1.** Snapshot of the feature extraction framework.

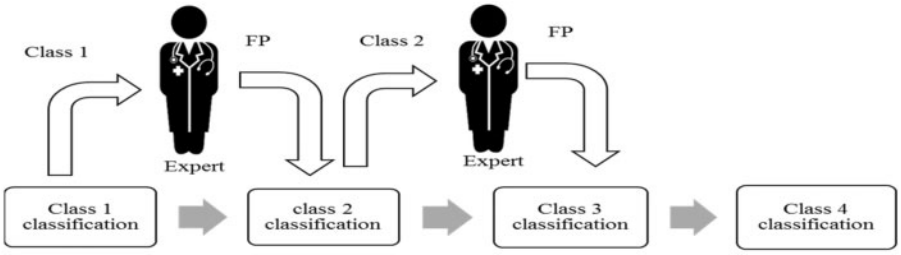


**Figure 2.** One-class-at-a-time classification approach.

**Input:** T, L, C
**Output:** Model

1. Given the training examples T = $[t_1,t_2....t_n]$, training class labels L = $[l_1^t,l_2^t....l_n^t]$
2. Given the validation examples V = $[v,v_2....v_m]$, validation class labels L' = $[l_1^v,l_2^v....l_n^v]$
3. Let, C = $[c_1,c_2....c_c]$ denote the set of all possible class values that the labels L and L' can take.
4. Sort C based on a heuristic (e.g. severity) to get a vector C': C' = $[c'_1,c'_2....c'_c]$
5. Initialize: $T^1$= T, $V^1$ = V
6. Let A = $[a_1,..... a_l]$ be the set of classifiers at each level with maximum of i
7. While i ≤ (c-1) {
8. $T = T^i$; $V = V^i$
9. $B_i^T = [b_1^t,b_2^t,....b_n^t]$ such that $b_k^t = 1$ if class $(l_k^t) = c'_i$ else 0 for any k < n
10. $B_i^V = [b_1^v,b_2^v,....b_m^v]$ such that $b_k^v = 1$ if class $(l_k^v) = c'_i$ else 0 for any k < m
11. Relabel the class labels of training instances L and validation instances V such that $l_n = 1$ if
12. While j ≤ (l) {
13. Train classifier $a_j$ on set T using features $F_i$ for binary classification $(B_i^T)$
14. Predict the binary classification $P_{ij}$ for the instances in the validation set V using $a_j$ on and the feature set $F_i$ such that $P_{ij} = [p_{1ij} = F_1(T,B_i^T,a_j,V), .. p_{mij} = F_m(T,B_i^T,a_j,V)]$
15. Return the accuracy $Acc_{ij} = \sum_{k=1}^{m}(p_{kij} - b_k^v)$ / m of the classifier
16. }
17. Select the classifier $a_f$ such that $Acc_{if} = \max(Acc_{ij})$ for any j ≤ l
18. $MODEL_i = a_f$ (Identified classifier at for the stage i)
19. $T^i = T^i - T^i_{(class(l_k^t)= c'_i)}$
20. $V^i = V^i - V^i_{(class(l_k^v)= c'_i)}$
21. Return (Model)
22. }

**Figure 3.** One-class-at-a-time cascading model-building and application algorithms.

new model is trained on the training data $T^i - T^i_{(class(l_k^t)= c'_i)}$ that excludes the records included in the class $c'_s$. The accuracy of the classifier $(a_j)$ is accessed on the validation set based on the predictions $P_{ij}$ from the trained model. The model-building procedure returns a set of best models (MODEL) at each stage of the cascade for a given sequence of the classes C'. The generated One-class-at-a-time cascading model is then applied to the test dataset to evaluate performance.

Cascading algorithms are usually shown in sequence in the order in which classification models are applied.[4] In Figure 4, we use this representation to describe the algorithm presented above. For the final predictions, the records in the holdout test dataset $Test^1$ are first classified ($c'_1 = 1/0$) using the "Model 1" from the model sequence (MODEL) obtained in the model-building stage. The records classified as $c'_1$ are then assigned to $P_1$ and are excluded from $Test^1$ to get new input ($Test^2$) for the trained "Model 2." The cascade continues until $c'_n$ is classified. The final prediction "P" is the union of predictions at each stage.

### Evaluation

To evaluate the performance of the One-class-at-a-time cascading models, we used de-identified psychiatry notes provided in the I2B2 challenge to predict symptom severity. We evaluated our proposed model with multi-class classification models such as Nearest Neighbor, RBF SVM, Decision Tree, Random Forest, Naïve Bayes, Quadratic Discriminant Analysis, Linear SVM, Linear Discriminant Analysis, and AdaBoost.

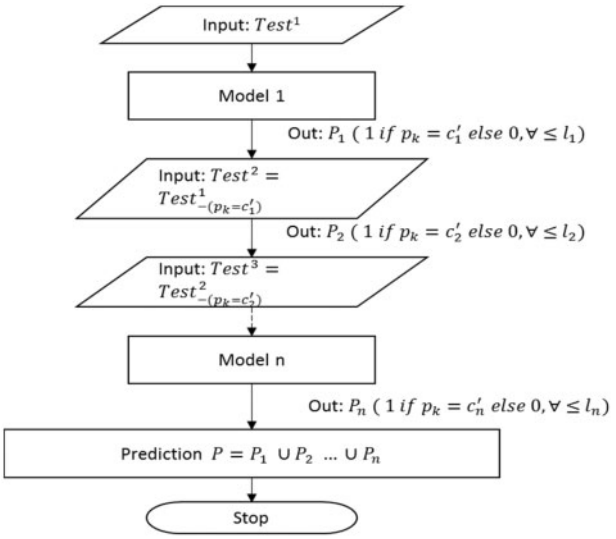


**Figure 4.** One-class-at-a-time prediction flowchart.

## RESULTS

### Cascading vs non-cascading classification results

Results in Table 1 indicate that for most models, the performance of the cascading models is better than non-cascading models, as the cascading models provide higher overall accuracies compared to non-cascading classifiers. Overall, this seems to suggest the potential for cascading models in this domain.

These results focused on the overall classification performance of One-class-at-a-time and hence pertain to models in which a different algorithm is used at each classification. However, the cascading literature also notes that subsequent classification models can be different (and in fact are often more complex). Hence, in our One-class-at-time cascading model illustrated in Figure 4, several algorithms are tested at each stage, and the best algorithms are then combined to build the final model.

### Cascading classification results – varying severity order

Figure 5 presents the results of applying our approach based on using severity levels. As noted above, each stage used multiple algorithms before picking the best one. For example, stage 1 may use a Decision Tree to predict "Severe/Non-Severe," stage 2 may then use SVM to predict "Moderate/Non-Moderate," and stage 3 may then use Decision Tree (again) to predict "Mild/Absent." What is important here is that the algorithm picked in each stage is determined based on performance of each classifier in the validation dataset at each stage.

The main results here suggest that the order of cascading can impact overall accuracy, as well as the performance within each class. In this case, the "decreasing severity" models appeared to be doing better than the "increasing severity models." This finding was as per our initial intuition, as we expected the severe cases to be more easily classifiable. Moreover, what we see here is that "absent" cases are less predictable than severe, and thus, removing the "severe" cases first appears to be a good strategy. As both "increasing severity" and "decreasing severity" are meta-algorithms that still eventually classify a patient as absent, mild, moderate, or severe, both approaches can still be used to identify severe cases first for triaging purposes based on overall classification accuracy.

**Table 1.** Performance comparison

| Classifier | Accuracy |
|---|---|
| Nearest Neighbors | 33% |
| Naïve Bayes | 39% |
| RBF SVM | 42% |
| AdaBoost | 47% |
| Quadratic Discriminant Analysis | 50% |
| Decision Tree | 53% |
| Random Forest | 55% |
| Linear SVM | 61% |
| Linear Discriminant Analysis | 61% |
| One-class-at-a-time | 77% |

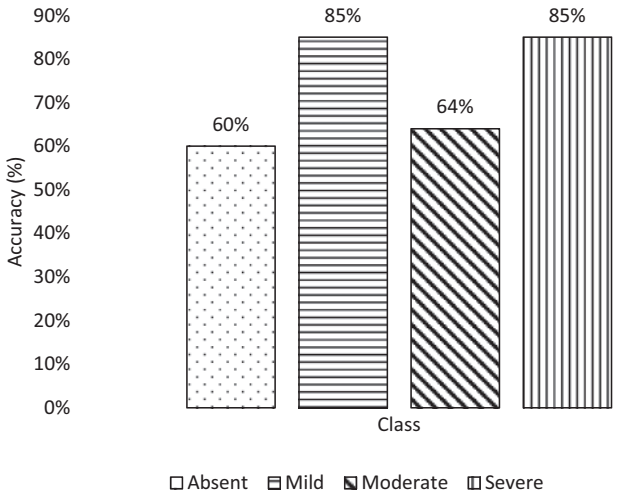


**Figure 5.** One-class-at-a-time (performance for different classes).

Using the increasing severity cascading model as a basis, we compare model performance with feature selection using domain knowledge based on classes being predicted at each level. These are consistent with the general ideas in the cascading literature that using different feature sets in different cascading stages is likely to be

useful. Overall, our results indicated the importance of One-class-at-time cascading models in psychiatric symptoms classification. Incorporating classification order and domain knowledge in feature selection, One-class-at-a-time cascading models appeared to have promising results in this domain. To improve the overall performance of the proposed method, we searched over different groups of features in each stage in the model used for evaluation. Our proposed technique outperformed the existing state-of-the-art multiclass classification techniques, including ensemble techniques such as AdaBoost.

Using an iterative process, we were able to determine 1) the order of the severity classification that provides the highest accuracy and 2) the classifiers that provide maximum accuracy at each stage. Results indicated that the best classification sequence is mild, followed by absent, moderate, and severe. The highest performance was achieved by classifying the mild category from the rest using the AdaBoost classifier, the absent category using the Decision Tree classifier, and moderate category from severe using the Linear SVM classifier.

Error analysis indicated a high rate of false positive cases in the mild category. One plausible explanation is that the number of mild samples in the dataset dominates the other categories. This can also be due to misclassification of the absent category to the mild category, which is a typical case of over-triage.

Overall, evaluation of the One-class-at-a-time classifier showed high classification accuracy, especially for the severe and mild cases (85%). These results have important implications for the practice of how classifiers can be integrated with clinical decision support systems to reduce the rates of 1) under-triage (along with associated poor health outcomes indicated[2]) and 2) over triage (which in turn can help improve the utilization of limited resources and staff[1,2]).

### Statistical test for robustness check

To test for the significance of our results, we conducted McNemar's test for a pair of classifiers.[14–16] For each of the samples, we determined the prediction from each of the classifiers. We evaluated the classification as correct if the predicted label is the actual label; otherwise, we considered it to be an incorrect classification. Following this process, we obtained the contingency table shown in Table 2.

McNemar's test is defined as follows:

$$\chi^2 = \frac{(b-c)^2}{b-c}$$

The *P*-value for the McNemar test is significant, which means that the performance of the One-class-at-a-time classifier is significantly better than random classification, as shown in Table 3. We also tested the performance of our classifier using the Kappa test.[17,18] Kappa is a measure of agreement between the outcomes from 2 models after accounting for the fact that models may agree or disagree on an outcome simply by chance. A kappa of 1 indicates perfect agreement, while that of 0 indicates agreement due to chance. The Kappa value of the One-class-at-a-time classifier is 0.67, which is higher compared to other multiclass classifiers.

We also compared our results with the performance of other approaches published recently using same dataset in Table 4. The inverse normalized macro-averaged mean absolute errors score ($INMAE^M$)[12] was used to evaluate the overall model classification accuracy across the 4 classes (severe, moderate, mild, and absent). Although our overall accuracy is lower compared to other competing approaches, our unique approach to classify one class at a time is appropriate with respect to domain requirement of medical triage.

**Table 2.** Contingency table for standard McNemar's test

| | Classifier 2 (correct) | Classifier 2 (incorrect) |
|---|---|---|
| Classifier 1 (correct) | a | b |
| Classifier 1 (incorrect) | c | d |

**Table 3.** Robustness test results for classifiers

| Classifier | Accuracy | McNemar's test (*P*-value) | Kappa value |
|---|---|---|---|
| Nearest Neighbors | 33% | .74 | 0.07 |
| Naïve Bayes | 39% | <.01 | 0.21 |
| RBF SVM | 42% | NA | 0.00 |
| AdaBoost | 47% | <.01 | 0.22 |
| Quadratic Discriminant Analysis | 50% | NA | 0.10 |
| Decision Tree | 53% | <.01 | 0.32 |
| Random Forest | 55% | NA | 0.18 |
| Linear SVM | 61% | NA | 0.32 |
| Linear Discriminant Analysis | 61% | .94 | 0.22 |
| One-class-at-a-time | 77% | .07 | 0.67 |

**Table 4.** Performance comparison with other research work[10]

| Algorithm/Team | $INMAE^M$ Score |
|---|---|
| SentiMetrix Inc. (best between 2 submissions) | 0.86 |
| University of Texas at Dallas | 0.84 |
| University of Kentucky (best among 3 submissions) | 0.83 |
| Med Data Quest Inc. | 0.83 |
| University of Pittsburgh | 0.82 |
| One-class-at-a-time | 0.77 |

## DISCUSSION

Given the increased interest in machine learning approaches today, there are many approaches based on combining models. One specific method of doing so, cascading, appears to be relatively rarely used, although in earlier applications of pattern/image recognition, this showed significant potential. Motivated by an important healthcare application, in this work, we show the potential of using cascading models for predicting the severity of patient cases.

Our results were consistent with results in the cascading literature that 1) overall cascading models can increase performance compared to non-cascaded ones, 2) feature selection integrated into the cascading stages can be valuable, and 3) using different classifiers at each stage is likely better than using the same classifier throughout. We achieved an overall accuracy of 77%, as shown in Table 1. The accuracy of the prediction also varies with the stage. The algorithm performs highest in mild and severe categories, followed by moderate and absent categories, respectively.

In a novel approach compared to what has been done in the machine learning literature, we manipulated the order of cascading based on the severity of patient cases to study its impact on overall performance. We found that the order mattered, although not in the exact manner hypothesized. The importance of the order mapped more with predictability than severity. In our study, the "severe" cases were most predictable and were therefore the ones that were most valuable to seek out first.

We applied these ideas to a real dataset provided by the I2B2 challenge, in which the cases involved real psychiatric patient records to be classified based on their severity. This is an application in which significant human effort is required, and an accurate classification is very important. Severity misclassification leading to under- or over-triage can have important implications, both in terms of economic costs of treatment, as well as for patient health outcomes.

While the One-class-at-a-time cascading method used in this research focused on positive valence disorders, our classification system was based on severity of disorder. As all mental health disorders can range in severity, this triage system could theoretically be applied to any disorder. However, in order to be replicated for other disorders, appropriate datasets would be required and new feature sets extracted. Specifically, the features included in general information, history, severe consequences, and other would be applied to all disorders, while disorders, symptoms, and treatments would require updating the feature sets.

Incorporating effective machine classification, as illustrated in this paper, can lead to significant process improvements. This approach can be integrated into current healthcare processes where triaging is common, potentially improving the quality of care and reducing costs. Nevertheless, constraints in learning algorithms are needed to further optimize classification accuracy depending on the clinical setting. In settings where under-triage is common, for example, high under-classification costs could be assigned in the model to ensure the under-triage rate is kept to a minimum.

## CONTRIBUTORS

VKS, US, BP, and LB were involved in conceptualization, and VKS, US, and LB in design and implementation of the One-class-at-a-time algorithm. AI and SKS provided domain knowledge. All contributed to manuscript drafting and revisions.

## FUNDING

This research received no specific grant from any funding agency in the public, commercial or not-for-profit sectors.

*Conflict of interest statement*. The authors have no competing interest to declare.

## ACKNOWLEDGMENTS

We acknowledge the support of the grants NIH P50 MH106933 (to PI: Isaac Kohane) and NIH 4R13LM011411 (to PI: Ozlem Uzuner) that funded the competition (CEGS N-GRID - 2016 shared task).

## REFERENCES

1. Gall C, Wetzel R, Kolker A, Kanter RK, Toltzis P. Pediatric triage in a severe pandemic: maximizing survival by establishing triage thresholds. *Crit Care Med* 2016; 44 (9): 1762–8.
2. Rotondo MCC, Smith R. Resources for optimal care of the injured patient (1sted.). *Am Coll Surg* 2014. doi:10.1016/j.jamcollsurg.2005.06.003.
3. Newgard CD, Staudenmayer K, Hsia RY, *et al*. The cost of overtriage: more than one-third of low-risk injured patients were taken to major trauma centers. *Health Aff (Millwood)* 2013; 32 (9): 1591–9.
4. Farrohknia N, Castrén M, Ehrenberg A, *et al*. Emergency department triage scales and their components: a systematic review of the scientific evidence. *Scand J Trauma Resusc Emerg Med* 2011; 19 (1): 42.
5. Chung J. An exploration of accident and emergency nurse experiences of triage decision making in Hong Kong. *Accid Emerg Nurs* 2005; 13 (4): 206–13.
6. Cai X, Perez-Concha O, Coiera E, *et al*. Real-time prediction of mortality, readmission, and length of stay using electronic health record data. *J Am Med Inform Assoc* 2016; 23 (3): 553–61.
7. Pudil P, Novovicova J, Blaha S, Kittler J. Multistage pattern recognition with reject option. In: Proceedings of 11th IAPR International Conference on Pattern Recognition. Vol. II. Conference B: Pattern Recognition Methodology and Systems. 1992: 92–95. doi:10.1109/ICPR.1992.201729.
8. Oliveira LS Jr, Britto AS, Sabourin R. Improving cascading classifiers with particle swarm optimization. In: Eight International Conference on Document Analysis and Recognition (ICDAR'05). 2005.
9. Alpaydin E, Kaynak C. Cascading classifiers. *Kybernetika* 1998; 34 (4): 369–374.
10. Bishop CM. Pattern recognition and machine learning. Springer publication; 2006.
11. Elkan C. The foundations of cost-sensitive learning. *Int Jt Conf Artif Intell* 2001; 973–8. http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.29.514
12. Filannino M, Stubbs A, Uzuner Ö. Symptom severity prediction from neuropsychiatric clinical records: overview of 2016 CEGS N-GRID shared tasks Track 2. *J Biomed Inform* 2017. doi:10.1016/j.jbi.2017.04.017.
13. Overall JE, Gorham DR. The brief psychiatric rating scale. *Psychol. Rep* 1962; 10: 799–812.
14. Walker G. *Common statistical methods for clinical research with SAS examples*. Cary, NC: SAS Institute; 2010.
15. McNemar Q. Note on the sampling error of the difference between correlated proportions or percentages. *Psychometrika* 1947; 12 (2): 153–7.
16. Aphinyanaphongs Y. Text categorization models for high-quality articleretrieval in internal medicine. *J Am Med Inform Assoc* 2004; 12 (2): 207–16.
17. Uzuner O, Zhang X, Sibanda T. Machine learning and rule-based approaches to assertion classification. *J Am Med Inform Assoc* 2009; 16 (1): 109–15.
18. Chapman W, Dowling J, Wagner M. Generating a reliable reference standard set for syndromic case classification. *J Am Med Inform Assoc* 2005; 12 (6): 618–29.