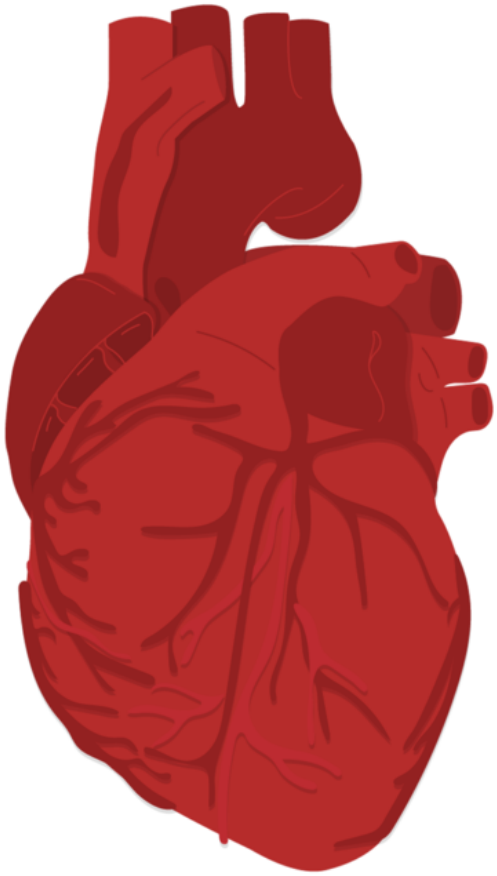
EPSRC
Engineering and Physical Sciences Research Council

InFoMM
Industrially Focused Mathematical Modelling

EPSRC Centre for Doctoral Training in Industrially Focused Mathematical Modelling



Congestive heart failure triage prediction model

James Morrill

UNIVERSITY OF OXFORD

REVON

Contents

1. Introduction

Motivation

Early detection of CHF exacerbations.

Heart failure is a chronic, progressive condition in which the heart muscle is unable to pump enough blood to meet the body's needs for blood and oxygen. It affects at least 26 million people worldwide, is cited as a contributing factor in 1 in 9 deaths in the US, and costs the nation $39.2 billion per year. With an ageing population, the epidemic of heart failure is going to increase in the coming decades with estimates that, by 2030, more than 8 million people in the US will have the condition, accounting for a 46% increase in prevalence [1]. Acute congestive heart failure is the rapid onset of symptoms and signs of heart failure that involves a worsening of clinical status; these flare-ups (or exacerbation's) account for nearly all of the $39 billion spent per year in the US primarily due to hospitalisations resulting from the worsening clinical status. Early detection of such flare ups is therefore essential to reduce the economic burden placed on healthcare systems worldwide, as well as to reduce the unnecessary stress and uncertainty patients experience when self diagnosing.

Background

One way to improve early detection of exacerbations is through the use of mobile applications. Currently the gold standard for a patient at home to determine whether they are having an exacerbation is through the use of "action plan" checklists, patients refer to a document when they are feeling concerned about their symptoms [2]. This will direct the patient, depending on the severities of their symptoms, to (i) continue treatment as normal, (ii) call their physician, or (iii) go immediately to the emergency room. While the medical guidance in these checklists has been useful to educate patients, the method of delivering that guidance through a hard-coded list lacks rigour, validation, and robustness at the level of the individual patient. This results in more frequent calls to the doctors office and visits to the emergency room than are strictly necessary. This might be alleviated by a more specialised diagnosis method that captures the complexities and the interplay between different vital signs and symptoms. Revon Systems are developing an application that can be used as a diagnosis tool for patients at home that will utilise machine-learning techniques to improve diagnosis of exacerbation. Revon have already had success in the prediction of chronic obstructive pulmonary disease (COPD) exacerbation, developing an algorithm that predicts patient state with an accuracy often better than the top performing doctor [3].

Revon are developing an at home diagnosis tool for patients with CHF.

Our aim is to expand on the work performed by Revon on COPD and apply it to predict exacerbations for patients with CHF. From an individual patient case containing information on their profile, symptoms and vitals signs, the goal is to predict a triage category for the patient, predict whether the patient is having an exacerbation, and to give the patient a recommended treatment plan. Revon have a 99-case validation set of data, in which each case was reviewed by nine Doctors; the ground truth is taken to be the majority ruling on any given patient case.

Glossary of terms

- **Confusion matrix:** a table used to describe classification performance.
- **Congesitve heart failure (CHF):** a weakness of the heart that leads to a buildup of fluid in the lungs and surrounding body tissues.
- **Exacerbation:** a sudden flare up of symptoms that requires a change in treatment plan.
- **Training set:** the data used to train the algorithm.
- **Triage:** the assignment of degrees of urgency to wounds or illnesses to decide the order of treatment of a large number of patients or casualties.
- **Validation set:** the data used to evaluate algorithm performance and not used in training.

2. Data description and feature extraction

The process of generating cases, training and validating the algorithm is explained in detail in [3]. Revon perform an exhaustive literature search alongside discussions with field experts to create a list of patient features (e.g. weight gain, edema, dyspnea) that are of highest importance when evaluating whether a patient is having an exacerbation. The final list of features is confirmed by doctors to contain all that they believe to be of importance when evaluating the state of the condition. The final feature space is split into patient profile, symptoms and vitals signs, and we give a brief example of the features in Table 1.

Category	Feature	Type
Patient profile	Age	years - continuous
	Ejection fraction	% - continuous
	Base weight	lb - continuous
Symptom profile	Symptoms worse?	categorical list
	Medication compliance	categorical list
	Current symptoms	categorical list
Vitals profile	Current weight	lb - continuous
	Current blood pressure	mmHg - continuous

Table 1 – Reduced feature list (total features $\tilde{1}30$)

More conversations are had with pulmonologists to gain an understanding of what typical values and ranges of these features are seen in cases of exacerbation. These rules are then used to generate synthetic cases that lie in, and well represent, the feature space of realistic cases. These synthetic cases are sent back to the pulmonologists who give their opinions on how well this represents reality and feedback what should be changed in the model. This process of generating cases, sending to pulmonologists and rewriting rules to generate improved cases, is performed until the experts are happy that the cases well represent what is seen in reality. After the feedback process the final set contains 2499 patient cases, of which 2400 are selected for training the algorithm (the training set) and 99 cases for validating (the validation set). The training set is shuffled and split into six sets of 400 cases, each of which is sent to a different doctor to review. Every case in the validation set is sent to each of these doctors plus an additional three doctors who did not provide any information into the training data, which gives a total of nine opinions on each of the validation cases. The doctors provided information on the following:

1) Profile, Symptom and Vitals features each ranked from 1-5 (least severe to most severe).
2) Exacerbation assessment, **Yes** or **No**.
3) Triage value of 1-4 where,
    1) **Okay** - No additional treatment required,
    2) **Plan** - Continue your medication plan as normal and check back in 1-2 days,
    3) **Doctor** - Call your physician,
    4) **ER** - Go to the emergency room.

The recipe for processing the data is as follows. The data are given in raw spreadsheet form. New features are generated from linear combinations of previous features. We find that features such as weight gain and O_2 saturation gain are highly predictive of both exacerbation and triage, and so we choose to include the current state and the gain, dropping the base value (as otherwise we will have collinear features). All continuous variables (barring age and BMI)

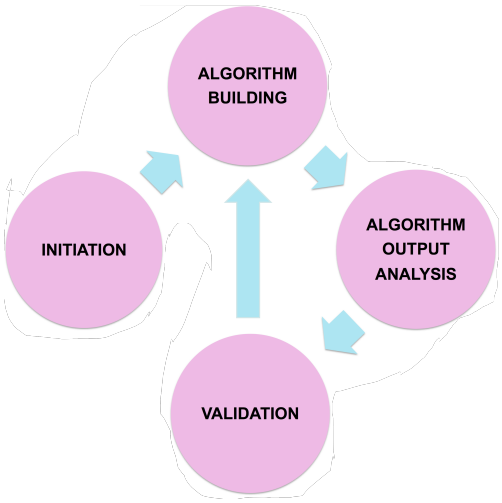


Figure 1 – Algorithm training process.

contain a number of unknown features to account for the fact that it will not always be the case that all patient information is known so there will be missing values. We bin the features using suitable bins, chosen through exploratory data analysis. These are optimised through a gridsearch cross validation method that we will explain later. Collinear features are removed through a combination of variance inflation factor (VIF) analysis and feature selection, where VIF is used to identify collinear features and then the most predictive of the collinear features are kept by evaluating the most important features via a feature selection method.

3. Algorithm training

The strategy we use to find the optimal prediction model is shown in Figure 1. Initially, several candidate supervised machine learning algorithms are selected, including support vector machines, logistic regression, naive Bayes, linear discriminant analysis, KNN, a variety of gradient boosted and ensemble decision tree methods, a multi-layer perceptron neural net, and a variety of different ensembles of such classifiers under both soft and hard voting rules to generate a final prediction.

Prediction of triage is a four-class classification problem, that is, we attempt to determine which of the classes {1, 2, 3, 4} the patient belongs. After a number of initial iterations, we observe that some algorithms (in particular linear models) were better at predicting 1s and 4s but less good at distinguishing between 2s and 3s, whereas other classifiers had the opposite problem. Thus we chose to train two classifiers: one to determine 1s and 4s from the collective 2s and 3s, then a second to determine 2s and 3s if the previous algorithm determined it as not being a 1 or 4. A five fold cross validation grid search is used to find the optimum hyperparameters for the classifiers. The top-performing algorithms of each class are selected based on how they perform when making predictions using a standard evaluation method [4]. We achieve best results using a linear discriminant analysis classifier for the prediction of 1s and 4s with an ensemble of gradient boosted trees and a Bernoulli naive Bayes classifer for the prediction of 2s and 3s. For prediction of exacerbations, the best score is achieved using a gradient boosted classifier with a reduced feature set generated via the Boruta feature selection method.

4. Results

We validate our algorithm by comparing our prediction of triage and exacerbation against consensus decision from a panel of physicians on the 99 hypothetical patient cases (the

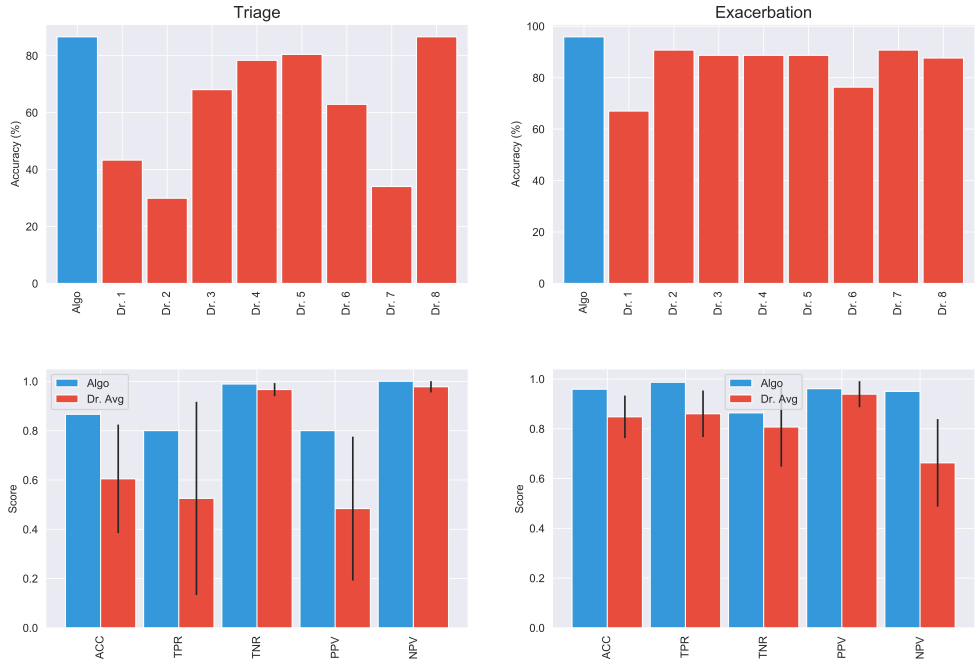


Figure 2 – Performance comparison of algorithm and individual physicians at predicting the consensus of the validation sets for triage (a), exacerbation (b) along with performance metric scores against the average doctor for triage (c) and exacerbation (d).

validation set). Each individual physician and the algorithm are tested for how often their particular recommendation for a patient case matched the majority opinion. In cases of ties, the more conservative medical decision (higher/more serious category) is accepted as the correct one. The 99 validation cases were removed from the case set prior to training, which made them statistically diverse, clinically relevant, and truly out-of-sample.

In Figure 2, we display the accuracy of the algorithm against all other doctors in the validation set, along with plots of different performance metrics against the average doctor. The accuracy here is taken to be the number of correct predictions out of the total number of cases. We see that the algorithm has accuracy equalling that of the top performing doctor and is significantly better than other doctors in the triage cases. Our algorithm also achieves the top score (with 95 of 99 cases correct) for exacerbation predictions. We also show, in Figures 3 and 4, the confusion matrices of the algorithm against the top performing doctor.

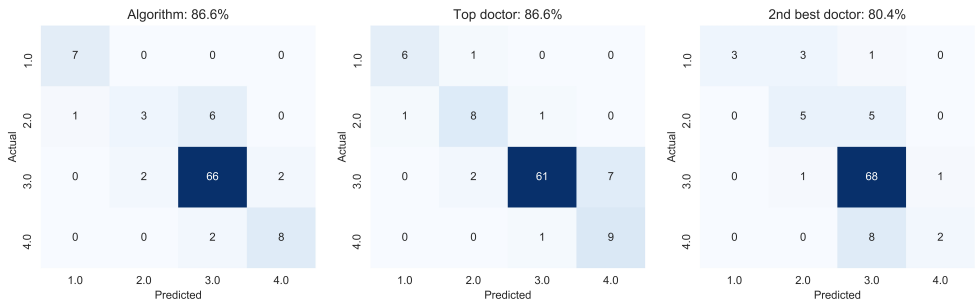


Figure 3 – Confusion matrices of triage prediction from the top algorithm against the top 2 performing doctors.

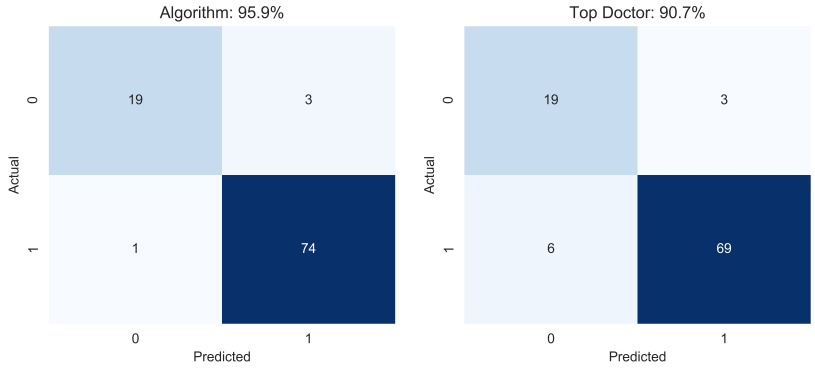


Figure 4 – Confusion matrices of exacerbation prediction from the top algorithm against best performing doctor.

Triage	Exacerbation
CurrentSymptoms_None	WeightGain(6, ∞]
SymptomsWorse?_No	WeightGain(2, 6]
O2SatGain(-∞, -4.0]	CurrentSymptoms_LegOrStomachSwelling
WeightGain(6.0, ∞]	CurrentSymptoms_Dyspnea
CurrentO2Sat(87.0, 89.0]	SymptomsWorse?_No
CurrentSymptoms_Dyspnea	O2SatGain(-∞, -4.0]
CurrentSymptoms_LegOrStomachSwelling	CurrentSymptoms_ChestPain
CurrentBPSystolic(-∞, 69.0]	AdditionalInfo_None
AdditionalInfo_Can'tGetOutOfBedOrOffTheCouch	CurrentSymptoms_None
CurrentO2Sat(-∞, 87.0]	CurrentO2Sat(92.0, 96.0]
AdditionalInfo_None	CurrentHR(45.999, 81.0]

Table 2 – Top 10 most important features as given from the algorithm.

A similar level of performance of the algorithm to the top physician can be seen from the confusion matrices. Both triage no one more then one value away from the majority, since all values lie within one position from the leading diagonal. The performance of the algorithm on the ER cases is better than that of the top doctor, since only two patients are overtriaged compared to seven (which is very costly). It does undertriage two 4s compared to the one undertriage by the top doctor, but if we analyse these cases they are both split decisions between 3 and 4 so the appropriate triage of these is unclear. The algorithm is also significantly better than the number two performing doctor. In terms of exacerbation, the algorithm is certainly better than the top doctor with two more correct classifications and a similar confusion matrix.

In Table 2, we list the most important features for predi ting triage and exacerbation. From the literature we know that these features are sensible and align with those known to be used by doctors to decide on exacerbation and triage decisions. This is extremely encouraging and suggests that the algorithm is finding sensible relationships in a manner close to that a doctor would and so gives evidence to the claim that this can provide a level of support equal (or better) to that of a doctor in an at-home situation.

5. Discussion, conclusions & recommendations

We have demonstrated the effectiveness of machine learning models to predict patient exacerbation and offer appropriate triage information when compared to a consensus physician opinion on a 99-case validation set. We have analysed a number of different performance metrics showing at least comparable and generally better performance of the algorithm when compared to the top performing doctor. From our feature importance study, we find that our algorithm puts higher weight to features that physicians consider to be the most important. While the algorithm is not intended to be a substitute for physician examinations, our study shows that they provide highly accurate at-home support which can direct patients to the right care.

The validation set is generated through a majority rule from nine expert opinions, whereas the training set has a single verdict for each entry. The training set labelling has not gone through this majority rule, and so is fundamentally different from the data in the validation set. If more data were collected so that a majority rule method could be applied to the training set, then the majority rule could be applied there also which we would then better represent the form of the data expected from the validation set and thus more accurate algorithms could be developed. This data is very expensive and time consuming to collect, and gathering another eight opinions on the training data for example will require over 19000 more case evaluation from doctors which is impractical. However, collection of three more opinions per case may be enough to provide significant improvement with a more feasible collection goal.

Our algorithm has so far only been tested on hypothetical patient cases. It would be important to prove the efficacy of the algorithm against real patient cases where a set of physicians will triage the same set of patients to form a new validation set and the algorithm can be tested against real patient data.

The model is set up to predict whether a patient is currently experiencing an exacerbation. However, if time dependent data could be gathered, a similar process could be set up to predict the likelihood of future exacerbations and to offer advice on to how best to avoid such things happening.

Sumanth Swaminathan, Chief Data Scientist of Revon said: *"Revon's primary mission is to revolutionize at-home care for patients with chronic illnesses. The approach we take uses machine-learning algorithms embedded in mobile applications to detect disease flare-ups and council patients to the right level of care at the right time. James' work focused on predicting exacerbations in heart failure patients. In his work, he eagerly embraced a predictive modelling project that required sophistication in machine-learning methods, programming, statistical analysis, and clinical medicine. His algorithms showed impressive out-of-sample prediction accuracy and will be the critical feature of our cardiovascular apps to be deployed for patient use in early 2019."*

References

[1] Jeff Voigt, M. Sasha John, Andrew Taylor, Mitchell Krucoff, Matthew R. Reynolds, and C. Michael Gibson. A reevaluation of the costs of heart failure and its implications for allocation of health resources in the united states. *Clinical Cardiology*, 37(5):312–321, 2014. ISSN 19328737. doi: 10.1002/clc.22260.

[2] Heart Foundation. Heart failure action plan and daily checks, 2018. URL `https://www.heartfoundation.org.nz/resources/heart-failure-action-plan`.

[3] Sumanth Swaminathan, Klajdi Qirko, Ted Smith, Ethan Corcoran, Nicholas G. Wysham, Gaurav Bazaz, George Kappel, and Anthony N. Gerber. A machine learning approach to triaging patients with chronic obstructive pulmonary disease. *PLoS ONE*, 12(11):1–21, 2017. ISSN 19326203. doi: 10.1371/journal.pone.0188532.

[4] Jason Brownlee. A Gentle Introduction to k-fold Cross-Validation, 2018. URL `https://machinelearningmastery.com/k-fold-cross-validation/`.