Application of Machine Learning to Support Self-management of Asthma with mHealth

Kevin C.H. Tsang[1]*, Hilary Pinnock[2], Andrew M. Wilson[3], and Syed Ahmar Shah[1]

***Abstract*— While there have been several efforts to use mHealth technologies to support asthma management, none so far offer personalised algorithms that can provide real-time feedback and tailored advice to patients based on their monitoring. This work employed a publicly available mHealth dataset, the Asthma Mobile Health Study (AMHS), and applied machine learning techniques to develop early warning algorithms to enhance asthma self-management. The AMHS consisted of longitudinal data from 5,875 patients, including 13,614 weekly surveys and 75,795 daily surveys. We applied several well-known supervised learning algorithms (classification) to differentiate stable and unstable periods and found that both logistic regression and naïve Bayes-based classifiers provided high accuracy (AUC $>$ 0.87). We found features related to the use of quick-relief puffs, night symptoms, frequency of data entry, and day symptoms (in descending order of importance) as the most useful features to detect early evidence of loss of control. We found no additional value of using peak flow readings to improve population level early warning algorithms.**

Keywords: Asthma, self-management, machine learning, mHealth, big data

I. Introduction

Asthma is a variable condition, affecting around 5.4 million people in the UK [1]. Every 10 seconds in the UK alone, someone has an asthma attack a few of which will be life-threatening [1]. The dual focus in the management of patients with asthma is on symptom control and the prevention of asthma attacks. The most common symptoms of asthma are wheezing, cough, chest tightness and shortness of breath. For most patients, the majority of the time, the condition is stable, manageable, and these symptoms are either absent or mild. However, after exposure to triggers (which vary between patients), these symptoms can get worse and lead to an attack (sustained worsening of symptoms that require emergency treatment such as oral steroids, or hospitalisation if not treated promptly).

Currently, there is no cure for asthma. However, existing management strategies, such as the use of "preventer" inhalers, can be used to control the condition. Supported self-management (including an action plan) significantly reduces the risk of an asthma attack [2]. A key component of self-management is monitoring which may be passive requiring minimal effort on the part of the patient (and thus preferable) or active requiring conscious effort which may be burdensome (or boring) reducing adherence. Mobile health (mHealth) offers a promising platform for combining passive and active approaches to deliver an engaging and effective self-management system for asthma.

There have been several efforts to use mHealth technologies to support asthma self-management [3]. myAsthma is a National Health Service (NHS) approved mobile app developed by My mHealth, which includes instructional videos about inhaler techniques, tracks symptoms and peak flow, provides local weather forecasts and stores action plans [4]. AsthmaMD has similar features, logging user asthma activity, peak flow, medications and triggers, has paperless action plan storage, and can provide custom notifications [5]. However, to date, no effective digital self-management solution for asthma exists that has been widely adopted. This is partly because existing solutions are rarely sufficiently engaging to enhance adherence to monitoring and also lack personalised algorithms to provide real-time tailored feedback based on symptoms and other parameters. Consequently, our long-term goal is to develop an effective and engaging mHealth system that facilitates self-monitoring and uses personalised algorithms to provide timely and appropriate feedback (early warning) to patients. To make progress towards achieving our goal, we have employed a publicly available mHealth dataset to apply and benchmark machine learning techniques to develop early-warning algorithms.

II. Methods

We first describe the Asthma Mobile Health Study (AMHS) dataset that was used in our study, followed by the methodology developed to analyse the data. The key methodological steps are "data pre-processing and labelling", "feature extraction", "feature selection", "classification", and "model evaluation". The flowchart in Fig. 1 provides an overview.

A. *Asthma Mobile Health Study*

We used the AMHS [6][7] dataset in our study. The AMHS was conducted via Asthma Health App, an Apple app designed using ResearchKit for the purpose of the study, and contains data (daily and weekly surveys, asthma history, demographic, EuroQol 5-dimensions 5-levels (EQ-5D-5L) survey, and location data) collected, often sporadically, by 5,875 US participants over 21 months. The nature of the study provided a wide geographic coverage across the US.

[1]K.C.H. Tsang and S.A. Shah are with the Asthma UK Centre for Applied Research, and Usher Institute of Population Health Sciences and Informatics, University of Edinburgh, Edinburgh, EH16 4UX, United Kingdom.

[2]H. Pinnock is with Asthma UK Centre for Applied Research, Allergy and Respiratory Research Group, and Usher Institute of Population Health Sciences and Informatics, University of Edinburgh, Edinburgh, EH8 9AG, United Kingdom.

[3]A.M. Wilson is with Asthma UK Centre for Applied Research, and Norwich Medical School, University of East Anglia, Norwich, NR4 7TJ, United Kingdom.

*Corresponding author email: `K.C.H.Tsang@sms.ed.ac.uk`

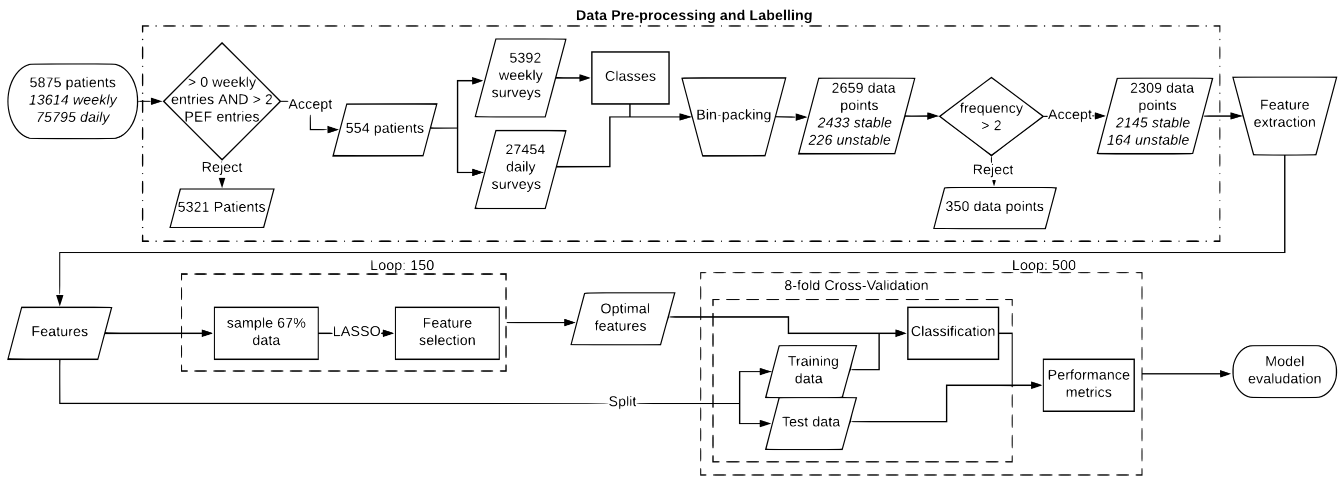

Fig. 1. Flowchart of data processing

Participants filled in daily and weekly questionnaires regarding their asthma. The AMHS had also obtained informed participant's consent through the app, which ensured they understood the risks, benefits, and options of study participation [7]. Furthermore, these data were made available for further research by the authors in 2018 [6].

B. Data Pre-processing and Labelling

The AMHS dataset contains data from 5,875 patients with 75,795 daily survey and 13,614 weekly survey entries. In this study, we analysed a subset of patients with at least one weekly survey entry and at least three daily survey entries with peak flow readings. Information about a patient's condition (stable or unstable) and symptoms were derived from the weekly and daily survey respectively.

The data selected from the daily survey used in the model included day and night symptoms, inhaler usage, asthma triggers and peak flow; see Table I for details of the daily survey questions. These are consistent with the criteria used clinically to assess asthma control in the Royal College of Physicians "3 Questions" [8][9] and defined by the Global Initiative for Asthma (GINA) [10].

The total number of asthma triggers were normalised per patient, because the number of self-reported triggers varied between patients and some may experience more symptoms. The peak flow values were also normalised per patient, this is common clinical practice; if a patient's peak flow drops below 80% of their best, they are said to be unstable [10].

1) Labelling Classes: We considered a patient to have been unstable in the period corresponding to a weekly survey if they had answered "true" in at least one of the three weekly survey questions describing an unscheduled use of healthcare resource in that week: seen asthma doctor other than a regular visit, visited the emergency room, or admitted to the hospital. We refer to these weekly survey entries as "unstable events".

We subsequently labelled periods corresponding to daily survey responses into stable, unstable and transient classes using the information from weekly surveys. All entries within a 14-day period before the unstable event were deemed to be in the unstable class. We also defined a 14-day recovery period (the transition period) beginning from the unstable event (see Fig. 2) (previous work identified a mean recovery time of approximately two weeks [11]). The patient was deemed to be in stable condition for the remainder of the time during monitoring.

TABLE I
DAILY SURVEY QUESTIONS TAKEN AS INPUT

Question	Input Type
In the last 24 hours, did you have any daytime asthma symptoms (cough, wheeze, shortness of breath or chest tightness)?	Boolean
In the last 24 hours, did you have any nighttime waking from asthma symptoms (cough, wheeze, shortness of breath or chest tightness)?	Boolean
Except for use before exercise, how many total puffs of your quick-relief medicine did you take over the past 24 hours?	Integer
Did any of the following cause your asthma to get worse today (check all that apply)?	Multiple choice 1-22 *
Enter your peak flow today (L/min)?	Integer
Did you take your asthma control medicine in the last 24 hours?	One of 4 **

* 1) A cold, 2) Exercise, 3) Being more active than usual (walking, running, climbing stairs), 4) Strong smells (perfume, chemicals, sprays, paint), 5) Exhaust fumes, 6) House dust, 7) Dogs, 8) Cats, 9) Other furry/feathered animals, 10) Mold, 11) Pollen from trees, grass or weeds, 12) Extreme heat, 13) Extreme cold, 14) Changes in weather, 15) Around the time of my period, 16) Poor air quality, 17) Someone smoking near me, 18) Stress, 19) Feeling sad, angry, excited, tense, 20) Laughter, 21) I don't know what triggers my asthma, 22) None of these things trigger my asthma
** 1) Yes, all of my prescribed doses, 2) Yes, some but not all of my prescribed doses, 3) No, I did not take them, 4) I'm not sure

2) Packaging Daily Survey Data: Patient adherence to monitoring is a major challenge in chronic disease management such as asthma, leading to sporadic patient data input patterns in mHealth studies [12]. Therefore, we employed a greedy bin-packing forward algorithm, which collated nearby similar (of the same class and separated by no more than two calendar weeks) 14-day periods to form larger blocks

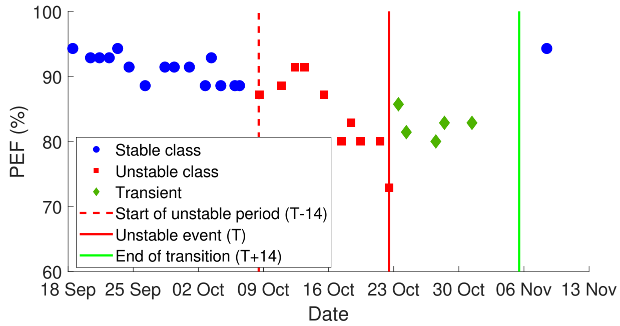

Fig. 2. Data class label surrounding unstable event

and provide more representative summary variables. The pseudo-code for the bin-packing algorithm is described in Algorithm 1. The maximum bin capacity C of the algorithm is the same as the initial time period used to class unstable data points; in the analysis, it was 14 data entries.

Algorithm 1: Bin-packing Algorithm

Data: Classed daily survey week of entry from one patient $data = \{(w_i^e, w_i^c, f_i, y_i)\}$, capacity C;
$w_i^e \in \mathbb{N}$: event week, number of unstable events preceding the entry,
$w_i^c \in [1, 53]$: calendar week of daily survey entry,
$f_i(\leq C)$: frequency of daily entries in calendar week,
$y_i \in$ {stable (0), unstable (1)}: class

begin
  Initialise bin load $f_{\text{bin}} = 0$, bin number $N = 1$;
  Sort data increasingly in w^e, then y, then w^c;
  Add first entry to bin: $f_{\text{bin}} = f_{\text{bin}} + f_1$;
  $y_{\text{bin}} = y_1$;
  $w^c_{\text{bin}} = w^c_1$;
  for $i = 2, \ldots, \#data$ **do**
    if $(y_i \neq y_{bin})$ *OR* $(3 \leq |w_i^c - w_{bin}^c| \leq 49)$ **then**
      New bin: increment N;
      $f_{\text{bin}} = 0$;
    end
    if $f_{bin} + f_i \leq C$ **then**
      Add to bin: $f_{\text{bin}} = f_{\text{bin}} + f_i$;
    else
      New bin: increment N;
      $f_{\text{bin}} = f_i$;
    end
    $y_{\text{bin}} = y_i$;
    $w^c_{\text{bin}} = w^c_i$;
  end
end

C. Feature Extraction

We performed linear regression on the set of daily survey data after the bin-packing, and estimated four summary variables per set, for each of the six daily survey items: mean, gradient of a linear fit, absolute gradient and R-squared (coefficient of determination). A linear fit was chosen to display the overall trend seen over a period of time. Fig. 3 illustrates linear regression applied to peak flow readings contrasting stable and unstable periods of a single patient in the study. The proposed feature extraction method allowed a direct comparison between periods of time despite the disparity in data availability and sampling frequency. In this work, we set a minimum frequency requirement of three data points before applying linear regression. Lastly, we also used the number of data points available in each bin (called frequency henceforth) as an additional feature providing us with 25 potential features (four summary variables from each of the six daily survey items, and frequency) in total.

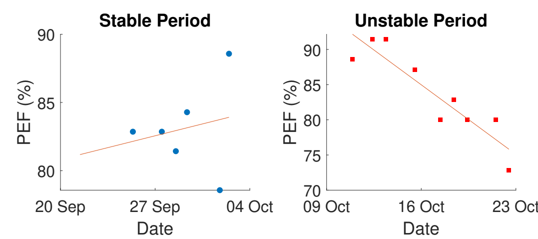

Fig. 3. Linear fit (feature extraction)

D. Feature Selection

The least absolute shrinkage and selection operator (LASSO) method was employed to rank and select a handful of input features. LASSO is a well-known technique for binary classification that combines fitting a cost function with regularisation. The regularisation performed in LASSO not only helps us avoid over-fitting but also helps in ranking input features based on their predictive power. By gradually increasing the amount of regularisation (by varying a single parameter), we can eliminate features in succession (so that features with less predictive power will be eliminated first) [13].

In this work, we varied the amount of regularisation in 100 steps (from parameter value of 0, corresponding to no regularisation and therefore using all features leading to an over-fitted model, to a heavily regularised model that knocks out all features). In order to determine an average ranking for the features, we repeated the LASSO procedure 150 times, using a different randomly selected 67% subset of the data each time. For each of these subsets, further subsets were made for ten-fold cross-validation. Cross-validation is a method to avoid over-fitting using data sub-sampling; k-fold cross-validation involves randomly splitting the data into k sets, then training on $k-1$ sets and testing on the remaining set [14].

E. Classification

This work aimed to predict if a patient is likely to have an unstable event. In the context of our study, this corresponds to developing an algorithm that could classify whether a patient is in a stable or in an unstable condition. We used a number of well-known machine learning classifiers, both linear and nonlinear, and both probabilistic and deterministic models. The models used in this study consisted of decision trees, logistic regression, naïve Bayes, and support vector machine (SVM). Decision trees find an optimal sequence of binary decisions on different features to classify the data, which creates a nonlinear decision boundary. Logistic regression learns a sigmoid-based discriminant function that uses a linear combination of features as input. Naïve Bayes is a probabilistic model and assumes independence between all variables; the largest posterior probability determines the class. SVM attempts to find the maximum margin hyperplane to separate data, which is a linear decision boundary [14].

For comparison, the base model used logistic regression and only the mean of the six daily survey data variables as features.

F. Performance Metrics

The area under the receiver operating characteristic curves (AUC-ROC) is a standard comparison metric for binary classification. The ROC curve reflects the obtainable balance between sensitivity (true positive rate (TPR), the proportion of unstable periods correctly classified) and the specificity (true negative rate (TNR), the proportion of stable periods correctly classified).

The skewed nature of the data made the geometric mean accuracy (GMA) a more suitable metric than accuracy (proportion of correct predictions over the total predictions) for identifying the optimal threshold to maximise the sensitivity and the specificity. A set of GMA can be evaluated from the points on the ROC curve, then the maximum of this set represents the best threshold for the given data.

$$\mathrm{GMA} = \sqrt{\mathrm{TPR} \times \mathrm{TNR}}, \tag{2}$$

III. Results

After pre-processing and labelling, the dataset used for subsequent analysis consisted of 2,309 periods with 2,145 (92.9%) corresponding to the stable class and 164 (7.1%) corresponding to the unstable class. The 2,309 periods amounted to 25,412 daily surveys (24,079 in stable class and 1,333 in unstable class) covering 55,509 days of patient monitoring; suggesting, on average, a patient completed a daily survey every 2.2 days. The flowchart in Fig. 1 shows the various steps of data processing.

A. Feature Ranking

The ranking associated with using all 25 features are in Table II. The median optimal model size from the 150 rankings was six, and all had used the same features. However, the rankings between the top three varied within the 150 rankings. The top six features in decreasing order of importance being: quick-relief puffs (mean), quick-relief puffs (absolute gradient), night symptoms (absolute gradient), night symptoms (mean), frequency, and day symptoms (mean).

The optimal model, according to the "one-standard-error" rule [15], used six features. Notably, the night symptoms were ranked higher than day symptoms, which has clinical face validity.

TABLE II

LASSO RANKING AND WEIGHT IN OPTIMAL MODEL

Rank	Feature	Weight
=1	quick-relief puffs (mean)	0.18
=1	quick-relief puffs (absolute gradient)	0.14
=1	night symptoms (absolute gradient)	0.17
4	night symptoms (mean)	0.14
5	frequency	-0.22
6	day symptoms (mean)	0.20
7	day symptoms (absolute gradient)	0
8	number of triggers (gradient)	0
=9	peak flow (absolute gradient)	0
=9	peak flow (gradient)	0

B. Classifiers

Each of the models were trained and tested using the eight-fold cross-validation dataset. The cross-validation was repeated for 500 times, to observe the behaviour of the models over different cross-validation sets. Note that the ROC curve was dependent on the training-test set segmentation for cross-validation. Fig. 4 shows the distribution of the model performances over 500 separate cross-validations. The SVM model varied most over the different validation sets. The median performance figures of 500 evaluations are listed in Table III.

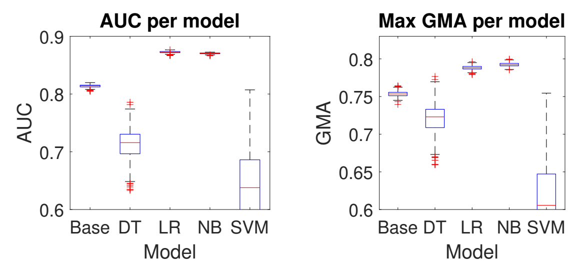

Fig. 4. Boxplots of AUC and maximum GMA of models over 500 evaluations, the best models were LR and NB.

TABLE III

PERFORMANCE METRICS

Classifier	AUC	GMA	Sensitivity	Specificity
Base	0.814	0.752	0.805	0.702
DT	0.716	0.723	0.683	0.766
LR	0.873†	0.788	0.817	0.760
NB	0.871	0.792†	0.866	0.725
SVM	0.638	0.606	0.591	0.620

†maximum in column

For better understanding of the characteristics of the classifiers, consider the weights used in the logistic regression model, see Table II. The value of the weight determines how much a variable affected the decision. The sign of the weight corresponds to the direction of the relation to the classes. For example, night symptoms (mean) has a positive weight; thus, more symptoms were correlated to a higher likelihood of being in the unstable class.

C. Top Classifier

Using median GMA as criteria, the naïve Bayes classifier was the best performing algorithm followed closely by logistic regression. The median model over 500 evaluations was the most representative naïve Bayes classifier. Each point on the ROC corresponds to selecting a different threshold for classification. In this study, we considered the optimal threshold to be the one with the largest GMA.

The optimal model had a GMA of 0.792, confusion matrix displayed in Table IV, with a sensitivity of 0.866, a specificity of 0.725, an AUC of 0.871, and the ROC curve displayed in Fig. 5.

TABLE IV
CONFUSION MATRIX NAÏVE BAYES MODEL

		Predicted Class	
		stable	unstable
True Class	stable	1555	590
	unstable	22	142

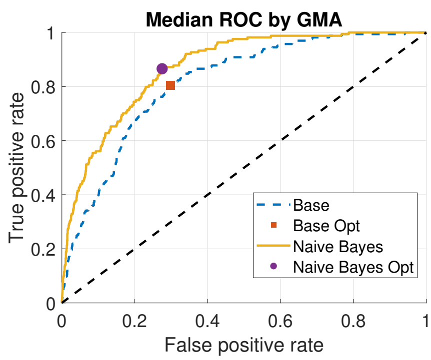

Fig. 5. ROC curves of the base model vs best model determined by highest median GMA (naïve Bayes)

IV. DISCUSSION AND CONCLUSION

Using the AMHS dataset we provided evidence of the utility of machine learning methods to aid asthma self-management, where the results of the data-driven methods aligned with clinical understanding. More specifically, we found that both a probabilistic (naïve Bayes) and a discriminant (logistic regression) classifier could provide high accuracy (AUC > 0.87) for early warning. Our work has shown the potential of the method which collates irregularly sampled data to form summary variables and allowed comparison between periods of time with different data availability. This work also demonstrated that the performance of prediction models could be further enhanced with features that capture gradient over a time period. Besides, this work also found that features using peak flow readings did not provide additional value over and above other self-reported features used in the study (derived from the use of quick-relief puffs, night and day symptoms).

A limitation in the data was the identification of the unstable event, using the weekly survey, with the resolution of an unstable event being a week. Also, the app was limited to Apple device users, who are not representative of the whole US population, thus introducing potential bias in the data.

Our study used self-reported data, and we anticipate further improvement over the reported performance if objective measures were available as feature inputs for prediction. Our future work aims to expand the analysis of this dataset by using the 3-digit ZIP code prefixes to link historical weather data, building demographic, geographic, and seasonal sub-models, exploring links between emotions and symptoms, and testing more complex models. Moreover, the data in the transition period between unstable and stable periods and temporally outlying data were not used in this analysis; future models could incorporate these data points using multi-scale models.

ACKNOWLEDGMENT

These data were contributed by users of the asthma app as part of the Asthma Mobile Health Application study developed by The Icahn School of Medicine at Mount Sinai and described in Synapse [doi:10.7303/syn8361748].

REFERENCES

[1] Asthma UK. (2019, Sep. 27). Asthma facts and statistics [Online]. Available: https://www.asthma.org.uk/about/media/facts-and-statistics/

[2] H. Pinnock, H.L. Parke, M. Panagioti, L. Daines, G. Pearce, E. Epiphaniou, et al., "Systematic meta-review of supported self-management for asthma: a healthcare perspective," BMC Medicine, vol. 15, no. 1, pp. 64, Mar. 2017.

[3] P. Tinschert, R. Jakob, F. Barata, J. Kramer, and T. Kowatsch, "The potential of mobile apps for improving asthma self-management: a review of publicly available and well-adopted asthma apps," JMIR mHealth and uHealth, vol. 5, no. 8, pp. e113, Aug. 2017.

[4] My mHealth. (2020, Jan. 06). myAsthma [Online]. Available: https://mymhealth.com/myasthma

[5] AsthmaMD. (2020, Jan. 06). AsthmaMD [Online]. Available: https://www.asthmamd.org/

[6] Y. F. Y. Chan, B. M. Bot, M. Zweig, N. Tignor, W. Ma, C. Suver, et al., "The Asthma Mobile Health Study, smartphone data collected using ResearchKit," Scientific Data, vol. 5, pp. 180096, May 2018.

[7] Y. F. Y. Chan, P. Wang, L. Rogers, N. Tignor, M. Zweig, S. G. Hershman, et al., "The Asthma Mobile Health Study, a large-scale clinical observational study using ResearchKit," Nature Biotechnology, vol. 35, no. 4, pp. 354, Apr. 2017.

[8] M. G. Pearson, and C. Bucknall, Measuring clinical outcome in asthma: a patient-focused approach. Clinical Effectiveness & Evaluation Unit, Royal College of Physicians, 1999.

[9] British Thoracic Society. (2019, Jul. 01). SIGN158 British guideline on the management of asthma [Online]. Available: https://www.sign.ac.uk/assets/sign158.pdf

[10] Global Initiative for Asthma. (2019, Jun. 01). Global strategy for asthma management and prevention [Online]. Available: www.ginasthma.org

[11] C.G. Jung, and H.S. Park, "Factors predicting recovery from asthma exacerbations," Allergy, Asthma & Immunology Research, vol. 8, no. 6, pp. 479, Nov. 2016.

[12] S. Hamine, E. Gerth-Guyette, D. Faulx, B.B. Green, and A.S. Ginsburg, "Impact of mHealth chronic disease management on treatment adherence and patient outcomes: a systematic review," Journal of Medical Internet Research, vol. 17, no. 2, pp. e52, Feb. 2015.

[13] R. Tibshirani, "Regression shrinkage and selection via the lasso," Journal of the Royal Statistical Society: Series B (Methodological), vol. 58, no. 1, pp. 267-288, Jan. 1996.

[14] C. M. Bishop, Pattern recognition and machine learning. Springer, 2006.

[15] T. Hastie, R. Tibshirani, and J. Friedman, The elements of statistical learning. New York: Springer Series in Statistics, 2001.