Classification of biophysical changes during food allergy challenges

Niall Twomey[1], Stephen Faul[1], Deirdre Daly[2], Jonathan O'B Hourihane[2], William P Marnane[1]
[1] Department of Electrical and Electronic Engineering, University College Cork, Ireland
[2] Department of Pediatrics and Child Health, University College Cork, Ireland
Email: niallt@rennes.ucc.ie

***Abstract*—This paper details the process of oral food challenges ('allergy tests') and steps followed to investigate whether automatic classification of the tests is possible. It has been observed by trained staff that the mood and physiological signals of a subject being tested for allergies can change during the test if they are sensitive to the allergen they are being tested against. Data from thirteen subjects was recorded, and thirteen features were extracted from each of these datasets. The changes in the features were then analysed over the course of each test. It was noted that when a subject failed the challenge, some of the features extracted from the ECG trace changed suddenly near the time that the test was stopped. Threshold classification was employed, and ROC curves were generated. Some features gave rise to ROC areas of over 0.97 on certain subjects. An average ROC area of 0.57 was computed over all subjects and all features due to wide inter subject variability.**

Keywords – allergy, ecg, hrv, feature extraction, classification

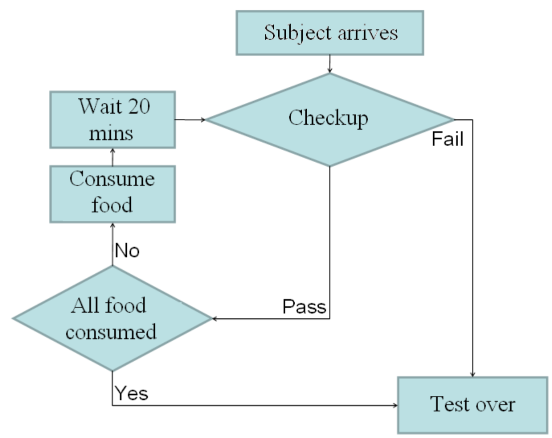


Fig. 1: Flowchart of oral food challenge

I. INTRODUCTION

The gold-standard in allergy classification is the Oral Food Challenge (OFC) [1]. OFCs involve the controlled ingestion of a food type that might provoke an allergic reaction on a subject. Skin and blood tests are performed prior to an OFC to indicate the sensitivity of a subject to the food type, but these are not always accurate.

A portion of food is prepared for the challenge, and is divided into incrementing sizes ($\frac{1}{16}$, $\frac{1}{8}$, $\frac{1}{4}$, etc). Starting with the smallest it is consumed piece by piece in periodic intervals until it has been consumed fully, or until a subject reacts to the food type. Figure 1 shows the typical flow of the challenges. If a subject reacts to the food they are being tested against they have failed the test, otherwise they have passed. Typical ages for the subject being tested are between six months and ten years.

It has been observed by nursing staff performing the challenges that the behavior and heart rate of subjects who fail the test can change during the test. A platform for monitoring the subjects and classifying potential allergies during the challenges is proposed. Early classification of a reaction to the potential allergen would reduce the stress on the subject and their family, reduce the discomfort of the child being tested, and reduce the time spent in the hospital for the test.

For data aggregation the SHIMMER [2] (Sensing Health with Intelligence, Modularity, Mobility, and Experimental Reusability) wireless mote was used to transmit the ECG signal trace via a Bluetooth link. The SHIMMER was programmed with custom-written firmware and it streamed raw two-channel ECG signals to a PC for storage.

Heart rate variability (HRV) is a measure of the changes in heart rate over time. With knowledge of the times at which the heart beats, various HRV features can be extracted from the signals, characterising the trace with various classes of feature (e.g. time and frequency domain features, etc). Thirteen features were extracted from data recorded during food challenges.

Classification was then employed to automatically detect the presence of an allergic reaction.

II. PROCEDURE

A. Hardware

The SHIMMER wireless device was employed to collect the data for the oral food challenges. Ethical approval was attained for the use of the SHIMMER in a hospital environment, and informed parental consent was also obtained on a subject-by-subject basis.

The SHIMMER was held onto the body via a custom-made and child-friendly neoprene strap and the ECG leads were configured in the Eindhoven triangle configuration. The ECG signal trace was sampled at 256 Hz and the signals were streamed to a PC via a Bluetooth connection.

TABLE I: Subject characteristics

Index	Gender	Age	Allergen	Result
1	Male	1.5 years	Wheat	FAIL
2	Male	6 years	Peanut	
3	Male	9 years	Egg	
4	Male	13 months	Milk	
5	Male	8 years	Peanut	
6	Female	9 years	Peanut	
7	Male	6 years	Egg	PASS
8	Male	10 years	Egg (cake)	
9	Female	4 years	Soy	
10	Male	6 years	Peanut	
11	Female	1.5 years	Milk	
12	Female	7 months	Milk	
13	Male	13 months	Milk	

B. Subjects

The ECG of thirteen subjects was recorded. The result of the oral food challenge is cast as either a pass or fail. To pass the food challenge means that the subject is not allergic to one portion of the type of food they were tested against. To fail the food challenge means that the subject is allergic to the food type and that it should be avoided in the subject's diet. Out of the thirteen subjects in this study six failed their tests and seven passed. Table I groups the gender, age, test food and challenge result of the subjects.

The procedure of the food challenges is shown in Figure 1. The subjects of the test are generally thought to have reacted to a food type before a food challenge is recommended. They are given skin and blood tests to gouge their sensitivity towards the food type in question, but the gold standard means of determining the sensitivity is the oral food challenge – the controlled ingestion of the supposed allergen. The goal of the challenge is for the subject to consume a full portion of the food (without reacting) in incremental sub-portions until the whole is reached.

When the subject arrives and after an initial checkup, the first portion of the food is given to the subject. The first portion is always the smallest, and in some cases it can simply involve rubbing of the food in question on the subject's lips. After the portion has been consumed the subject rests for 20 minutes under the supervision of staff. After this time period (or during the time period if the subject appears to be reacting badly to the food type) the subject's heart rate, blood pressure and oxygen saturation are checked. The test may be stopped if the results of the checkup are not satisfactory, and the subject has then failed the test. Otherwise the portion of food is doubled. This procedure is repeated until a portion is fully consumed, after which the subject waits for one last 20 minute period. If they have not reacted they have passed the test.

C. Annotations

While the data was being recorded annotation files were created in real time (by consulting with the nursing staff overseeing the procedure), making note of the times checkups were performed and some of the physiological results of the checkups, etc. Once the tests were complete each QRS point was manually labeled on the ECG trace. This was done in order to verify that any changes in heart rate variability features in the analysis in later sections were as a result of changes in the heart rate rather than an inaccuracy resulting from an automatic QRS detector.

The time the test was called a fail will be termed the 'call time' in this paper. Prior to this time is a second time where the reaction itself has begun but the test has not yet been stopped, termed 'reaction time'. Determination of this time is important for classification performance analysis.

D. Features

With the recorded data and the extracted QRS points various features were extracted. The features that were extracted were chosen due to their inclusion by the task force on heart rate variability analysis [3], and they are tabulated below.

- Time-domain features:
  - Mean heart rate
  - Standard deviation of heart rate
  - Coefficient of variation of heart rate
  - RMSSD of heart rate
    Root mean squared successive difference.
  - PNN25/PNN50 of heart rate
    Percentage of successive QRS points that differ by more than 25ms or 50ms.
  - Sequential trend analysis
  - Poincar CSI/CVI The cardiac sympathetic index and the cardiac vagal index.
- Frequency-domain:
  - Low frequency power
  - High frequency power
  - High to low power ratio

The frequency domain features were extracted by using the Lomb periodogram [5]. This allows frequency domain powers to be extracted from non-periodically sampled signal traces, such as the heart rate. It is defined by (1).

$$P_x\left(f\right) = \frac{1}{2\sigma^2}\left\{ \frac{\left[\sum_{n=1}^{N}(x(t_n)-\overline{x})\cos(2\pi f(t_n-\tau))\right]^2}{\sum_{n=1}^{N}\cos^2(2\pi f(t_n-\tau))} + \frac{\left[\sum_{n=1}^{N}(x(t_n)-\overline{x})\sin(2\pi f(t_n-\tau))\right]^2}{\sum_{n=1}^{N}\sin^2(2\pi f(t_n-\tau))} \right\} \quad (1)$$

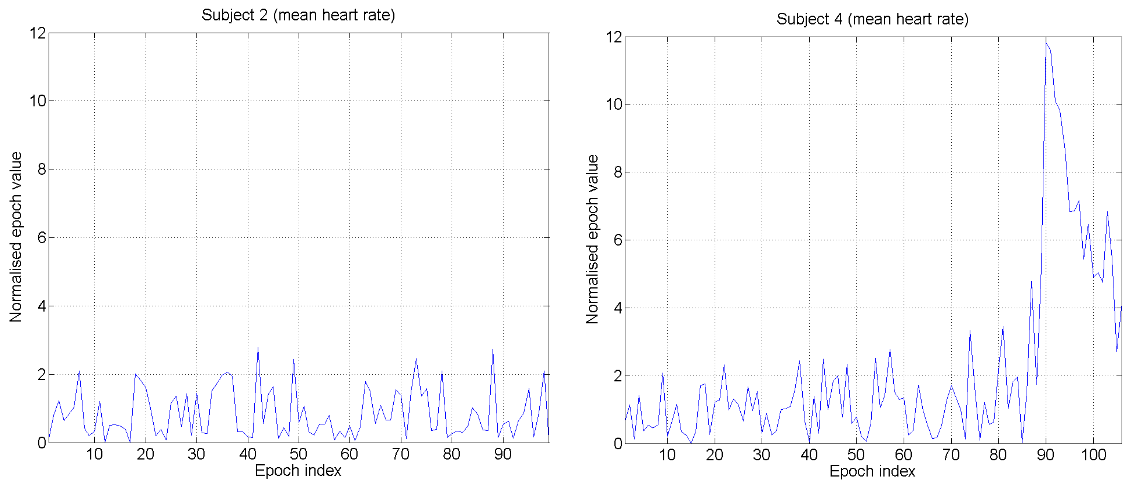
where $\bar{x}$ and σ^2 are the mean and variance of the series. τ makes the series insensitive to time shift, and is defined by (2).

$$\tau = \frac{1}{4\pi f}\arctan\left(\frac{\sum_{n=1}^{N}\sin(4\pi f t_n)}{\sum_{n=1}^{N}\cos(4\pi f t_n)}\right) \quad (2)$$

Each feature was extracted over one minute, non-overlapping epochs. After feature extraction the features were normalised, for classification purposes. A wide variability in subject heart rate was seen, with the resting heart rate of some subjects being less than 70 while a resting heart rate of over 110 was seen with others. Due to this variation the features were self-normalised. The first ten minutes of the extracted features from each subject were taken as a normalisation baseline. The mean (μ) and standard deviation (σ) of this normalisation period were calculated. The feature (f) was normalised (to $\hat{\mathrm{f}}$) according to the following formula:

$$\hat{\mathrm{f}} = \mathrm{abs}\left(\frac{\mathrm{f}-\mu}{\sigma}\right) \quad (3)$$

This normalisation allows the same threshold to be used for every subject in the classification stage as each feature was normalised by the feature values at rest.



(a) Subject 2, heart rate

(b) Subject 4, heart rate

Fig. 2: Intersubject difference in feature changes

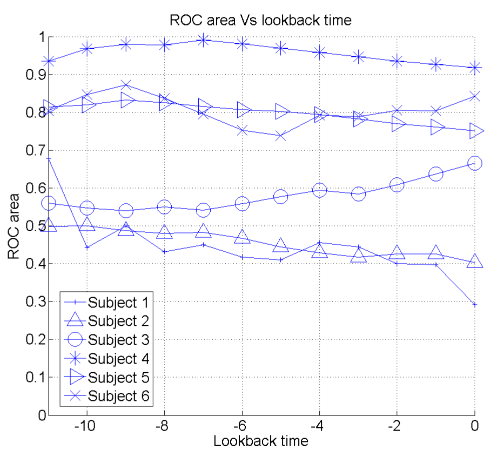


Fig. 3: Changes in ROC area for the mean heart rate as a function of the lookback time

E. Signal trends

The traits of the features were investigated at this point, and a range of characteristics were observed.

For the subjects who failed there were some occur

Two subsets of failing subjects were discovered. The physiological signals of one set (subjects 1, 2 and 3) did not vary noticeably over the course of the test, but the same signals of the second set (subjects 4, 5, and 6) did vary strongly.

There was no noticeable rise in heart rate or any associated HRV features for the subjects who passed the oral food challenges.

F. Classification

1) ROC curves: Receiver operating curves (ROC) were used to compute the performance of the classification. Thresholding classification was employed, which involves sweeping a threshold over the range of a feature. If any point of the feature is above the threshold they are classified as 'fail' (i.e. the classifier has classified the subject as failing the test), otherwise it is classified as a 'pass'.

For each feature true positives (TP) are the count of the number of times the feature has risen over the threshold after the reaction time. False negatives (FN) are the count of the number of times the feature remained below the threshold after the reaction time. True negatives (TN) are the number of times the feature remained below the threshold before the reaction time. False positives (FP) are the number of times the feature rose above the threshold before the reaction time. From these figures sensitivities and specificities can be computed.

Sensitivity (Sn) and specificity (Sp) are measures of the percentage accuracy of classification. The sensitivity of a classification is a measure the percentage of correctly classified fails, and the specificity is a computation of the percentage of correctly classified passes. They are defined in (4).

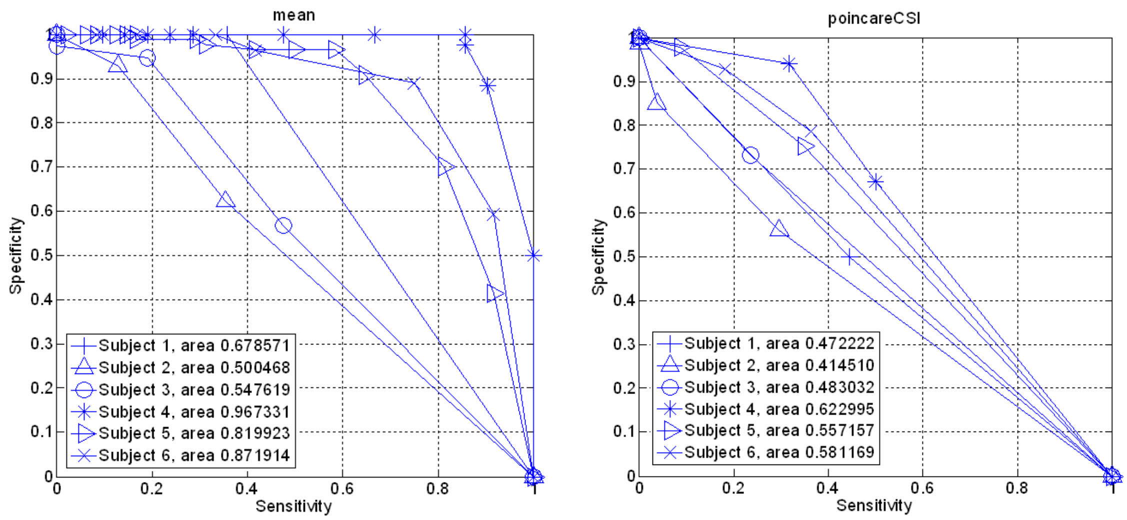
$$\mathrm{Sn} = \frac{\mathrm{TP}}{\mathrm{TP+FN}}, \quad \mathrm{Sp} = \frac{\mathrm{TN}}{\mathrm{TN+FP}} \tag{4}$$

When sweeping the threshold over the range of the feature various sets of sensitivities and specificities are computed. The ROC curves are the set of sensitivities and specificities that were computed. Computing the area below the ROC curve is indicative of the performance of the classifier, and Figures 4a and 4b show a sample of a set of ROC curves.

2) Reaction time: Due to the small number of subjects in this study it was important not to bias the results with subject-dependent processing, in particular when determining the reaction time for each subject. To remove this bias, a leave-one-out strategy was adopted.

To discover the reaction time for each subject a lookback search was performed. A variable reaction time was modified from the call time to the time of the preceding checkup in steps of one epoch. This reaction time was then used to compute a set of ROC areas for every feature. When one subject was being investigated for one feature, six sets of ROC areas were

978-1-4244-8132-3/10/$26.00 ©2010 IEEE



(a) Mean heart rate ROC curve

(b) High frequency component ROC curve

Fig. 4: Best and poorest performing features

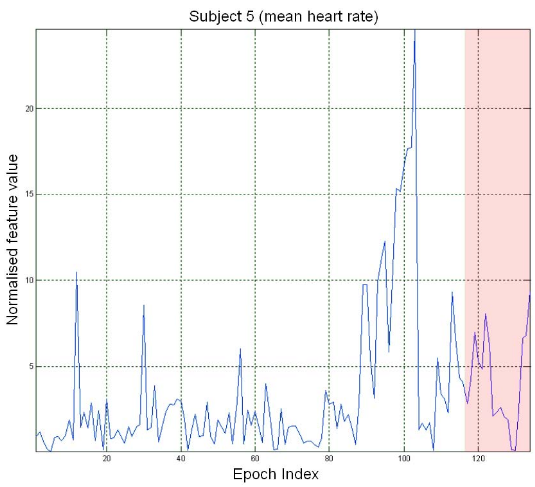


Fig. 5: Early detection (subject 5)

TABLE II: Feature ranking

features	mean	max	min	std
mean	0.731	0.9673	0.5005	0.1861
seqTrendNN	0.6534	0.7943	0.5029	0.1143
pnn25	0.6445	0.9632	0.4442	0.1808
nn50count	0.627	0.9656	0.5162	0.1722
poincareCVI	0.595	0.7446	0.4937	0.0843
pnn50	0.5881	0.693	0.4103	0.1187
lombFL	0.5618	0.7436	0.3829	0.1509
lombFVL	0.56	0.6331	0.4731	0.0758
seqTrendPP	0.5522	0.7807	0.4167	0.1247
poincareProduct	0.544	0.734	0.2917	0.1644
histogram	0.5425	0.6201	0.3889	0.0847
lombPflPfh	0.5364	0.6285	0.3749	0.0895
rmssd	0.5275	0.8005	0.1786	0.2098
lombFH	0.5218	0.623	0.4145	0.0781
cv	0.5131	0.6818	0.3704	0.1041
std	0.5125	0.5887	0.3077	0.1069
poincareCSI	0.4944	0.5951	0.3523	0.1041
Overall	0.5700	0.7407	0.4060	0.1253

available. The set of ROC areas of the five subjects that were not being investigated were averaged, and the lookback time which maximised the average ROC area was used as the reaction time for the sixth subject. The change in ROC area with respect to reaction can be seen in Figure 3.

G. Results

The mean ROC area over all features was calculated as 0.57. The mean heart rate performed best of all the features with an average ROC area of 0.731, but it can be seen from the standard deviation value in the first row in Table II and Figure 4a that the consistency of the feature is very subject-dependent – high standard deviations imply that there is a wide variance between the figures computed from each subject, and in Figure 4a there are two distinct groups of ROC curves giving areas of ~ 0.5 and ~ 0.8.

The mean heart rate is the best performing feature. The maximum ROC area for subjects 1, 2 and 3 was under 0.54, while the minimum ROC area for subjects 4, 5 and 6 was over 0.82, see Figure 4a.

Similar inter-patient variabilities presented with the other features. The poorest performing feature calculated was the cardiac sympathetic index of the Poincare analysis, see Figure 4b. The first three subjects (1, 2 and 3) present with ROC areas of between 0.41 and 0.46, but the other subjects present with distinctly larger areas of between 0.55 and 0.65. These values are low and ROC areas around this region might be considered a random choice.

This gap in performance between subjects 1, 2 and 3 and 4, 5, and 6 is consistently shown over the extracted features.

H. Temporal resolution

The result of the oral food challenges is the gold standard in allergy detection. However, the gold standard does not provide

978-1-4244-8132-3/10/$26.00 ©2010 IEEE

temporal information about how the reaction is progressing during the course of the test – i.e. there is no 'fuzzy' representation of a subject's current state of reaction. This means that some points flagged as false positives may not be false in reality.

Figure 5 show a time domain plot of the mean heart rate over time for subject 5. The highlighted region denotes the time period after the reaction time. In the un-highlighted zone there are sustained rises in heart rate. With subject 5 there was a normalised increase of over 20. At most thresholds these points would have been flagged as false positives due to their height and that they occurred before the reaction time. However, as the subject failed the food challenge it is logical to assume that these sustained changes in heart rate were as a result of a reaction to the food type. This effect leads to potentially artificially low classification results. This trait also presents with subject 6.

III. Discussion

A. Gold standard

The result of the oral food challenges is a binary classification of a subject's sensitivity to a food type at the end of the challenge. The results in Section II-G were calculated by counting the number of true and false positives and negatives that occurred at various thresholds before the reaction time. The lack of temporal resolution, discussed in Section II-H, results in an increase of the count of false positives, but with fore-knowledge of the result they might be safely classed as true positives. It is not possible to use this knowledge without biasing the results, but it should be noted that the results may be artificially low due to this.

There are two instances (with subjects 5 and 6) where it is suspected that the test could have been halted prior to the time it eventually was stopped. Had the features been analysed in real time, subject 5's test may have been stopped up to 30 minutes sooner, and subject 6's test may have been stopped 6 minutes sooner. The result of this would have been to ease the stress of the test on the subject and their family.

B. Epoch characteristics

The analysis of the features was done on an epoch-by-epoch basis. The epoch was chosen as 60 seconds in length, as this is a common epoch length in heart rate variability analysis. Variation of the length and the overlap of the epochs might have merit to investigate in future analyses. Varying epoch sizes may better emphasise changes in traits of some features more than others, while overlapping epochs will give better temporal resolution of the extracted features over every recording. This will increase the resolution of the ROC curves, and might allow better localisation of the onset of the allergic reaction.

C. Feature selection and classification

Table II tabulates the performance of every feature that was analysed. Some features have very high ROC areas with specific subjects, while others do not perform well. Adding feature selection classification might allow higher ROC areas to be achieved. Owing to the amount of inter-patient variability, see section II-G, feature selection may be able to exploit changes in some features to better classify the results of the test.

The classification that was employed in Section II-F is a primitive classifier. Advanced classification could be able to fuse the changes in various features together to improve the overall classification routine.

IV. Conclusion

It is possible to automatically classify the onset of allergic reactions in oral food challenges.

The quality of the classification varies between subjects and between extracted features. That the ROC areas for the mean heart rate over an epoch were over 0.8 with subjects 4, 5 and 6 suggests that this is a very strong candidate for the automatic classification of the food challenge. The mean heart rate is the best performing single classification feature that was extracted with an average ROC area of 0.742 over the six subjects. Six other features present with an ROC area near and over 0.6.

The ROC areas that were calculated may be artificially low in a number of cases. This is due to how the sensitivity/specificity is calculated. It is difficult to define the exact time at which the reaction has occurred. Due to this difficulty some segments which might be true positives (Figures 5 and 5) are flagged as false positives.

Improvements will be introduced with sophisticated classification and feature selection, and further research into improving the classification is justified due to the excellent performance of a number of the features over some of the subjects.

Acknowledgment

This work is supported by Science Foundation Ireland (SFI/07/SRC/I1169), and an enterprise partnership scheme with IRCSET and the TRIL Centre.

References

[1] Bindslev-Jensen C, Ballmer-Weber BK, Bengtsson U, Blanco C, Ebner C, Hourihane J et al. Standardization of food challenges in patients with immediate reactions to foods–position paper from the European Academy of Allergology and Clinical Immunology. Allergy 2004; 59(7):690-7.

[2] SHIMMER research homepage available at: http://www.shimmer-research.com/

[3] C.M.A. van Ravenswaaij-Arts, L.A.A. Kollee, J.C.W. Hopman, G.B.A. Stoelinga, and H.P. van Geijn, Heart Rate Variability, Annals of Internal Medicine, vol. 118, Mar. 1993, pp. 436-447.

[4] A. Temko, E. Thomas, G. Boylan, W. Marnane, and G. Lightbody. An SVM-based system and its performance for detection of seizures in neonates. In Engineering in Medicine and Biology Society, 2009. EMBC 2009. Annual International Conference of the IEEE, pages 26432646, Sept. 2009.

[5] N. Lomb. Least-squares frequency analysis of unequally spaced data. Astrophysics and Space Science, vol. 39, pp. 447462, 1976.

978-1-4244-8132-3/10/$26.00 ©2010 IEEE