ACOUSTIC ANALYSIS AND ASSESSMENT OF THE KNEE IN OSTEOARTHRITIS DURING WALKING

Costas Yiallourides[1], *Alastair H. Moore*[1], *Edouard Auvinet*[2], *Catherine Van Der Straeten*[2] *and Patrick A. Naylor*[1]

[1]Department of Electrical and Electronic Engineering, Imperial College London, UK

[2]MSk Lab, Department of Surgery and Cancer, Imperial College London, UK

ABSTRACT

We examine the relation between the sounds emitted by the knee joint during walking and its condition, with particular focus on osteoarthritis, and investigate their potential for non-invasive detection of knee pathology. We present a comparative analysis of several features and evaluate their discriminant power for the task of normal-abnormal signal classification. We statistically evaluate the feature distributions using the two-sample Kolmogorov-Smirnov test and the Bhattacharyya distance. We propose the use of 11 statistics to describe the distributions and test with several classifiers. In our experiments with 249 normal and 297 abnormal acoustic signals from 40 knees, a Support Vector Machine with linear kernel gave the best results with an error rate of 13.9%.

Index Terms— knee osteoarthritis, acoustic signals, feature extraction, discriminant analysis, classification

1. INTRODUCTION

Clinical detection of knee Osteoarthritis (OA) relies on a combination of patient reported symptoms and medical imaging. Magnetic Resonance Imaging (MRI) is a popular method that provides clear images of the joint but it is expensive and is unable to capture the knee during functional activity. X-ray imaging has the same limitation and is also problematic due to radiation exposure. Hence, new objective methods which are sensitive, low-cost and risk-free are required for the detection of pre-clinical OA to improve the selection of patients requiring further clinical testing, reducing therefore associated costs, and to facilitate effective intervention toward disease management. In normal, well lubricated knee joints, a protective space separates the bones which have smooth surfaces due to a layer of cartilage [1]. They move freely and the level of sound emitted is low. In OA knees the lubrication process is degraded and the protective space reduces, resulting in increased friction that accelerates the wear of cartilage [2]. This increased friction makes the knee more noisy during motion.

Using knee sounds for diagnostic purposes is well documented in the literature. The movement protocols most often used were knee flexion-extension and sit-to-stand. Auscultation based Phonoarthrography (PAG) utilises acoustic microphones in the audible frequency range to record sounds. Work on PAG reported that the spectral activity of pathological knees spanned the entire audible frequency range and the acoustic power increased with severity of cartilage damage [3–5]. Significant work was directed to the development of Vibroarthrography (VAG) as an alternative to PAG which relies on accelerometers which are sensitive at frequencies below 1 kHz, to pick up mechanical vibrations. Several algorithms were proposed for classifying the knee VAG signals, according to pathological conditions, using linear prediction modelling [6, 7], time-frequency analysis [8] and wavelet decomposition [9]. Features used include waveform variability parameters, spectrogram features, statistical features [10], fundamental frequency, mean amplitude of pitches and their jitter and shimmer [11, 12]. Various classifiers were considered, from early neural network architectures [8, 13] to maximal posterior probability decision criterion [14], bagging ensemble and multiple classifier system based on adaptive weighted fusion [9]. Recently the use of Acoustic Emission (AE) in the ultrasound frequencies for assessing knee joints was explored [15]. It was demonstrated, using Principal Component Analysis, that AE measures of healthy and OA knees form separate clusters and concluded that the latter produce substantially more AE events with higher peak magnitude and average signal level [16, 17].

The novel contributions of our work are the use of: 1) dynamic functional activity which is essential for understanding pathology development and progression and 2) the Bhattacharyya distance extended for gamma distributions for statistical analysis with a comprehensive experimental validation of the discriminant power of various features including modulation and pulse features which are novel in this context.

2. DATA ACQUISITION

Adults reporting no knee pain in the last 2 weeks were recruited. Knees were classified by clinicians as: 1) normal (clinically healthy), 2) abnormal (OA). Exclusion criteria were: aged $<$18 years, previous surgery, unable to provide consent. AE signals were acquired with a sampling frequency of ≥ 44.1 kHz and downsampled to 16 kHz for subsequent processing, using a contact microphone (Basik Pro Schertler, 20 Hz – 20 kHz), attached over the patella, during walking on a treadmill instrumented with force plates.

The assessment commenced with a 5 minute warm-up on the treadmill followed by data acquisition at progressive speeds on a flat level until maximum walking speed was achieved (defined as the maximum pain-free speed where one foot was always in contact with the ground). Maximum speeds achieved per subject range from 2.5 to 9 km/h.

Data used in this work originates from 40 knees, of which 19 are normal (15 patients) and 21 are abnormal (18 patients).

3. PRE-PROCESSING AND FEATURE EXTRACTION

We assume that sounds related to abnormality are repeatable and appear within time periods of 20 seconds. We then divide each knee signal into non-overlapping segments and label them according to the knee condition, resulting in 249 normal and 297 abnormal segments. Other time periods that do not violate the aforementioned assumptions could also be used but would alter the number of segments. All recorded signals are normalised to have equal Root Mean Square (RMS) level.

Let $\mathbf{F} = \{f_1, ..., f_N\}$ be the feature set containing (in order) (a) 19 Mel-frequency Cepstral Coefficients (MFCC) and their first and second derivatives, (b) magnitude of Short Time Fourier Transform (STFT) with 257 bins, (c) 257 modulation magnitude spectrum values and (d) 3 pulse waveform parameters totalling $N = 574$ features. MFCCs are successfully used in speech recognition and music genre classification but have never been used, to our knowledge, for knee signals. MFCC and |STFT| features are calculated in 32 ms frames with 50% overlap. Modulation magnitude spectrum is obtained as in [18], using a window of 6 acoustic time frames.

In our study we observed occasional short acoustic pulses in the recorded signals. We define a pulse with the parameters $\{\alpha_0, t_0, r_0\}$ where α_0 is the amplitude threshold used for discarding unwanted variations, t_0 is the minimum duration (16 samples) and r_0 is the rest time ($r_0 = t_0$), a window within which the waveform amplitude is below α_0 and determines the end of the pulse. After 250 Hz high-pass filtering the segment, the variance is computed in 50 ms frames with 50% overlap. We identify frames with variance less than the 10th percentile as noise-like segments and estimate their sample amplitude distribution. We then obtain the inverse CDF and compute α_0 at a probability of $1 - \eta$ where η is a variable that controls the value of α_0. In our experiments we found $\eta = 10^{-14}$ to be a good choice. Changing any of the values of $\{\alpha_0, t_0, r_0\}$ would alter the number of pulses detected. The results in this paper use the above-mentioned values.

Following the above we denote the filtered signal as $s_0(n)$ and find the peaks $\{p_1, ..., p_c\}$ in $|s_0(n)|$ that exceed $|\alpha_0|$. To avoid spurious threshold crossings we discard those that have less than 2 samples above $|\alpha_0|$ in a window of 5 samples centred at the peaks. We denote j and l as the repetition indices and the unknowns $\{b_j^i, v_l^i\}$ as the start and stop pulse samples with $b_0^i = v_0^i = p_i$. To find them we repeat: $b_j^i = \left[\min(b_{j-1}^i - n); n = 0, ..., t_0; j = 1, 2, ...\right]$ and $v_l^i = \left[\max(v_{l-1}^i + n); n = 0, ..., t_0; l = 1, 2, ...\right]$ subject to $|s_0(b_{j-1}^i - n)| > |\alpha_0|$ and $|s_0(v_{l-1}^i + n : v_{l-1}^i + n + 1)| > |\alpha_0|$ respectively, until the inequalities do not hold for any n of the current indices in which case we stop and add r_0 to v_l^i. At the end we merge any overlapping pulses. Fig. 1 shows a snapshot of $s_0(n)$ with the identified pulses. Clearly there can be many and with various waveforms. We therefore extract from each: the peak-to-peak amplitude, duration and energy.

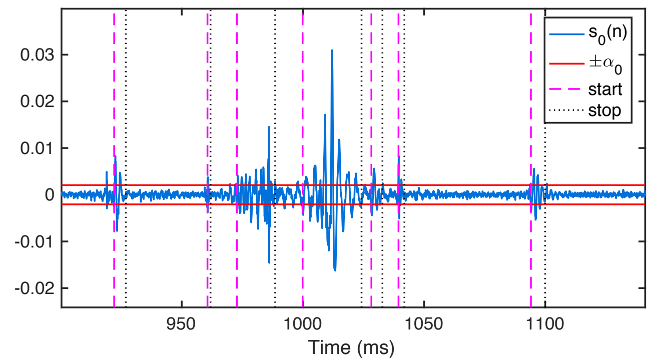

Fig. 1. High-pass filtered signal with the identified pulses

4. FEATURE ANALYSIS

We analyse the discriminant power of each of the 574 features aiming to obtain insights into the nature and the fundamental differences between normal and abnormal knee signals. We first perform a statistical analysis on the separability of the feature distributions that leads to a discriminant analysis.

4.1. Statistical analysis of feature distributions

Let $\mathbf{F}_x = \{x_1^{d_1}, x_2^{d_2}, .., x_N^{d_N}\}$ and $\mathbf{F}_y = \{y_1^{m_1}, y_2^{m_2}, .., y_N^{m_N}\}$ denote the features, as in $\mathbf{F}$, of normal and abnormal segments respectively where (d_i, m_i) are their dimensions. We are interested in finding whether $x_i^{d_i}$ and $y_i^{m_i}$ generate dissimilar sample distributions such that a classifier with low error rate could be designed. Given the short time frames however, it is unlikely that in every such frame the prevalence of OA signatures in OA segments is high and hence, overlap is expected.

Using the two-sample Kolmogorov-Smirnov Test (KST) [19], we test for every $i = 1, ..N$ the null hypothesis H_0 that $x_i^{d_i}$ and $y_i^{m_i}$ originate from the same continuous distributions. Results show that, at 5% significance level, H_0 was rejected for 565 features implying that their statistical differences are significant. To further compare the distributions we use the Bhattacharyya distance [20], given by

$$D_B = -\log_e \left[\int_{-\infty}^{+\infty} \sqrt{p_1(x)p_2(x)} dx \right] \tag{1}$$

where $p_1(x)$ and $p_2(x)$ are the two continuous distributions in question. Assuming that MFCCs follow normal distributions then (1) can be easily simplified [21]. The rest of the features are one-sided, hence we assume that exponential distributions for modulation spectrum features and gamma distributions for |STFT| and pulse features give better approximations. We extend (1) for gamma to obtain

$$-\log_e \left[\frac{\Gamma((\alpha_1+\alpha_2)/2)}{\left[\Gamma(\alpha_1)\Gamma(\alpha_2)(\beta_1^{\alpha_1}\beta_2^{\alpha_2})(\frac{1}{2\beta_1}+\frac{1}{2\beta_2})^{\alpha_1+\alpha_2}\right]^{\frac{1}{2}}} \right] \quad (2)$$

where $\Gamma(.)$ is the gamma function and $\{\alpha_1, \beta_1, \alpha_2, \beta_2\}$ are the parameters of the two distributions. The expression for exponential distributions can be easily obtained from (2).

We estimate the model parameters from the sample populations and compute $D_B(p_1(x_i^{d_i}), p_2(y_i^{m_i}))\ \forall\ i = 1, ..., N$. The results in Fig. 2 support the KST outcome but also indicate that STFT and modulation domain features for mid to high frequencies exhibit the largest separations and so, have higher discriminant power. Low D_B values mean significant overlap but a closer look at the distributions reveals that in spite of this, the shapes and tails can be different. Using the whole distribution may be hindering some class differences. We use the mean, variance, kurtosis, skewness, max, min and the 10th, 25th, 50th, 75th, 90th percentiles to represent each feature distribution and examine the effectiveness of this parameterisation in the subsequent discriminant analysis.

4.2. Discriminant analysis of features

Bayes' rule for minimum error states that a vector $\boldsymbol{y} = \{y_1, ..., y_p\}$ is assigned to class ω_j if $p(\omega_j|\boldsymbol{y}) > p(\omega_k|\boldsymbol{y})$ $\forall\ k = 1, ..., K; k \neq j$ [21]. Using Bayes' theorem we obtain

$$p(\boldsymbol{y}|\omega_j)p(\omega_j) > p(\boldsymbol{y}|\omega_k)p(\omega_k). \quad (3)$$

For Linear Discriminant Analysis (LDA) we assume that the classes follow multivariate normal (MVN) distributions with equal covariances Σ and $p(\boldsymbol{y}|\omega_k)$ given by

$$\frac{1}{\sqrt{(2\pi)^p|\Sigma|}} \exp\Big[-\frac{1}{2}(\boldsymbol{y}-\boldsymbol{\mu_k})^T\Sigma^{-1}(\boldsymbol{y}-\boldsymbol{\mu_k})\Big]. \quad (4)$$

By substituting (4) into (3), taking the log and ignoring constant terms across classes, we obtain the discriminant function

$$h_k(\boldsymbol{y}) = \log(p(\omega_k)) - \frac{1}{2}\boldsymbol{\mu_k}^T\Sigma^{-1}\boldsymbol{\mu_k} + \boldsymbol{y}^T\Sigma^{-1}\boldsymbol{\mu_k}. \quad (5)$$

It has been shown that (5) is robust to deviations from the covariance equality assumption [22]. However, we also perform Quadratic Discriminant Analysis (QDA) where $\Sigma_1 \neq \Sigma_2$ to search for non-linear discriminants. The MVN assumption was tested using the Henze-Zirkler test, [23], on each f_i in **F**, with $p = 11$ in (4), giving p-values <0.05 for all features, rejecting therefore the assumption. However, based on previous findings that LDA performs robustly for certain tasks even when the data is not MVN we employ it in the following [24].

The parameters $\{\Sigma, \boldsymbol{\mu_k}, p(\omega_k)\}$ are calculated from the sample data for each class k over a training set. For LDA, Σ is computed as the unbiased estimate of the pooled within-class covariance matrix. New observations $\boldsymbol{y}$ are assigned to the class for which $h_k(\boldsymbol{y})$ is largest. For QDA the appropriate equation derived from (3) using (4) for different Σ_i is used.

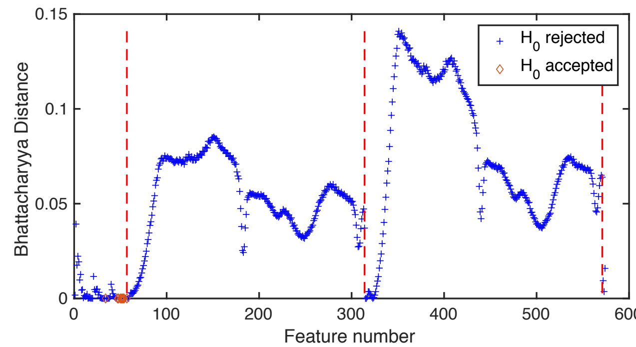

Fig. 2. Bhattacharyya distance for each feature. Dashed lines separate the feature types. Left to right are the MFCC, acoustic magnitude, modulation magnitude and pulse features.

Cross-validation testing was performed 100 times for each f_i in order to reduce the variance of the estimator, averaging the results at the end. Five groups were randomly created each time having normal to abnormal knees ratio of $\frac{3}{5}, \frac{3}{5}, \frac{3}{5}, \frac{5}{3}, \frac{5}{3}$. Each group is then made up with the segments of its constituting knees. Some variability in the group sizes exists as some knees have more segments than others.

For performance assessment we compute the error rate (E_r), $F_{0.5}$ score (a variation of F_1) and Matthew's Correlation Coefficient (MCC). From a clinical perspective, the false prediction of abnormal segments as normal is worse than the contrary. $F_{0.5}$ emphasises this error type more than F_1 and is thus preferred. MCC is regarded as a balanced measure ranging from -1 (prediction totally different from observation), to 1 (perfect prediction), with 0 stating random prediction [25].

Results are reported in Table 1. To aid the comparison we included an average score $S = \big[(1 - E_r) + F_{0.5} + MCC\big]/3$ computed using the average metric values. Due to space lim-

Feature	E_r		$F_{0.5}$		MCC		S
	$\overline{x}$	std	$\overline{x}$	std	$\overline{x}$	std	
MFCC 2 (f_2)	19.4	2.47	78.9	2.77	60.8	4.95	73.4
|STFT| 11 (f_{68})	23.3	2.65	74.1	2.79	53.3	5.35	68.0
|STFT| 10 (f_{67})	24.9	2.56	72.4	2.67	50.1	5.22	65.9
|STFT| 12 (f_{69})	25.3	2.87	72.1	2.97	49.1	5.90	65.3
|STFT| 13 (f_{70})	26.5	2.61	70.7	2.68	46.7	5.41	63.7
|STFT| 9 (f_{66})	26.7	3.27	70.5	3.50	46.3	6.79	63.3

Table 1. Average cross-validation results for the best features. All values are $\times 10^{-2}$ and $\overline{x}$ refers to average metric values.

Classifier	E_r		$F_{0.5}$		MCC		S	$\theta_{0.5}$	θ_{mcc}	Features used
	Mean	Std	Mean	Std	Mean	Std				
LDA	0.157	0.0215	0.817	0.0227	0.688	0.0435	0.783	0.70	0.10	$\{f_2, f_{66} - f_{70}\}$
SVM (linear)	0.139	0.0189	0.841	0.0213	0.722	0.0378	0.808	0.70	0.45	$\{f_2, f_{66} - f_{70}\}$
SVM (gaussian)	0.247	0.0309	0.719	0.0316	0.515	0.0607	0.662	0.65	0.55	$\{f_2\}$
CART	0.178	0.0244	0.796	0.0279	0.646	0.0478	0.755	0.75	0.25	$\{f_2\}$

Table 2. Results show the best performance per classifier.

itations we report only the top 6 features and provide a plot in Fig. 3 that shows the performance of all, obtained as the maximum S score of both methods. Non-linear boundaries are more suitable for 517 features since QDA performed better. Evidently, the most discriminant features are predominantly at low frequencies coming from the cepstral and STFT domains. In general, the performance of MFCC varies from very poor to the best and there is no evidence of increased discrimination in their first and second derivatives. Using the proposed set of statistics has improved the discriminability of some features for which the KST H_0 was accepted. STFT and modulation features have similar structure in S that somewhat resembles Fig 2. It peaks after a few bins and then follows a negative trend. Differences at high frequencies are small, resulting in poorer scores. Pulse duration scores the highest amongst the pulse features but moderately overall.

Given that the MVN assumption is not valid, the approximation to (3) is not optimal. Nevertheless, we showed that linear and non-linear hyperplanes separating the classes exist and are captured by LDA and QDA with good performance.

5. EXPERIMENTS

We experiment with several classifiers using different subsets of **F** that are constructed by setting thresholds on the performance parameters, denoted as $[\theta_{er}, \theta_{0.5}, \theta_{mcc}]$. We can either choose suitable values or search over a range and compare the results. Intuitively, θ_{er} can be set equal to the error rate we obtain when we always predict the smallest class, that is 0.456. By keeping θ_{er} constant and varying $\theta_{0.5}$ and θ_{mcc} in the range $[0, 0.05, ..., 1]$ we construct all possible subsets. Cross validation procedure was employed per classifier as before. Prior to this, we scale the training data by subtracting the mean and normalising the variance. The same scale values are applied to the test set. In Table 2 we report the best results (according to S) for each classifier.

Comparing with Section 4 results, classification performance is improved except for the case of Support Vector Machine (SVM) with gaussian kernel. On average, SVM with linear kernel performs better and with less variability, achieving an S score of 0.808. If we use these results and classify each knee using majority vote on the segments, we obtain S=0.777 with average $[\mathrm{E}_r, \mathrm{F}_{0.5}, \mathrm{MCC}]$ equal to

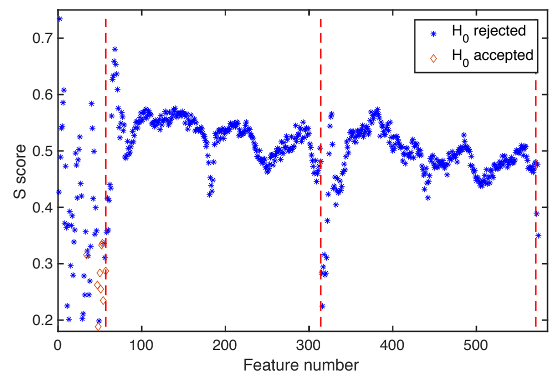

Fig. 3. Performance of all features in **F**. H_0 refers to the KST.

$[0.175, 0.849, 0.658]$, given again by linear SVM.

We notice that the best results were achieved with just a few features and in two cases with only one. Even Classification and Regression Tree (CART), which uses an inherent feature selection method tied to the classifier model, performs best with a single feature. The above are possibly attributed to the small dataset given that one feature adds 11 variables, the feature space eventually becomes sparse. With too many variables the classifier is likely to overfit the training data and fail to generalise to new data resulting in poor performance.

6. CONCLUSION

In this paper we addressed the discriminant analysis of various features for the 2-class classification of knee sounds during walking. We statistically evaluated the feature distributions and motivated the use of 11 summary statistics to describe them. With LDA and QDA we found that MFCC and STFT features at low frequencies are the most discriminant. Selecting only 6 of these by thresholding and using a linear kernel SVM, we obtained $[\mathrm{E}_r, \mathrm{F}_{0.5}, \mathrm{MCC}] = [0.139, 0.841, 0.722]$. From our results we can deduce that knee sounds obtained during walking, with a contact microphone attached over the patella, show a strong indication that they could be used as a non-invasive and a cost effective method to discriminate between normal and abnormal knees, where abnormality in our case was clinically defined as OA.

7. REFERENCES

[1] R. P. Welsh, "Knee joint structure and function," *Clinical orthopaedics and related research*, vol. 147, pp. 7–14, 1980.

[2] N. Lane and D. J. Walace, *All about osteoathritis: the definitive resource for arthritis patients and their families*. Oxford University Press, 2002.

[3] M. L. Chu, I. A. Gradisar, M. R. Railey, and G. F. Bowling, "An electroacoustical technique for the detection of knee joint noise," *Med. Res. Eng.*, vol. 12, no. 1, pp. 18–20, 1976.

[4] M. L. Chu, I. A. Gradisar, L. Zavodney, and G. F. Bowling, "Computer aided acoustical correlation of papathologic cartilage generated noise," in *30th Ann. Conf. Eng. Med. Biol.*, 1977, p. 175.

[5] M. L. Chu, I. A. Gradisar, and R. Mostardi, "A noninvasive electroacoustical evaluation technique of cartilage damage in pathological knee joints," *Med. Biol. Eng. Comput.*, vol. 16, pp. 437–442, 1978.

[6] R. M. Rangayyan, S. Krishnan, , G. D. Bell, C. B. Frank, and K. O. Ladly, "Analysis of vibroarthrographic signals with features related to signal variability and radial-basis functions," *IEEE Trans. Biomed. Eng.*, vol. 44, no. 11, pp. 1068–1074, Nov. 1997.

[7] S. Krishnan, R. M. Rangayyan, G. D. Bell, C. B. Frank, and K. O. Ladly, "Adaptive filtering, modelling and classification of knee joint vibroarthrographic signals for non-invasive diagnosis of articular cartilage pathology," *Med. Biol. Eng. Comput.*, vol. 35, no. 6, pp. 677–684, Nov. 1997.

[8] K. S. Kim, J. H. Seo, J. U. Kang, and C. G. Song, "An enhanced algorithm for knee joint sound classification using feature extraction based on time-frequency analysis," *Comput Methods Programs Biomed*, vol. 94, pp. 198–206, 2009.

[9] Y. Wu and S. Krishnan, "An adaptive classifier fusion method for analysis of knee-joint vibroarthrographic signals," in *International Conference on Computational Intelligence for Measurement Systems and Applications*, 2009, pp. 190–193.

[10] R. M. Rangayyan and Y. F. Wu, "Screening of knee-joint vibroarthrographic signals using statistical parameters and radial basis functions," *Med Biol Eng Comput*, vol. 46, no. 3, pp. 223–32, Mar. 2008.

[11] K. S. Kim, S. O. Lee, J. H. Seo, and C. G. Song, "Classification of joint pathology using an acoustical analysis of knee joint sound," in *Biomedical Circuits and Systems Conference (BioCAS)*, Nov. 2006.

[12] K. S. Kim, J. H. Seo, and C. G. Song, "An acoustical evaluation of knee sound for non-invasive screening and early detection of articular pathology," *Journal of Medical Systems*, vol. 36, pp. 715–722, 2012.

[13] R. M. Rangayyan and Y. Wu, "Analysis of vibroarthrographic signals with features related to signal variability and radial-basis functions," *Ann Biomed Eng*, vol. 37, pp. 156–163, Jan. 2009.

[14] Y. Wu, S. Cai, S. Yang, F. Zheng, and N. Xiang, "Classification of knee joint vibration signals using bivariate feature distribution estimation and maximal posterior probability decision criterion," *Entropy*, vol. 15, pp. 1375–1387, 2013.

[15] B. Mascaro, J. Prior, L.-K. Shark, J. Selfe, P. Cole, and J. Goodacre, "Exploratory study of a non-invasive method based on acoustic emission for assessing the dynamic integrity of knee joints," *Med. Eng. Phys.*, vol. 31, pp. 1013–1022, 2009.

[16] L.-K. Shark, J. Goodacre, and H. Chen, "Knee acoustic emission: A potential biomarker for quantitative assessment of joint ageing and degeneration," *Med. Eng. Phys.*, vol. 33, pp. 534–545, 2011.

[17] H. Chen, "Discovery of acoustic emission based biomarker for quantitative assessment of knee joint ageing and degeneration," Ph.D. dissertation, University of Central Lancashire, 2011.

[18] K. Paliwal, K. Wojcicki, and B. Schwerin, "Single-channel speech enhancement using spectral subtraction in the short-time modulation domain," *Speech Communication*, vol. 52, no. 5, pp. 450–475, 2010.

[19] F. J. Massey, "The kolmogorov-smirnov test for goodness of fit," *Journal of the American Statistical Association*, vol. 46, no. 253, pp. 68–78, 1951.

[20] A. Bhattacharyya, "On a measure of divergence between two statistical populations defined by their probability distributions," *Bull. Calcutta Math. Soc.*, vol. 35, p. 99109, 1943.

[21] S. Theodoridis and K. Koutroumbas, *Pattern Recognition*. Elsevier, 2009.

[22] T. J. O'Neil, "Error rates of non-bayes classification rules and the robustness of fisher's linear discriminant function," *Biometrika*, vol. 79, no. 1, pp. 177–184, 1992.

[23] N. Henze and B. Zirkler, "A class of invariant consistent tests for multivariate normality," *Communications in Statistics - Theory and Methods*, vol. 19, no. 10, pp. 3595–3617, 1990.

[24] R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern Classification*, 2nd ed. John Wiley and Sons, 2001.

[25] P. Baldi, S. Brunak, Y. Chauvin, C. A. F. Andersen, and H. Nielsen, "Assessing the accuracy of prediction algorithms for classification: an overview," *Bioinformatics*, vol. 16, no. 5, pp. 412–424, 2000.