Acoustic Emissions From Loaded and Unloaded Knees to Assess Joint Health in Patients With Juvenile Idiopathic Arthritis

Sevda Gharehbaghi, *Graduate Student Member, IEEE*, Daniel C. Whittingslow, *Graduate Student Member, IEEE*, Lori A. Ponder, Sampath Prahalad, and Omer T. Inan, *Senior Member, IEEE*

***Abstract*—Objective: We studied and compared joint acoustical emissions (JAEs) in loaded and unloaded knees as digital biomarkers for evaluating knee health status during the course of treatment in patients with juvenile idiopathic arthritis (JIA). Methods: JAEs were recorded from 38 participants, performing 10 repetitions of unloaded flexion/extension (FE) and loaded squat exercises. A novel algorithm was developed to detect and exclude rubbing noise and loose microphone artifacts from the signals, and then 72 features were extracted. These features were down-selected based on different criteria to train three logistic regression classifiers. The classifiers were trained with healthy and pre-treatment data and were used to predict the knee health scores of post-treatment data for the same patients with JIA who had a follow-up recording. This knee health score represents the probability of having JIA in a subject (0 for healthy and 1 for arthritis). Results: Post-treatment knee health scores were lower than pre-treatment scores, agreeing with the clinical records of successful treatment. Regarding loaded versus unloaded knee scores, the squats achieved a higher score on average compared to FEs. Conclusion: In healthy subjects with smooth cartilage, the knee scores of squats and FEs were similar indicating that vibrations from the friction of articulating surfaces do not significantly change by the joint load. However, in subjects with JIA, the scores of squats were higher than the scores of FEs, revealing that these two exercises contain different, possibly clinically relevant, information that could be used to further improve this novel assessment modality in JIA.**

***Index Terms*—Wearable technologies, knee joint health, joint acoustic emissions, vibroarthrography, supervised learning, feature selection.**

I. Introduction

ARTHRITIS is the inflammation of a joint often associated with symptoms of swelling, heat, pain, and stiffness [1]. The most common form of childhood arthritis is juvenile idiopathic arthritis (JIA), which refers to all forms of arthritis that appear before 16 years of age and are of unknown origin [2]. The precise etiology of JIA is poorly understood; however, research studies indicate that it is an auto-immune disease with multiple genetic and environmental risk factors involved [1]. It has a prevalence of up to 150 cases per 100 000 children in North America [3]. JIA becomes a chronic condition in about half of the cases, and continues afflicting the patient for several years or even a lifetime [2], [4]. This type of arthritis has a heterogeneous presentation and few reliable biomarkers which makes diagnosis, and quantifying treatment efficacy, difficult [3], [5]. Furthermore, there is a limited access to pediatric rheumatologists, who are specially trained for diagnosing and treating JIA, where only 1 in 4 children with JIA are able to regularly see a pediatric rheumatologist in the U.S. [6], [7].

JIA is classified based on the number of affected joints, clinical and laboratory features as well as family history [1]. The knee is the most commonly affected joint, which is a hinge type synovial joint protected by articular cartilage and lubricated with synovial fluid [8]. The cartilaginous surfaces of a normal knee are smooth and slippery [9], whereas in an arthritic knee, the synovial membrane surrounding the joint becomes inflamed, the smooth cartilage degenerates, and—if left untreated—bony erosions and density loss may occur [8]–[10]. Therefore, early diagnosis with effective treatment is necessary to prevent long-term effects [11].

During healthy joint movement, the inter-joint articular friction produces vibrations or sounds, which are referred to as vibroarthrographic (VAG) signals or joint acoustic emissions (JAEs) [9], [12]. Vibrations generated by the articulating surfaces of degenerated cartilage are expected to be different from the JAEs of healthy cartilage [8]. JAEs carry important information about the joint health, and these signals can be

Manuscript received August 28, 2020; revised February 7, 2021; accepted May 8, 2021. Date of publication May 18, 2021; date of current version September 3, 2021. This research was supported in part by National Science Foundation under Grant Number 1749677. The work of Dr Sampath Prahalad was supported in part by The Marcus Foundation Inc., Atlanta, GA, USA. *(Corresponding author: Sevda Gharehbaghi.)*

Sevda Gharehbaghi and Omer T. Inan are with the School of Electrical and Computer Engineering, Georgia Institute of Technology, Atlanta, GA 30332 USA (e-mail: sevda@gatech.edu; omer.inan@ece.gatech.edu).

Daniel C. Whittingslow is with the Coulter Department of Biomedical Engineering, Georgia Institute of Technology, Atlanta, GA 30332 USA (e-mail: d.c.whittingslow@emory.edu).

Lori A. Ponder and Sampath Prahalad are with the School of Medicine of Emory University, Atlanta, GA 30322 USA, and also with the Children's Healthcare of Atlanta Inc., Atlanta, GA 30322 USA (e-mail: laponde@emory.edu; sprahal@emory.edu).

Digital Object Identifier 10.1109/JBHI.2021.3081429

measured non-invasively and used in the classification of normal versus abnormal joints [13]–[17]. Previous studies demonstrated that JAEs could be used as a digital biomarker for JIA diagnosis through evaluating knee health during the course of treatment [18]–[20]. However, in all prior studies, the analysis was limited to JAEs generated strictly with unloaded flexion/extension exercises (FE).

We hypothesized that, while JAEs recorded during unloaded FE tasks for patients with JIA post-treatment closely matched healthy knees, JAEs recorded during loaded more complex multi-joint weighted movements involving knee and hip flexion / extension (i.e., squats) would still be significantly different from healthy controls. The rationale for this hypothesis was that loaded movements would result in greater joint contact forces [16], [21]–[24], and thus the frictional interaction of articulating surfaces within the knee would be increased; any roughness in surfaces would thus result in different acoustic characteristics during the movement. Accordingly, in this paper, we analyze for the first time JAEs extracted during squats from patients with JIA compared to healthy controls, and for a sub-set of the same patients with JIA following several months of treatment. In addition to addressing our scientific hypothesis, we believe that there is practical value in demonstrating that JAEs measured during squats can differentiate JIA from healthy, and pre-treatment JIA from post-treatment: squats are commonly used in clinical settings to study movement since they can be performed with minimal equipment, and they include a sit-to-stand component with demonstrated clinical value [14], [21], [25], [26]. If JAEs derived from squats hold merit for assessing knee health in JIA, then JAEs could ultimately be extracted during routinely performed sit-to-stand exercises from patients at home and during everyday settings with a wearable smart brace [27]. To further increase this potential for translating the approaches described here to the home, we developed and validated a novel method for identifying and removing the signal artifacts to improve the robustness of this sensing technique.

II. Materials and Methods

A. Human Subject Protocol and Subject Demographics

This work builds upon our prior studies [20], [28], which were approved by the Georgia Institute of Technology and the Emory University Institutional Review Boards (#00081670). In this work, knee JAEs were acquired from 38 study participants including 20 subjects who were diagnosed with JIA by a pediatric rheumatologist and 18 healthy controls with no history of JIA or acute knee injuries. All subjects had BMI in the normal range and were able to ambulate without assistance. The group with JIA consists of 17 females (13.2 $\pm$ 2.1 years old, BMI 20.4 $\pm$ 4.2 kg/m^2) and 3 males (10.7 $\pm$ 3.8 years old, BMI 17.4 $\pm$ 2.0 kg/m^2), and the healthy control group consists of 15 females (12.4 $\pm$ 3 years old, BMI 20.8 $\pm$ 2.8 kg/m^2) and 3 males (13 $\pm$ 4.6 years old, BMI 17.5 $\pm$ 1.9 kg/m^2). To measure longitudinal changes in the knee JAEs during the course of treatment, 10 subjects with JIA had a follow-up recording, 3–6 months after initial measurements. The demographics and physical characteristics of the participants are presented in Table I,

TABLE I
Demographic Data for Study Participants

	JIA	Healthy
# Subjects	20	18
# Females (% of group)	17 (85%)	15(83%)
# Males (% of group)	3 (15%)	3 (17%)
Age (mean $\pm$ σ, in $years$)	12.9 $\pm$ 2.5	12.5 $\pm$ 3.2
Weight (mean $\pm$ σ, in kg)	49.0 $\pm$ 12.6	51.1 $\pm$ 12.3
Height (mean $\pm$ σ, in cm)	155.9 $\pm$ 10.7	158.5 $\pm$ 16.9
BMI (mean $\pm$ σ, in kg/m^2)	20.0 $\pm$ 4.0	20.6 $\pm$ 2.7

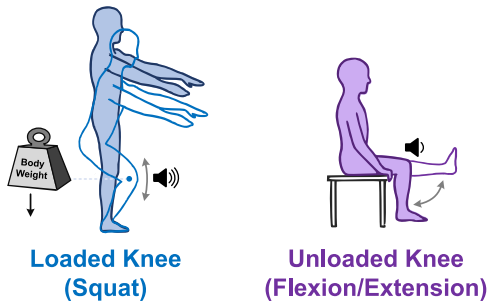

Fig. 1. Loaded and unloaded exercises to excite the knee acoustic emissions.

which shows a fairly balanced dataset among the two groups of healthy controls and patients with JIA. Note that JIA is more common in females [29] and the demographics of this study correspond with this distribution.

In this protocol, subjects were asked to perform ten repetitions of two exercises shown in Fig. 1: loaded squats while bearing the body weight, and unloaded FEs while seated on a height-adjustable stool without foot contact with the ground. Subjects performed each exercise by following an instructional cartoon that encouraged a movement cycle to be completed every four seconds through the full range of motion (RoM). The JAEs from each knee were recorded by two uniaxial accelerometers (Series 3225F7, Dytran Instruments Inc., CA, USA) which were attached 2 cm medial and lateral to the distal patellar tendon of each knee using double-sided adhesive pads (Rycote Microphone Windshields Ltd, Stroud, Gloucestershire, GL5-1RN, U.K.). These accelerometers, acting as contact microphones, have a wide bandwidth of 2 Hz–10 kHz and a high sensitivity of 100 mV/g. An inertial measurement unit (IMU) (BNO055, Adafruit Industries, NY, USA) was also attached around the subjects' ankles to record the joint motion while subjects were performing the exercises. Knee joint vibrations were sampled at 100 kHz via a data acquisition system (USB-4432, National Instruments, TX, USA) and stored on a computer for further signal processing in MATLAB (MATLAB, MathWorks, MA, USA).

B. Signal Processing and Feature Extraction

The signal processing pipeline of knee sounds in this work is depicted in Fig. 2, where, following pre-processing, the recorded signals from the contact microphones were divided

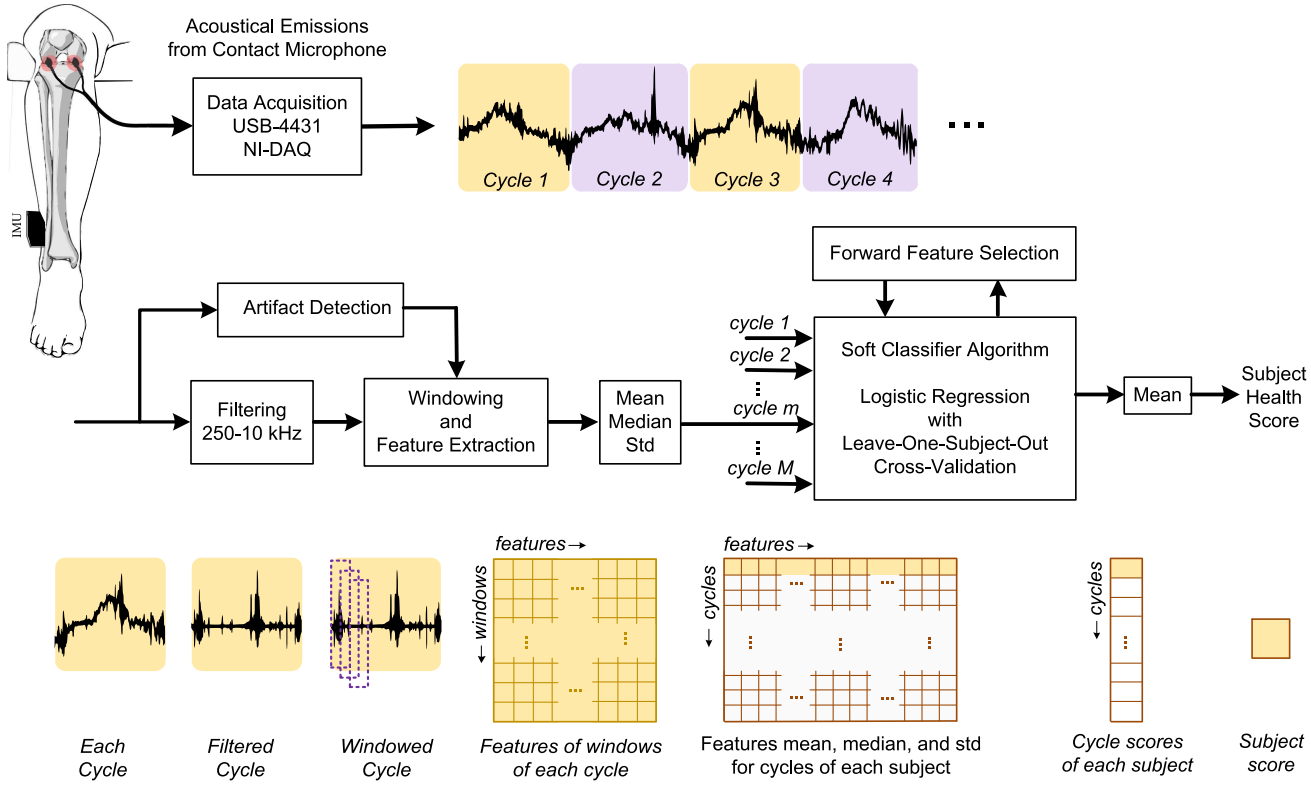

Fig. 2. Overall block diagram and system pipeline: The recorded JAEs were filtered, windowed, and features were extracted for all the cycle frames. An artifact detection algorithm also processed each cycle to identify the affected frames by artifacts and later exclude those frames. Then the mean, median, and standard deviation of each feature were stored in a larger feature matrix with all the cycles of a subject. After the feature matrix was calculated for all the subjects, a soft classifier was trained and evaluated with a leave-one-subject-out cross-validation algorithm. A forward feature selection block down-selected the features and modified the classifier. The output of the classifier is the predicted cycle scores, which are then averaged over each subject to get a single subject health score. The score is the probability of having JIA (0 for healthy and 1 for JIA).

into movement cycles (∼4 seconds each) based on the IMU data. The knee sounds contain high-energy and short-duration acoustic signals with "spike-like" waveforms, having a broad frequency spectrum that are mostly limited to 10 kHz [15], [18]. Therefore, the pre-processing of these joint sounds comprised digital filtering with a Kaiser-window bandpass filter (250 Hz–10 kHz) to reduce the unwanted noise and interference. Then, each cycle was divided into 200 ms frames with 50% overlap, where each frame was long enough to contain multiple JAE signatures. This window size was heuristically chosen to provide several frames per movement cycle while maintaining the low frequency content of the signal [20]. In addition, a novel automated algorithm was developed to detect the artifacts and rubbing noise associated with loose microphone contacts [17], and exclude them from the analysis to increase the reliability and accuracy of results.

Table II shows the list of 72 features that were extracted from each frame. The first category is temporal features which consists of the signal energy, zero crossing rate (ZCR), RMS amplitude, and energy entropy (f_1–f_4). The second category is spectral features, which includes the spectral centroid, spectral spread, spectral entropy, spectral flux, spectral roll-off, spectral crest, spectral decrease, spectral flatness, spectral kurtosis, spectral slope, spectral skewness, harmonic ratio, fundamental frequency, mean frequency, 12 chroma features, and 29 bandpower features, which are the signal power in 29 distinct log-scaled frequency range between 250 Hz–10 kHz (f_5–f_{59}). The third category is Mel-frequency cepstral coefficients (MFCCs) (f_{60}–f_{72}), which define the overall shape of the signal spectral envelope and are widely used in speech recognition and music information retrieval. A detailed description of these features can be found in [15], [16], [20], [30].

TABLE II
A LIST OF ALL EXTRACTED FEATURES

Feature Sets	Features Name
Temporal	Energy, Zero-Crossing Rate, RMS Amplitude, Entropy of Energy
Spectral	Spectral Centroid, Spectral Spread, Spectral -Entropy, Spectral Flux, Spectral Roll-Off, Spectral Crest, Spectral Decrease, Spectral -Flatness, Spectral Kurtosis, Spectral Slope, Spectral Skewness, Fundamental Frequency, Harmonic Ratio, Mean Frequency, #12 Chromas, #29 Band powers
MFCC	#13 Mel-Frequency Cepstrum Coefficients

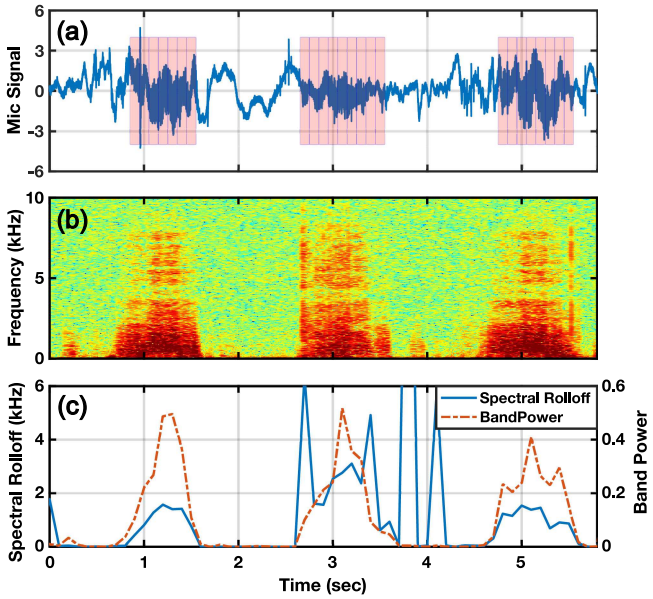

Fig. 3. a) waveforms of knee JAEs and detected windows with rubbing artifacts, b) the spectrogram of the signal, in which the high power levels of the artifacts saturated the color-bar rendering normal JAEs barely visible, and c) the two features, bandpower and spectral roll-off, used to detect the artifacts.

To summarize, each movement cycle (either FE or squat) was about 4 seconds, containing approximately 40 frames, and each frame was processed to extract 72 features. Then, the mean, median, and standard deviation of each feature for all frames of a movement cycle were calculated and stored in the feature matrix, providing a total of 216 (72×3) features for each movement cycle.

Later on, these features were down selected, and the classifier was trained to predict the knee scores of healthy controls and subjects with JIA. In this dataset, the number of available movement cycles is $\sim$2700 since there are 2 datasets $\times$ 38 subjects $\times$ 4 microphones $\times$ 8 to 10 movement cycles per recording. For predicting the knee scores, the classifier was trained with the cycles of all the held-in subjects (2 datasets $\times$ $\sim$36 cycles $\times$ 37 subjects) and the trained classifier was tested on the cycles of the held-out subject (2 datasets $\times$ $\sim$ 36 cycles $\times$ 1 subject).

C. Artifact Detection

Removing noise and artifacts is an essential step towards improving the accuracy and robustness of JAE recordings. A previous study showed that loose microphone contact can introduce additional noise and artifacts to the recorded signals [17]. In some measurements, we noticed that microphone contacts were loose for parts of the recording, and the recorded signal had similar patterns to those of [17]. Therefore, we developed an automated algorithm to identify and exclude these noisy regions affected by the artifacts. Applying this algorithm on all knee sounds increased the robustness of our analysis against such artifacts that are likely to corrupt measurements taken in daily life with wearable devices.

Fig. 3(a) shows the waveforms of knee JAEs with highlighted rubbing artifacts, and Fig. 3(b) presents the spectrogram of the signal which contains strong high-frequency components during the occurrence of these artifacts. We analyzed the spectral and bandpower features and found the following two features were effective in artifact detection when considered together: 1) band power in the range of 0.3–10 kHz to measure the signal power, and 2) spectral roll-off to find the frequency below which 90% of the signal energy is concentrated. Fig. 3(c) shows these two features for the same recording. The rubbing artifact detection algorithm designates a signal frame as "artifact" when both the bandpower and spectral roll-off values were above certain thresholds. The threshold values were selected heuristically based on the signal acquisition system. Of importance, this method of artifact detection is not necessarily the only applicable technique, but other audio features (e.g. MFCCs and spectral entropy) were also able to detect these high-power and high-frequency frames, as those also quantify the spectral power distribution.

The artifact detection algorithm was used for the unfiltered data, which was divided into 200 ms frames with 50% overlap (similar to the windows from feature extraction). Then, the features of the frames contaminated with artifacts were removed from the feature matrix before calculating the mean, median, and standard deviation of each feature for frames of a movement cycle. This increased the reliability of our analysis by relying on the features of denoised frames.

D. Forward Feature Selection

Down selecting the number of features is an important step in reducing model complexity, the computational load, and the possibility of overfitting [31]. Feature-selection is an iterative process performed on a particular classifier and dataset to improve one or several of the classifier parameters, such as validation accuracy and area under the curve (AUC). Forward feature selection (FFS) and backward feature selection (BFS) are two of the most commonly used algorithms in practice, where FFS begins with an empty set of features and in each iteration adds the feature that best improves the desired metric [32]. On the other hand, BFS begins with a model with all features, and it removes the feature without which the model has the highest performance [33]. In this work, FFS was used as it is computationally less expensive and, during this process, the classifier performance was evaluated with leave-one-subject-out cross-validation (LOSO-CV). In the FFS process, the validation accuracy generally tends to increase until it reaches the optimum set of features, and then, the performance starts to drop as more features were added to the model which indicates overfitting. Therefore, feature selection not only helps with reducing the number of features in the model while maximizing the performance, but also decreases the chance of overfitting which makes the model more generalizable.

In this work, the knee sound features for squat and FE exercises could be combined or analyzed separately. The goal of this analysis was to compare JAEs of the squat exercise with those of the FE exercise. Thus, to be consistent with squat and

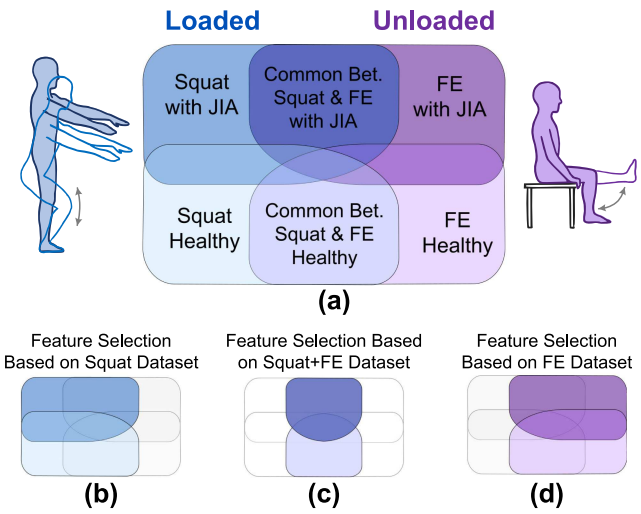

Fig. 4. Feature selection criteria. a) conceptual diagram illustrating the similarities and differences of JAEs of squat versus FE for healthy and JIA groups. Feature selection based on only squat dataset (b), combined squat and FE datasets (c), and only FE dataset (d).

FE knee scores, a single classifier had to predict the scores for both exercise recordings, which means that the training and testing datasets should include both squat and FE knee sounds to make a fair comparison. However, the testing accuracy can be calculated separately for each movement. Nevertheless, the goal of feature down-selection was to find the features that maximized the difference between healthy controls and subjects with JIA.

Fig. 4(a) shows a conceptual diagram indicating that squat and FE knee sounds can have similar or different characteristics and feature values. Thus, if the features were down-selected only based on the squat dataset or the FE dataset, those two feature-sets would not necessarily be the same. We studied this issue more carefully, where the feature down-selection was performed based on 1) only the squat dataset, 2) only the FE dataset, or 3) a combined dataset of both squat and FE recordings. To elaborate more on this, feature selection based on only squat or FE data maximized the difference between the scores of healthy controls and subjects with JIA of that specific dataset. Whereas, in a combined dataset, the classifier tried to assign a score of 0 for both healthy groups (FE and squat) and a score of 1 for both groups with JIA, resulting in similar scores for both exercises. Analyzing these three different feature-sets would pronounce the important features for squat dataset, FE dataset, or the common features between the two [see Figs. 4(b–d)].

E. Classifier Training and Cross-Validation

A supervised learning method was used to assign a health score to knee JAEs based on the down-selected features. The feature matrix has M rows of movement cycles and N columns of features. These features were standardized to have zero mean and unity standard-deviation for each feature (column), and then they were imported to a soft classifier with corresponding knee sound labels (0 for healthy and 1 for arthritis). A binary logistic regression classifier was trained with these features which has a mathematical model formulated as

$$y = \frac{e^{\beta_0+\beta_1 x_1+\cdots+\beta_N x_N}}{1+e^{\beta_0+\beta_1 x_1+\cdots+\beta_N x_N}}, \tag{1}$$

where $x_1, x_2, \ldots, x_N$ are the selected features, $\beta_0, \beta_1, \ldots, \beta_N$ are the classifier coefficients, and y is the classifier output [34]. This logistic regression classifier converts the knee sound features to a probability score between 0 and 1 using the logistic sigmoid function in a way that it can best fit a relationship between the labels and the given features [34]–[36]. Accordingly, the classifier assigned a knee health score for each movement cycle, which was the estimated probability that a given knee sound belonged to an involved knee with JIA. Thus, a score of zero indicated a healthy subject and a score of one was a subject with JIA.

As a result, the calculated knee health score was expected to be higher for subjects with JIA compared to healthy controls, and a threshold was required to classify the predicted probabilities into the two classes [35]. The threshold was set to a default score of 0.5 as a starting point and then adjusted through an optimization process to find the optimum threshold for the best performance (e.g., highest validation accuracy). Thus, any knee sound with a score of less than the threshold was considered a healthy case and a knee sound with a score of greater than the threshold was considered as a case with JIA. In practice, the knee health scores of all movement cycles ($\sim$ 36 cycles) for a subject were averaged first, and then the subject was classified to minimize the effects of noise.

To evaluate the classifier performance, accuracy is a reliable metric when the dataset is fairly balanced, and the problem requires binary classification. In this work, the accuracy was measured by LOSO-CV, in which the classifier was trained with the data of all subjects except one, and then the score for the movement cycles of that held-out subject were predicted and compared with its ground truth labels. This process was repeated for all subjects, and the overall validation accuracy was calculated for the classifier. The calculated labels of the movement cycles were compared with the ground truth to compute the cycle-wise validation accuracy. In addition, the predicted scores of a subject's movement cycles were averaged and compared with the ground truth to calculate a subject-wise validation accuracy, which usually has a higher value than the associated cycle-wise accuracy and is more robust against noise.

There were a few important details in the classifier cross-validation process that were considered: 1) When standardizing the features before each training step, the mean and standard-deviation were calculated based on only the training dataset (data from held-in subjects excluding the testing subject) to make sure the testing data has no effect on the classifier training, then the testing data was standardized with the calculated training mean and standard-deviation values. 2) The follow-up recordings of the subjects with JIA were *excluded* in the training process to ensure that these follow-up data were solely used for testing the algorithm. Our hypothesis was that the predicted scores of the follow-up recordings should decrease with a successful treatment.

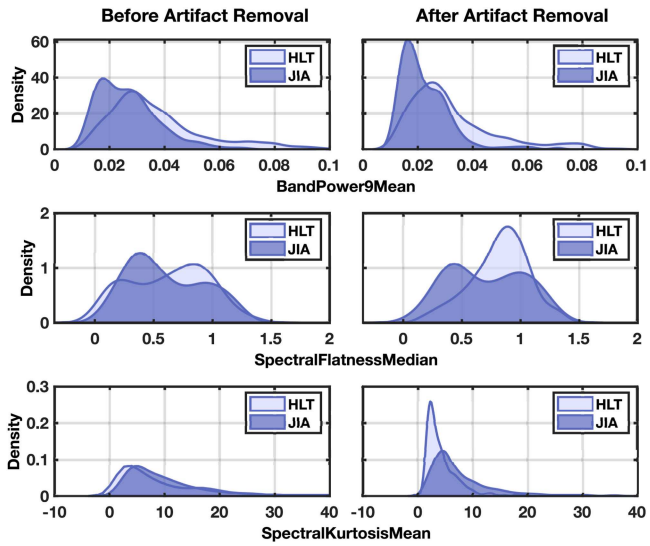

Fig. 5. Kernel density distributions of three sample features before and after artifact removal: bandpower 9 mean (0.69–0.78 kHz), spectral flatness median, and spectral kurtosis mean. The distinction between healthy and JIA group is improved after artifact removal.

F. Feature Importance Ranking

A logistic regression classifier was trained based on the knee sound features to predict the probability of JIA, where the classifier was modeled by a logistic sigmoid function. The coefficients of this classifier, denoted as $\beta_0, \beta_1, \ldots, \beta_N$ in (1), were estimated using the Maximum Likelihood Method [34], [36], and the coefficients explain the size and direction of the relationship between each feature variable and the predicted score. Since the feature variables were standardized to have zero mean and unity standard deviation, the most important features were the ones with the largest coefficient magnitudes. Thus, importance of the features can be found by sorting the absolute value of the coefficients in a descending order, and the most important features were the ones with the highest absolute values. Note that formal model validation with testing data was not needed in this case, as our goal was not to generalize the model, and the relative importance of features is subject to change with the subset of feature variables included in the classifier.

III. Results and Discussion

A. Artifact Detection

In Section II-C, We discussed a novel algorithm to detect and exclude the high-power and high-frequency artifact frames using the two features of bandpower in the range of 0.3–10 kHz and spectral roll-off. Removing the artifacts reduced the corruption due to rubbing noise and improved signal quality. A median of 6–8% of the cycle duration was removed by the artifact detection algorithm and that was similar between the FE and Squat datasets. On a small portion of the recorded cycles (4% of the datasets), more than 50% of the cycle duration was detected to be affected by the artifacts and in those extreme contaminated cases, we removed the whole cycle.

TABLE III
The Performance of Classifiers

	FFS Based on Squat	FFS Based on Squat + FE	FFS Based on FE
Sbj. AUC (Squat)	0.91	0.92	0.93
Sbj. AUC (FE)	0.93	0.89	0.93
Cyc. AUC (Squat)	0.87	0.89	0.86
Cyc. AUC (FE)	0.83	0.84	0.83
Sbj. Val. Acc. (Squat)	88.2%	92.2%	88.2%
Sbj. Val. Acc. (FE)	89.5%	92.1%	89.5%
Cyc. Val. Acc. (Squat)	80.1%	84.3%	80.0%
Cyc. Val. Acc. (FE)	76.5%	82.3%	78.0%
Cyc. Train Acc.	83.5%	86.7%	83.6%
Opt. Threshold (Squat)	0.57	0.75	0.65
Opt. Threshold (FE)	0.45	0.61	0.46

Sbj. stands for subject-wise, Cyc. stands for cycle-wise, Acc. stands for accuracy, Val. stands for validation, and Opt. stands for optimum.

Fig. 5 illustrates the kernel density distributions before and after artifact removal for features such as bandpower 9 mean (0.69–0.78 kHz), spectral flatness median, and spectral kurtosis mean. In these plots, the light and dark density plots correspond to healthy controls and subjects with JIA, respectively, where the differences between the joint sounds of the two groups were heightened after artifact removal. Other features, such as MFCC1 and spectral entropy, were also able to detect the contaminated frames by the artifacts, as those were also quantifying the spectral power distribution and signal energy. Excluding these contaminated frames from data before calculating the mean, median, and standard deviation of frame features for each movement cycle, increased the robustness to the measurement condition and environmental factors. Thus, the classifier was trained based on the features of uncorrupted frames, which led to 8% validation accuracy improvement in the classifier performance.

B. Forward Feature Selection

FFS algorithm was run on all the three mentioned cases with the goal of maximizing classifier AUC, which provides an aggregate measure of performance across all possible classification thresholds. By comparing the classifier parameters as a function of number of features, these trends were observed: When the classifier was trained with only a few features (5–10), the AUC and validation accuracy were still low, but both tended to increase as the number of features increased. With a moderate number of features (20–50), the AUC and validation accuracy improved and saturated, and both eventually dropped with introducing more features, which was an indication of the model overfitting. After comparing these results, the number of features was reduced from 216 to 44 for all three cases, where the classifiers had a relatively high performance.

Table III summarizes the performance of classifiers, where the subject-wise AUC values were between 0.89–0.93, and subject-wise validation accuracy values were in 88–92% range. As mentioned earlier, the subject-wise performance was expected

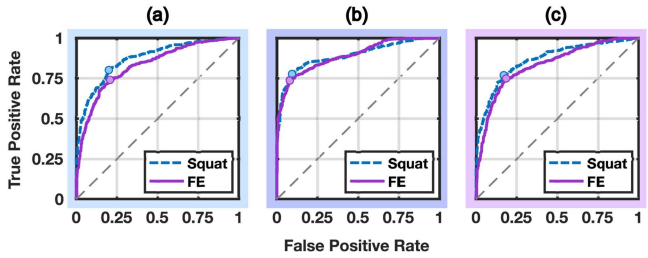

Fig. 6. Cycle-wise ROC curves with marked optimum thresholds of all the models, when features were selected based on a) only the Squat dataset ($AUC_{Squat} = 0.87$ and $AUC_{FE} = 0.83$), b) the Squat + FE datasets ($AUC_{Squat} = 0.89$ and $AUC_{FE} = 0.84$), and c) only the FE dataset ($AUC_{Squat} = 0.86$ and $AUC_{FE} = 0.83$).

TABLE IV
SUMMARY OF KNEE HEALTH SCORES

	FFS Based on Squat	FFS Based on Squat + FE	FFS Based on FE
JIA-Pre (Squat)	0.86 ± 0.09	0.91 ± 0.07	0.81 ± 0.16
JIA-Pre (FE)	0.80 ± 0.12	0.87 ± 0.11	0.77 ± 0.14
JIA-Post (Squat)	0.50 ± 0.09	0.40 ± 0.12	0.51 ± 0.11
JIA-Post (FE)	0.37 ± 0.11	0.35 ± 0.08	0.40 ± 0.09
Healhty (Squat)	0.24 ± 0.15	0.24 ± 0.19	0.23 ± 0.17
Healthy (FE)	0.21 ± 0.15	0.18 ± 0.15	0.17 ± 0.15

to be better than cycle-wise accuracy as it was averaged across all movement cycles of each subject. The cycle-wise AUC was in 0.83–0.89 range, and the cycle-wise validation accuracy was between 76–84%. The optimum thresholds of the classifiers were determined based on the optimal point of the receiver operating characteristics (ROC) as described in [37]. Fig. 6 illustrates the ROC curves and the optimum thresholds of all three models for FE and Squat groups. Of importance, several features were common across these three cases and the classifier accuracy was not very sensitive to the exact number of features. Thus, adjusting the number of features changed the accuracy by a few percent as long as the number of features was in a close range. Note that with this FFS algorithm, the selected features are not totally uncorrelated, and the Pearson's cross-correlation coefficients between the features in the same model were 0.37 ± 0.29, 0.33 ± 0.23, 0.31 ± 0.24 for the feature selection based on Squat, FE, and combined datasets, respectively.

C. Predicted Knee Health Scores and Feature Importance

Fig. 7 (a–c) shows the resulting scores for three sets of features explained in Section III-B, where the box plots and violin plots of squats (in blue) and FEs (in purple) were shown separately for pre- and post-treatment subjects with JIA, as well as healthy subjects. In these plots, a score of 0 corresponds to a healthy subject, a score of 1 corresponds to a subject with JIA; the score of each subject was calculated through averaging all the movement cycle scores of that subject, and a summary is reported in Table IV. With features selected based on squat dataset, the pre-treatment JIA knee scores reduced from 0.86 ± 0.09 and 0.80 ± 0.12 to 0.50 ± 0.09 and 0.37 ± 0.11 after 3–6 months of treatment, for squat and FE exercises respectively. The knee scores of the healthy group were 0.24 ± 0.15 and 0.21 ± 0.15, for squat and FE exercises respectively. Similarly, with FFS based on FE dataset, the pre-treatment JIA knee scores reduced from 0.81 ± 0.16 and 0.77 ± 0.14 to 0.51 ± 0.11 and 0.40 ± 0.09 after treatment, respectively. The healthy knee scores for squat and FE exercises were 0.23 ± 0.17 and 0.17 ± 0.15, respectively. Finally, when the FFS was performed based on the combined dataset, the pre-treatment JIA knee scores decreased from 0.91 ± 0.07 and 0.87 ± 0.11 to 0.40 ± 0.12 and 0.35 ± 0.08 after treatment, for squat and FE exercises respectively. Furthermore, the healthy knee scores were 0.24 ± 0.19 and 0.18 ± 0.15, for squat and FE exercises respectively.

Based on the results presented in Fig. 7, the followings are concluded: 1) The post-treatment scores were always lower than the pre-treatment scores, confirming the clinical records of successful treatment; 2) The squat knee scores were higher on average compared to FE knee scores for all groups in all three feature settings. Similarly, the optimum threshold levels of the classifiers for squat recordings were higher than those of the FE recordings. This implies that the effects of loading on the joint sounds is more pronounced in patients with JIA; 3) The difference between post-treatment squat and FE scores was statistically significant ($p < 0.05$ using two sample Kolmogorov–Smirnov test) when the FFS was performed on either the squat or the FE dataset (Fig. 7(a) and Fig. 7(c)), as the features were selected to maximize the separability of healthy and patient groups; and 4) The difference between post-treatment squat and FE scores was not statistically significant ($p>0.05$ using two sample Kolmogorov–Smirnov test) when the FFS was performed on the combined dataset (Fig. 7(b)), as the features were selected in a way to assign similar knee scores to both squat and FE exercises.

Feature importance was analyzed based on the feature coefficients of trained classifiers, $\beta_0, \beta_1, \ldots, \beta_N$ in (1). The top 15 features in each case were shown on Fig. 7 (d–f), and it is interesting to see that these features consist of several types of temporal, spectral, and MFCC features, which represents the necessity of a diverse feature set for a JIA versus healthy classification task. Some of these top features were common among the three cases; for instance, spectral crest and MFCC 10 were among the top four features and the top seven features in all the three settings, respectively. Note that these time-frequency features are to some extent correlated with each other and the feature rankings are highly dependent to the dataset, trained models, and the selected features in each model. In addition, it was interesting to see that our most important features were consistent with some of the previously reported features: ZCR, spectral spread, MFCC10, MFCC13, energy, spectral crest, and spectral entropy was common with [18] and [20].

The knee health scores evaluated based on the features selected from a combined dataset shows that machine-learning algorithms are capable of processing the knee JAEs regardless of the movement type and still providing a similar knee score. The optimum thresholds reported in Table III were chosen based on the classifiers' ROC to improve their performance. However,

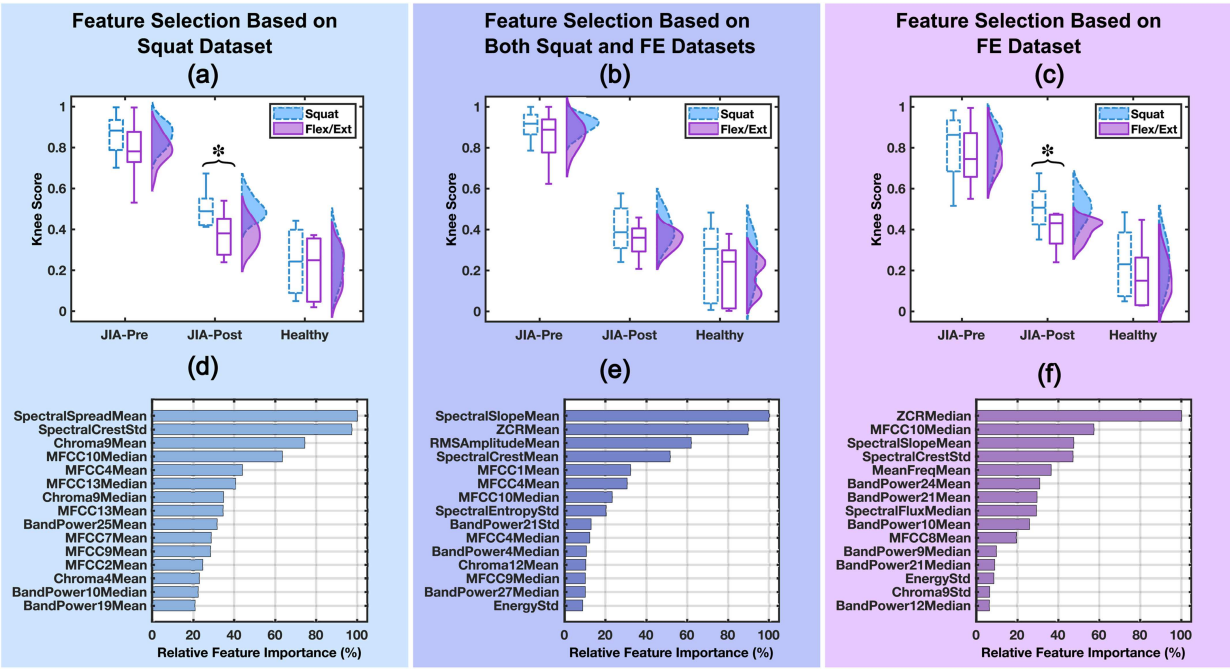

Fig. 7. Box plots, violin plots, and feature importance rankings, when the feature selection was performed based on the squat dataset (a and d), the combined dataset (b and e), and the FE dataset (c and f). The asterisk ($*$) represents statistical significance ($p < 0.05$) based on a two-sample Kolmogorov-Smirnov test.

performance of the classifiers was also calculated with a default threshold of 0.5, and the subject-wise validation accuracies only degraded by less than 5%, which shows that the classifiers were robust and not sensitive to the small changes of threshold levels.

The scores of healthy subjects between squat and FE were expected to be similar as shown in Fig. 7 (a–c), since healthy joints for kids have intact cartilage and minimal damage / wear-and-tear, and thus the loading of the joint, effecting how "hard" the internal surfaces were rubbed, does not substantially impact the JAEs produced. In subjects with JIA, treatment reduces synovitis quickly, but if the synovial thickness was still greater than a healthy joint, and / or if there is any damage to cartilage—even if minimal—then there could still be some residual "roughness" of the articulating surfaces. This roughness may be too small to make any change in the JAEs in the unloaded FE state, but when the knees were under pressure of the body load, that small amount of increased roughness can introduce more friction into the movement and concomitant changes in JAEs. This explanation supports the higher squat scores than the FE scores in all three feature selection cases. Although a detailed analysis on the joint force can further reveal the loading effect, due to limitations of collecting the data in a clinical environment and especially on children, this analysis only relied on the loading change from an unloaded suspended knee while doing the FE to the loaded knee of squats, and the anatomical information and kinematics data of the subjects were not collected due to the large number of sensors needed.

The clinical diagnosis of JIA is typically a diagnosis of exclusion, meaning that to reach the diagnosis doctors must first rule out a variety of other inflammatory and infectious etiologies, most notably septic/reactive arthritis and musculoskeletal injuries. There are no lab tests that are specific for JIA and instead a diagnosis is formed based on a constellation of non-specific inflammatory markers, patient history and physical exam. The gold standard of diagnosis is a trained physician's exam, which is what our algorithm has been compared against. Furthermore, imaging (X-rays or MRI) is not the standard method for diagnosis and if it shows pathology like generic inflammatory changes, they are still not specific for JIA. No technique exists currently for non-invasively diagnosing JIA - accordingly, the authors believe that the methodology presented here may have impact in the management and care of persons with suspected JIA.

IV. Conclusion and Future Work

In this study, we investigated knee JAEs from both unloaded and loaded exercises in a pediatric population with JIA. JAEs from both load-states can differentiate between healthy and arthritic joints and could thus be used as part of a diagnostic plan. However, diagnosis is only a small part of the potential application of this sensing modality. As discussed, there is a gap in the care of JIA management at least partially related to the shortage of pediatric rheumatologists. JAE sensing provides a novel type of non-invasive monitoring of joint health which could help bridge that gap. If JAEs are capable of elucidating longitudinal changes of JIA in relation to exacerbations or therapies, then treatment could be better personalized and titrated for each child. To that end, changes in JAEs following successful therapy were analyzed in this paper.

It had previously been shown that unloaded JAEs after successful treatment were similar to healthy JAEs. In this study, we found that in the loaded case JAEs similarly trended toward

healthy JAEs, but they did not reach the high level of correlation seen in the unloaded state. This suggests that by loading the joint, the JAEs were more sensitive to persistent changes in the micro-architecture of the articulating joint. These joints following treatment were qualitatively healthy but may have had latent changes related to the preceding period of inflammation and degradation. To further understand this difference in trend between unloaded and loaded JAEs following treatment, in the future more timepoints as well as a larger population with a more comprehensive joint work-up should be recorded and recruited. We believe that with a larger population recorded longitudinally, the full capabilities of JAE analysis will be better understood.

Furthermore, there were patients from all major categories of JIA among the participants of this study. Thus, the classifier appears to detect JIA compared to healthy controls regardless of the disease subtype. Although the dataset has a high variation in the disease severity, for this initial study the ground truth disease severity information was not available from the clinical collaborators since they only performed the standard physical examination as part of their assessment. Future work will examine further analyzing differences in joint acoustic emissions across subtypes of JIA and across disease severity. It is possible that this type of sensing could one day be used not only for diagnosis, but also for adverse event or prediction of acute flare-ups, subtypes, and optimization of therapy.

References

[1] M. W. Beresford, "Juvenile idiopathic arthritis," *Pediatr. Drugs*, vol. 13, no. 3, pp. 161–173, 2011.

[2] T. Beukelman *et al.*, "2011 American college of rheumatology recommendations for the treatment of juvenile idiopathic arthritis: Initiation and safety monitoring of therapeutic agents for the treatment of arthritis and systemic features," *Arthritis Care Res.*, vol. 63, no. 4, pp. 465–482, 2011.

[3] A. Ravelli *et al.*, "Juvenile idiopathic arthritis," in *The Heart in Rheumatic, Autoimmune and Inflammatory Diseases*. Elsevier, Cambridge, MA, USA, 2017, pp. 167–187.

[4] A. C. Brescia, *Juvenile Idiopathic Arthritis*, Apr. 2016. Accessed: May 23, 2020. [Online]. Available: https://kidshealth.org/en/parents/jra.html

[5] A. Consolaro *et al.*, "Development and validation of a composite disease activity score for juvenile idiopathic arthritis," *Arthritis Care Res.: Official J. Amer. College Rheumatol.*, vol. 61, no. 5, pp. 658–666, 2009.

[6] H. Srinivasalu and M. Riebschleger, "Medical education in pediatric rheumatology - unique challenges and opportunities," *Clin. Rheumatol.*, vol. 39, no. 3, pp. 643–650, 2020.

[7] "Arthritis foundation, addressing the pediatric rheumatology shortage," Accessed: Jul. 15, 2020. [Online]. Available: https://www.arthritis.org/advocate/federal/addressing-the-pediatric-rheumatology-shortage

[8] R. M. Rangayyan and Y. Wu, "Screening of knee-joint vibroarthrographic signals using statistical parameters and radial basis functions," *Med. Biol. Eng. Comput.*, vol. 46, no. 3, pp. 223–232, 2008.

[9] S. Krishnan, R. M. Rangayyan, G. D. Bell, and C. B. Frank, "Adaptive time-frequency analysis of knee joint vibroarthrographic signals for non-invasive screening of articular cartilage pathology," *IEEE Trans. Biomed. Eng.*, vol. 47, no. 6, pp. 773–783, Jun. 2000.

[10] A. Theodorakidis *et al.*, "What juvenile idiopathic arthritis?," 2017. Accessed: Apr. 17, 2020. [Online]. Available: https://www.aboutkidshealth.ca/Article?contentid=1049

[11] C. A. Wallace, "Current management of juvenile idiopathic arthritis," *Best Pract. Res. Clin. Rheumatol.*, vol. 20, no. 2, pp. 279–300, 2006.

[12] S. Shrivastava and R. Prakash, "Assessment of bone condition by acoustic emission technique: A review," *J. Biomed. Sci. Eng.*, vol. 2, no. 3, pp. 144–154 , 2009.

[13] Y. Wu, *Knee Joint Vibroarthrographic Signal Processing and Analysis*. Berlin, Germany: Springer, 2015.

[14] R. E. Andersen *et al.*, "A review of engineering aspects of vibroarthography of the knee joint," *Crit. Rev. Phys. Rehabil. Med.*, vol. 28, no. 1-2, pp. 13–32, 2016.

[15] S. Hersek *et al.*, "Acoustical emission analysis by unsupervised graph mining: A novel biomarker of knee health status," *IEEE Trans. Biomed. Eng.*, vol. 65, no. 6, pp. 1291–1300, Jun. 2018.

[16] H. K. Jeong, M. B. Pouyan, D. C. Whittingslow, V. Ganti, and O. T. Inan, "Quantifying the effects of increasing mechanical stress on knee acoustical emissions using unsupervised graph mining," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 26, no. 3, pp. 594–601, Mar. 2018.

[17] N. B. Bolus *et al.*, "A glove-based form factor for collecting joint acoustic emissions: Design and validation," *Sensors*, vol. 19, no. 12, 2019, Art. no. 2683.

[18] B. Semiz *et al.*, "Using knee acoustical emissions for sensing joint health in patients with juvenile idiopathic arthritis: A pilot study," *IEEE Sensors J.*, vol. 18, no. 22, pp. 9128–9136, Nov. 2018.

[19] B. Semiz *et al.*, "Change point detection in knee acoustic emissions using the teager operator: A preliminary study in patients with juvenile idiopathic arthritis," in *Proc. IEEE EMBS Int. Conf. Biomed. Health Informat.*, 2019, pp. 1–4.

[20] D. C. Whittingslow *et al.*, "Knee acoustic emissions as a digital biomarker of disease status in juvenile idiopathic arthritis," *Front. Digit. Health*, vol. 2, pp. 1–12 , 2020.

[21] B. Mascaro *et al.*, "Exploratory study of a non-invasive method based on acoustic emission for assessing the dynamic integrity of knee joints," *Med. Eng. Phys.*, vol. 31, no. 8, pp. 1013–1022, 2009.

[22] N. Tanaka and M. Hoshiyama, "Vibroarthrography in patients with knee arthropathy," *J. Back Musculoskelet. Rehabil.*, vol. 25, no. 2, pp. 117–122, 2012.

[23] C. N. Teague *et al.*, "Novel methods for sensing acoustical emissions from the knee for wearable joint health assessment," *IEEE Trans. Biomed. Eng.*, vol. 63, no. 8, pp. 1581–1590, Aug. 2016.

[24] K. L. Scherpereel, "Estimating knee joint load using acoustic emissions during ambulation," *Ann. Biomed. Eng.*, vol. 49, no. 3, pp. 1000–1011, 2021.

[25] K. S. Kim *et al.*, "An enhanced algorithm for knee joint sound classification using feature extraction based on time-frequency analysis," *Comput. Methods Programs Biomed.*, vol. 94, no. 2, pp. 198–206, 2009.

[26] T. F. Lee *et al.*, "Analysis of vibroarthrographic signals for knee osteoarthritis diagnosis," in *Proc. 6th Int. Conf. Genet. Evol. Comput.*, 2012, pp. 223–228.

[27] C. N. Teague *et al.*, "A wearable, multimodal sensing system to monitor knee joint health," *IEEE Sensors J.*, vol. 20, no. 18, pp. 10323–10334, Sep. 2020.

[28] S. Gharehbaghi *et al.*, "Joint acoustic emissions as a biomarker for knee health assessment in loaded and unloaded exercises," in *Proc. Amer. Soc. Biomech. Annu. Meeting*, 2020, p. 1.

[29] R. K. Saurenmann *et al.*, "Risk factors for development of uveitis differ between girls and boys with juvenile idiopathic arthritis," *Arthritis Rheumatism*, vol. 62, no. 6, pp. 1824–1828, 2010.

[30] T. Giannakopoulos and A. Pikrakis, *Introduction to Audio Analysis: A MATLAB Approach*. New York, NY, USA: Academic, New York, NY, USA, 2014.

[31] T. Hastie *et al.*, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer Sci. Bus. Media, New York, NY, USA, 2009.

[32] I. Guyon and A. Elisseeff, "An introduction to variable and feature selection," *J. Mach. Learn. Res.*, vol. 3, pp. 1157–1182, Mar, 2003.

[33] J. Tang, S. Alelyani, and H. Liu, "Feature selection for classification: A review," *Data Classification: Algorithms and Applications*. Boca Raton, FL, USA: CRC Press, 2014, pp. 37–64.

[34] A. Janghorbani *et al.*, "Prediction of acute hypotension episodes using logistic regression model and support vector machine: A comparative study," in *Proc. 19th Iranian Conf. Elect. Eng.*, 2011, pp. 1–4.

[35] A. Géron, *Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems*. O'Reilly Media, Sebastopol, CA, USA. 2019.

[36] M. Pohar *et al.*, "Comparison of logistic regression and linear discriminant analysis: A simulation study," *Metodoloski Zvezki*, vol. 1, no. 1, pp. 143–161 , 2004.

[37] MathWorks, performance curves, Accessed: Jun. 11, 2020. [Online]. Available: https://www.mathworks.com/help/stats/performance-curves.html