Acoustical Features as Knee Health Biomarkers: A Critical Analysis

Christodoulos Kechris[a,*], Jerome Thevenot[a], Tomas Teijeiro[a,b], Vincent A. Stadelmann[c], Nicola A. Maffiuletti[d], David Atienza[a]

[a]*Embedded Systems Laboratory (ESL), Ecole Polytechnique Federale de Lausanne (EPFL), Switzerland*
[b]*Basque Center for Applied Mathematics (BCAM), Spain*
[c]*Department of Research and Development, Schulthess Klinik, Zürich, Switzerland*
[d]*Human Performance Lab, Schulthess Klinik, Zürich, Switzerland*

Abstract—Acoustical knee health assessment has long promised an alternative to clinically available medical imaging tools, but this modality has yet to be adopted in medical practice. The field is currently led by machine learning models processing acoustical features, which have presented promising diagnostic performances. However, these methods overlook the intricate multi-source nature of audio signals and the underlying mechanisms at play. By addressing this critical gap, the present paper introduces a novel causal framework for validating knee acoustical features. We argue that current machine learning methodologies for acoustical knee diagnosis lack the required assurances and thus cannot be used to classify acoustic features as biomarkers. Our framework establishes a set of essential theoretical guarantees necessary to validate this claim. We apply our methodology to three real-world experiments investigating the effect of researchers' expectations, the experimental protocol and the wearable employed sensor. This investigation reveals latent issues such as underlying shortcut learning and performance inflation. This study is the first independent result reproduction study in the field of acoustical knee health evaluation. We conclude with actionable insights from our findings, offering valuable guidance to navigate these crucial limitations in future research.

Index Terms—Knee Acoustic Emissions, Knee Biomarkers, Causal Machine Learning, Shortcut Learning, Explainable-AI, Knee Arthritis

I. INTRODUCTION

Knee health evaluation by examining the vibrations of the joint has been proposed as an alternative to other diagnostic approaches such as medical imaging. This family of methods, also referred to as Vibration Arthrometry, are based on the premise that structural imperfections / damage to an unhealthy knee will generate abnormal vibrations [1]. Capturing and characterizing these abnormalities offers the opportunity to develop novel diagnostic tools. In one of the first medical reports on knee acoustic emissions [2], Blodgett set the tone: Finding patterns useful for diagnosis would prove to be a challenging task. He concluded his report by stretching the main difficulty of the acoustical modality: Unlike other diagnostic tools, such as X-ray imaging, the interpretation of the knee audio signal is significantly based on the *"imagination of the observer"*.

Blodgett's initial work effectively kicked off efforts to increase the diagnostic value of acoustic signals for knee

pathologies. In the 1990s, significant advances in sensor and signal processing technologies led to efforts to gain more insights and understanding of these audio recordings [3], [4], [5].

The introduction of Machine Learning (ML) shifted the field's scope from identifying, understanding and modelling audio signals to directly developing *features* relevant to several knee conditions, utilizing ML methods for knowledge discovery. The workflow usually adopted is the following [6], [1], [7], [8], [9] : (i) A vibration or acoustical sensor is placed on the skin surface near the region of the knee. (ii) The sensor records vibration/audio signals while the patient performs specific postular transitions, usually unloaded flexion-extension or loaded sitting-to-stand. (iii) After the recording session has finished, the acquired signals are filtered. Filtering usually involves bandpass filtering [8], or discarding regions of signals that are considered to be significantly affected by external artifacts [9]. (iv) In the final step, the features are extracted from the recordings and used as input to a ML algorithm to infer the health status of the recorded knee, by classifying it in a range between healthy and pathological with respect to a disease. The acoustic features or the model's output is then treated as a "biomarker" to evaluate the patient's knee health. A plethora of methods have emerged following this workflow, presenting optimistic results based on diagnosis classification accuracies as high as 96% [10].

Despite the overall positive reports, such approaches have not yet been adopted in medical practice [11]. Furthermore, the approach presented in knee acoustical biomarker reports is contradicted by the current medical standards: acoustical biomarkers are reported to be capable of successfully diagnosing knee pathologies without any additional modalities or clinical information. In contrast, medical guidelines recommend a multi-modal approach involving patient history, clinical and imaging examinations [12], [11]. In their 2013 review, Abott et al. [10] critiqued the stagnation of the Vibration Arthrometry field: the medical community approaches the field with skepticism, and knee acoustic emissions have not yet found their place in diagnostic applications. The authors argued that for the field to move forward, better methodological practices are needed to instill trust.

However, ten years later, the validity of vibration arthrometry as a diagnostic tool is based solely on the promising performance of reported ML models trained on data collected through retrospective studies [13]. Knowledge of

how pathological acoustical signals manifest is fairly limited. Nevertheless, classification performance alone is not enough to guarantee robustness in ML knowledge discovery [14], [15]. In the context of biomedical ML, external sources of information can inflate the model's performance [16], [17], [18] resulting in wrong conclusions. The model then learns shortcuts: unwanted patterns that do not correspond to the pathophysiology of the patients. The solution is to identify the data structures which lead the model to learn these spurious relations and isolate them. However, in vibration arthrometry, we are still far from what has been achieved in other biomedical fields where there is prior medical knowledge of how body structures as well as pathophysiological conditions are manifested in the corresponding modality. For example, in medical imaging, there is a general understanding of bone and soft tissue structures, information introduced by the machine, as well as common artifacts [19].

Such insights about knee acoustical signals are considerably limited: identifying the source for each component is not trivial. In fact, it is the goal of the entire acoustical-biomarker-discovery exploration. This realization bears two consequences: (i) Identifying possible sources of bias in knee-audio datasets is a significantly more challenging task than in other medical modalities, e.g. images. (ii) Consequently, attribution of statistical differences in acoustical features between populations to a pathophysiological source is not a straightforward process.

In this work, we contribute to the field of Knee Vibroarthrometry by introducing a novel causal framework as a tool to robustly validate acoustical knee biomarkers. We argue that the current state of the art in Knee Vibroarthrometry has failed to provide satisfying evidence in support of the field's main hypothesis: unhealthy knee joints produce abnormal vibrations. We begin our exploration by formulating a causal description of the task at hand. Through this theoretical investigation we highlight the main pitfalls of the state-of-the-art workflow and demonstrate that features extracted from knee audio recordings cannot serve as health biomarkers. We support our theoretical approach with our findings on real-world data. To the best of our knowledge, our work is the first independent result reproduction study. Finally, we show how this causal exploration allows for the proposal of a set of strict guarantees that should be met in order to reach robust conclusions.

II. THEORETICAL EXPLORATION

We will now attempt to construct a more formalized description of the problem and link this theoretical notation to the workflow adopted by proposed ML methods. An illustrative summary of our approach is presented in Figure 1.

We start with the health condition of the knee, which we will refer to as H. Being a complex structure, the joint can be affected by several conditions, such as osteoarthritis, meniscal or ligament tear, etc. [12]. However, to simplify our analysis, for now we will ignore the vast range of plausible conditions and treat them all the same. Hence, without loss of generality we assume that variable H can take two values: $Healthy$ and $Unhealthy$. A similar approach is also adopted in most publications where unhealthy knees are conditioned by a specific pathology, e.g. meniscus tear [20]. Furthermore, our approach can also be extended to multi-class classification [21].

The joint health, H, will affect the structural integrity of the knee [12], and the general consensus is that by extension the vibrations formed in the joint region during motion will also be affected. It is important here to emphasize that this consensus is a hypothesis that needs to be supported by evidence. In an ideal scenario, a vibration sensor should be placed on the affected component of the knee (e.g. meniscus or ligament) to capture these hypothesized vibrations. We refer to these ideal vibration signals as V. Thus the main hypothesis can be stated as:

$$H \rightarrow V \tag{1}$$

Unfortunately, such an ideal observation is practically infeasible due to the complexities involved in designing and implementing the experimental setup and doing so without interfering with the observed mechanism. In practice, the vibration sensor is placed non-intrusively on the surface of the skin, in strategic areas near the knee [1], [22], [23]. The goal of this experimental setup is to capture the best possible approximation of V, $\tilde{V}$. Thus, the Main Hypothesis of Eq. 1, transforms into the real-world hypothesis:

$$H \rightarrow \tilde{V} \tag{2}$$

This transformation bears the implicit assumption that $\tilde{V}$ approximates the ideal vibration V in a satisfactory manner. As such, Eq. 1 is equivalent to Eq. 2 under the approximation constraint:

$$\tilde{V} \approx V \tag{3}$$

On the contrary, if this constraint cannot be guaranteed, the equivalence between Eq. 1 and Eq. 2 cannot be guaranteed either. Of course, robustly validating Eq. 3 presents challenges since access to the actual V is infeasible. Thus, in a real-world setting, we could only require evidence hinting at this constraint. We will elaborate on this issue later on.

Having acquired the observed signal $\tilde{V}$, the next step is to extract features to describe its qualitative properties that are hypothesized to be linked to the patient's knee health status. An intermediate step of filtering is also usually employed before feature extraction; however, for simplicity let us assume that $\tilde{V}$ has already been filtered and contains the final vibration approximation signal accompanied by all necessary guarantees, i.e. Eq. 3. Let g be the function that maps the recording $\tilde{V}$ to a set of features X, then:

$$X = g(\tilde{V}) \tag{4}$$

In the final step a classifier, $f(\cdot)$ is trained:

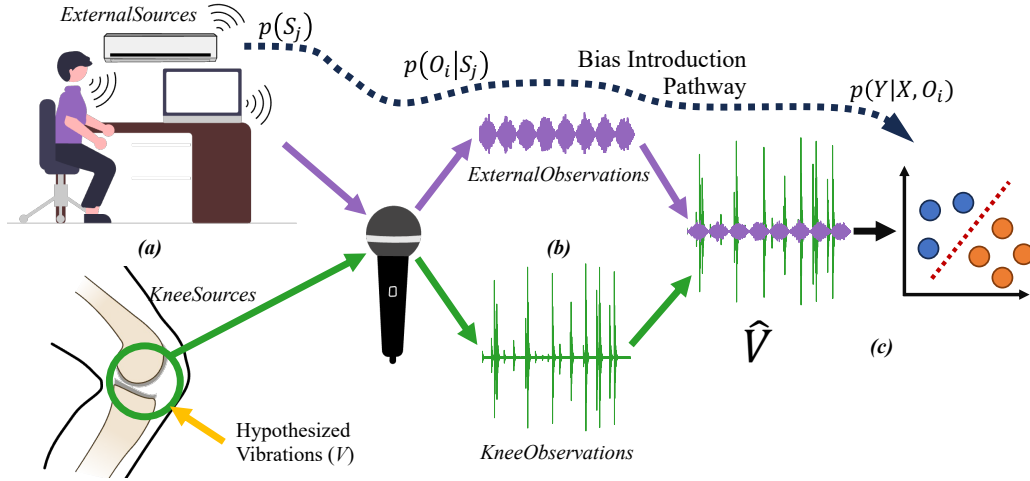$$Y = f(X) = f \circ g(\tilde{V}) \tag{5}$$

Fig. 1: Illustration of our proposed causal framework for investigating and validating acoustic knee biomarkers. **(a)** Multiple audio source may be present in the experimental environment, including the examined knee. **(b)** These source will potentially be picked up by the acoustical sensor and may influence the extracted acoustical features. **c** These external source may introduce bias in the final results, severely boosting the model's performance and leading to wrong conclusions.

so that a loss function between Y and H is minimized. There are many works achieving satisfying levels of classification accuracy [10]. Interpreting these high classification results is a non-trivial task. The most common argumentation is that they provide evidence supporting the original hypothesis Eq. 2, and by extension Eq. 1. The conclusion is indeed that (i) health deterioration of the knee can affect the vibrations it emits and (ii) the experimental setup is capable of capturing and identifying these vibrations.

Here, we propose a more reserved and pragmatic interpretation. In fact, experiments in small datasets show that Y, and by extension (Eq. 5) $\tilde{V}$, are correlated with H. As such, according to the Common Cause Principle [24], there exists a third variable, Z, that causes the dependence between $\tilde{V}$ and H and the conditioning of which renders $\tilde{V}$ and H independent. In the extreme case, Z can be the same as H, which is the case adopted by current literature. Making this claim however requires proper guarantees for Eq. 3 in order to reduce the possibility of a third external source of information. That in combination with high $\tilde{V} - H$ correlations could provide strong evidence in support of the argument that Z is indeed H further supporting the original hypothesis from Eq. 1 and 2.

Let's now elaborate on the classification task starting with Eq. 5 describing the output of the classifier inferring a knee's health by its acoustical observations as input. Observation X is inherently a multi-source signal, and thus exploration of Eq. 5 should take into consideration X's multi-source nature.

We begin our exploration with the observation $\tilde{V}$, and its representation X in an appropriate domain. A finite set of N sources $Sources = S_0, ..., S_{N-1}$ which are observable by our sensor contribute to $\tilde{V}$ (Figure 1a). For each source S_i our sensor observes a corresponding observation O_i forming the finite set of N observations $Observations = O_0, ..., O_{N-1}$. The distinction between the $Sources$ and the $Observations$ is equivalent to that between V and $\tilde{V}$. $Sources$ refer to real-world events, the knowledge of which requires perfect prior information about the world in which the experiments take place. On the contrary, $Observations$, is the set of signal sources that the sensor observes.

To give more insight on the introduced notation, let us consider the example of the knee acoustic sensor. The event $Source = Knee$ signifies that the knee produces a vibration during the acquisition of the observation $\tilde{V}$. $Observation = Knee$ refers to the event of the sensor capturing the knee and hence X can partly be attributed to the knee. Thus, the probability $p(Observation = Knee \mid Source = Knee)$, describes the sensor's ability to capture a knee acoustic event. Similarly $Observation = Knee \mid Source = \overline{Knee}$ is impossible to happen. Now imagine that in the experimental environment there is an additional external mechanism generating sound waves, for example, the room's ventilation or air-conditioner (A/C) (Figure 1a). Then $p(Observation = \overline{A/C} \mid Source = (Knee, A/C))$ describes the device's ability to filter out the external noise of the A/C.

For the classification task, given an acoustical observation $\tilde{V}$ we need to estimate the most probable knee health status Y:

$$p(Y \mid X) = p(Y \mid g(\tilde{V})) \tag{6}$$

We consider the multi-source scenario and the notation introduced in Eq. 6. Also, for now the observations O_i, O_j $\forall O_i, O_j \in Observations$ are considered disjoint and similarly all sources S_i, S_j $\forall S_i, S_j \in Sources$. In the real world this assumption cannot be guaranteed; in the previous example the A/C could operate at the same time as the subject mobilizes the knee joint. We will deal with this scenario later. For now, we can rewrite Eq. 6 as:

$$p(Y \mid X) = \sum_i p(Y \mid X, O_i)p(O_i \mid X) \tag{7}$$

The first term of the sum, $p(Y \mid X, O_i)$, describes a model trying to infer Y from a sample X attributed to observation O_i. The second term, $p(O_i \mid X)$, is essentially a source identification model, attributing the sample X to observation O_i. Through Eq. 7, the initial source-neglecting model is rewritten into a source-aware model that performs source separation and source-specific classification. Using Bayes' rule, we elaborate on the source identification term, $p(O_i \mid X)$:

$$p(O_i \mid X) = \frac{p(X \mid O_i) \cdot p(O_i)}{p(X)} \tag{8}$$

$p(O_i)$ denotes the probability that observation i is observed in our experimental setting using the sensor. Intuitively, we can consider that capturing an observation from source S_i requires that the actual source S_i occurs, $p(S_i)$, and that our sensor is capable of observing this event when it is happening, $p(O_i|S_i)$. As such, considering all possible sources S_j, we can split $p(O_i)$ as:

$$p(O_i) = \sum_j p(O_i \mid S_j)p(S_j) \tag{9}$$

Putting together Eq. 6, 7, 8 and 9:

$$p(Y \mid X) = \frac{1}{p(X)} \sum_i p(Y \mid X, O_i)p(X \mid O_i) \\ \cdot \sum_j p(O_i \mid S_j)p(S_j) \tag{10}$$

For the use-case of classifying the health status of a knee, H, from an acoustical sample X, we can split the set of Sources - Observations into two subsets: those that can be attributed to the knee, and those that are attributed to external phenomena, e.g. A/C :

$$p(Y \mid X) \propto \sum_{\substack{Knee \\ Observations}} p(Y \mid X, O_i)p(X \mid O_i) \\ \cdot \sum_j p(O_i \mid S_j)p(S_j) \\ + \sum_{\substack{External \\ Observations}} p(Y \mid X, O_i)p(X \mid O_i) \\ \cdot \sum_j p(O_i \mid S_j)p(S_j) \tag{11}$$

Eq. 10 interestingly reveals that: inferring a knee's health condition by its acoustic observations implicitly requires understanding of the knee behavior and the sensor's ability to capture it.

Eq. 11 shows how a model $f(X)$ inferring $p(Y \mid X)$ can still draw information from external sources as long as $p(Y \mid X, O_i)$ is informative for at least one of them, creating a Bias Introduction Pathway (Figure 1). This equation also reveals a set of strategies to achieve a robust experimental setting. In the Approximation Constraint, Eq. 3, we required knowledge of the ideal knee observation V, which was considered physically infeasible. Here, we require to minimize the effect of the second summation term, so as to ensure that no external source interferes with the classification task, breaking the Bias Introduction Pathway. This can be achieved by optimizing in three different directions:

1) $p(Y \mid X, O_i)$: External observations should be random and not correlated with knee health.
2) $p(O_i \mid S_j)$: The sensor should be shielded from interference sources, so that external signals do not affect knee recording.
3) $p(S_j)$: Ideally, there should not be any external sources of interference in the experimental setting. The only source of information present in the audio recordings, $\tilde{V}$ should be the knee mechanism.

For each external observation O_i in the world of our experiments, ensuring at least one of these conditions is enough to provide a robust dataset, safe from external biases. Achieving this is not trivial since one needs to first identify all possible sources of external interference and then study their effect on the experiment. However, this task is considerably more feasible than acquiring direct access to the ideal recording $\tilde{V}$.

An important assumption that was necessary to derive Eq. 10 was that the sources in the set of $Sources$ are pairwise disjoint, and similar for the observations. In the real world however this assumption does not necessarily hold true, and multiple events can occur during the recording of an observation $\tilde{V}$. To overcome this limitation, we define $Sources$ as follows. First, we define the finite set of all N sources occurring in the experiment S, the complementary set $\overline{S} = \{\overline{x}, x \in S\}$ and their union $A = S \cup \overline{S}$. Now consider one subset $M_i \subset A$ such that:

$$|M_i| = N \tag{12}$$

and

$$\forall x_i, x_j \in M_i, x_i \neq \overline{x_j} \tag{13}$$

Define the event S_i as the intersection of all elements of M_i, $S_i = \bigcap_{x_j \in M_i} x_j$. Due to Eq. 12 and 13, given two different sets S_i and S_j, they are disjoint. Finally the $Sources$ set is defined as the union of all possible S_i. The set of $Observations$ can be defined in a similar way. For example, in the case where we only consider the sources of the Knee and the A/C, instead of defining $Sources = \{Knee, A/C\}$, $Sources$ should be defined as:

$$Sources = \{(Knee, A/C), (\overline{Knee}, A/C), \\ (Knee, \overline{A/C}), (\overline{Knee}, \overline{A/C})\} \tag{14}$$

III. Real-world Lessons

So far, we have established a framework to describe and characterize the task at hand and the tools employed to tackle it, i.e. classification. However, our analysis has remained mostly theoretical. In this section we complement this exploration with three real-life study cases. Simultaneously we provide key takeaways from our investigation and literature regarding best practices when designing knee acoustic emission experiments and evaluating their results.

A. The Relevance of the Expectations

We begin with a counterfactual thought experiment involving the design of a study to identify and evaluate acoustical biomarkers. The question we raise is: Will acoustical features reveal knee structural differences, even when they should not? We show how the researchers' expectations regarding the experiment's output, i.e. how the samples are split in health/unhealthy groups, can influence the interpretation of the features. An illustration of the counterfactual thought-experiment is presented in Figure 2.
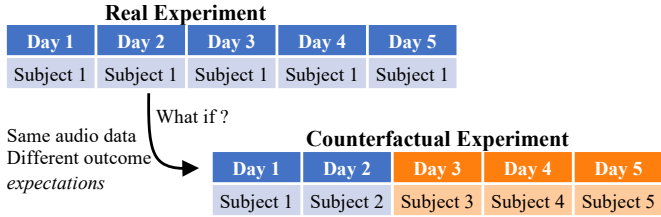

Fig. 2: Illustration of our counterfactual experiment. By changing the expectation on the output (asking what if) the same audio data are interpreted differently. This interpretation is not necessarily causally linked to the underlying knee mechanism we are trying to describe.

The exploration is set up as follows. We perform a real experiment with a single subject who underwent measurements on the right knee for five consecutive days. Following a similar workflow as the ones proposed in the relevant literature, we employed a wearable acoustical device to record knee audio signals. The wearable device uses two microphones to record signals from the sub-patella region: one on the medial and one on the lateral side of the knee. On each day the subject performed 6 unloaded reciprocal knee flexion-extensions, an exercise that is usually used for this purpose [1], [9], [8], [20], [5], [23], [25], [26], [27]. After collection, a set of audio features was extracted, following the work of [9], [8], [23], [26], [27]. All recordings were taken approximately at the same time of the day, between 12:00 and 12:30 and no injury was reported throughout the week. Hence, we have no reason to believe that there was any mechanical degradation in the subject's knee mechanism that would affect its acoustic emissions throughout the course of the experiment. Under these assumptions, any variance in the audio features is interpreted not as fluctuations in the knee mechanism but rather as variance in the measurement process.

Using the same audio recordings, we ask a counterfactual question, changing the researchers' expectations: What if the data had been recorded from 5 different subjects? In this scenario, the first 2 knees are healthy, while the next 3 are unhealthy. Now we are looking to identify the most relevant audio features that will serve as our proposed knee health biomarkers. After our analysis, we conclude that the Mel Frequency Cepstral Coefficient (MFCC) features [28], specifically MFCC 8 and MFCC 11, have a good discriminant validity, as they show a clear separation between the two populations (Figure 3). Furthermore, similar audio features have already been successful in discriminating healthy and

unhealthy subjects [9]. The study of this counterfactual thought experiment is concluded with very positive remarks. We have managed to identify promising audio biomarkers, capable of identifying unhealthy knees with a Leave-One-Subject-Out (LOSO) Cross Validation accuracy of 96%.
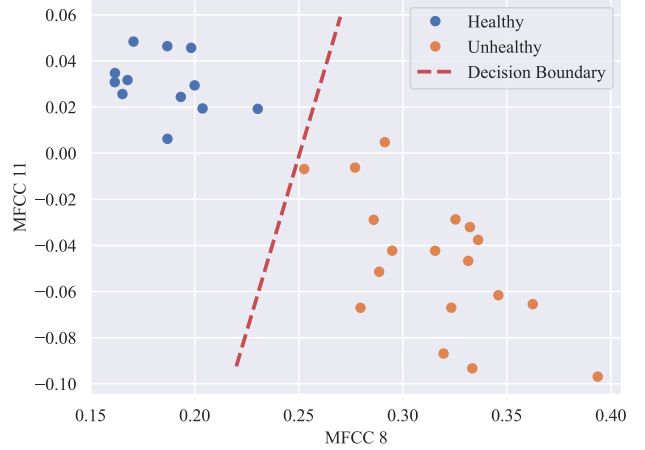

Fig. 3: Separation of Healthy / Unhealthy subsets based on MFCC8 and MFCC11.

In this experiment, by construction, we know that the audio features do not characterize biomechanical degradation. Yet, in the counterfactual case, the interpretation of the high classification accuracy led to (falsely) attributing the feature information to the knee mechanism.

Takeaway: High classification accuracy alone is not enough to establish knee health biomarkers.

Conclusions such as the one made in the counterfactual scenario fail to take into consideration the multi-source nature of the audio recordings: all the information contained in the signal has been attributed to the observed knees and their health status. However, referring back to Eq. 11, there are two terms that could potentially contribute to this classification: one whose source is the joint mechanism and another one which should be attributed to external sources, w.r.t. the joint sources.

Proper source attribution in Eq. 11 is crucial: by attributing all information to the knee we have managed to identify structural differences where we should not have. Thus, a causal exploration of the phenomena that affect the audio features identified as relevant biomarkers should not be considered as future work [23], [8], but rather a necessary step in the researchers' exploration.

Takeaway: Causal investigation on the audio features is imperative to claim them as biomarkers.

Such an investigation is not necessarily a trivial task. In the next two study-cases we present two examples of acoustical features causal study.

As a final note, two points need to be addressed. The first one is the population size which is admittedly small (5 subjects). Nevertheless, similar knee acoustic emission studies with comparable sample sizes exist in the literature [29], [20], [30], [31], [25], [32], [27]. Furthermore, a larger population size could help address the issue in this particular case: it decreases the probability that $p(Y \mid X, O_i)$ is informative for external sources. However, it does not necessarily guarantee lack of bias in the data. As we demonstrate in the next section a larger population size can still be affected by external sources.

Takeaway: Larger sample size helps, but cannot substitute proper causal investigation.

The second point is that the audio data remained the same in the counterfactual exploration with only the output - expectations changing. The counterargument here is that in the scenario with multiple subjects, Healthy and Unhealthy, other components related to the health variable would have been added in the audio features. This is also what the state of the art usually adopts as a direction. However it is crucial here to stress that this is just a hypothesis which needs to be validated. It should not be considered as prior information. What we have shown in this example is that this hypothesis is not necessary for the health classification accuracy to be high.

B. The Relevance of the Protocols

This case-study is a real-world case of a publicly available dataset [8]. 43 subjects were recruited for the study: 18 healthy and 25 with Juvenile Idiopathic Arthritis (JIA) [33]. Five of the 25 JIA subjects were not included in the public dataset. We replicate the workflow reported in the original publication:

1) Bandpass filter the audio signal in the range 250 Hz - 10 kHz.
2) Extract the most significant features, as reported in [8].
3) Train and validate a linear regression model using the LOSO strategy.

We achieved an average exercise-repetition accuracy of 51.28%, a significant drop from the reported 80.6%. To investigate further, we performed the following explorations: (i) manual inspection of the audio signals and (ii) per frequency-range classification.

Manual Inspection. We performed an audio-visual inspection of the recorded signals in the time-frequency domain. A constant frequency component centered around 33 kHz with a high prevalence was detected in the unhealthy population, a representative example is presented in Figure 4b. However, this component was not present in any of the healthy samples (Figure 4a), with the exception of one healthy subject. This signal is constant throughout the duration of the recording, even when the legs are inactive. Thus, we consider impossible the scenario that it is generated by the joint's internal mechanism. Rather, there has to be an external interference.

Takeaway: Thorough manual audio-visual inspection helps in the causal attribution investigation.

Per Frequency-Range Classification. We elaborate on the previous findings by determining the frequency range that is more informative for the knee's health status, i.e. it achieves higher classification accuracy. We split the available range 250 Hz - 50 kHz of the signal into 5 groups of 10 kHz. For each frequency range group, we follow the same procedure as in [8]: features are extracted for the corresponding frequency range, and a linear model is trained and validated using LOSO cross validation. Figure 4c presents the results: classification accuracy as a function of the frequency components in the inputs of the classifier. Average accuracy peaks in the 30 kHz - 40 kHz range at 77.9%, while for the rest of the groups it remains close to 50%.

As discussed in Section II, the Common Cause Principal is crucial in interpreting our results. The external constant interference, Z, causes the dependence between the classifier's output $g(\tilde{V})$ and the knee health status H. Since the knee cannot produce such an interference, Z is not the same as H. By filtering out the 33kHz component, effectively conditioning on Z, $\tilde{V}$ and H become independent. It is also important to note that the constraints resulting from Eq. 11, as discussed in Section II, are also violated. We can now summarize the Bias Introduction Pathway proposing solutions to breaking it. For the second sum of Eq. 11 describing the effect of external sources :

1) $p(S_j)$**:** There is obviously an external interference.
   Potential Solution: Remove the external interference from the experiment environment.
2) $p(O_i \mid S_j)$**:** The acoustic sensor observes this external interference.
   Potential Solution: Isolate the sensor and electronics from the detected interference. Depending on the nature of the external source filtering can also be performed post-hoc through digital filtering, if it is not affecting the actual knee components.
3) $p(Y \mid X, O_i)$**:** The external source is not random and provides information to the classifier about the health status of the knee.
   Potential Solution: Randomizing the observation of the external component w.r.t. the health status of the knee would prevent bias, even if external interference was present. For example, if it was present in half of the healthy and unhealthy populations and not present in the rest of the subjects.

Takeaway: The same environmental conditions should be guaranteed for the entire trial population.

Such inconsistencies may arise either from ambient noises, e.g. machinery present in the room, or from electronic component noise from within the signal acquisition setup, e.g. power source.

Regardless the solution, the first and most difficult step in tackling the bias issue is detecting and identifying it. In this case-study, manual inspection of the signals, using intuitive time-frequency representations was crucial in detecting the 33 kHz constant component. Equally important is observing its
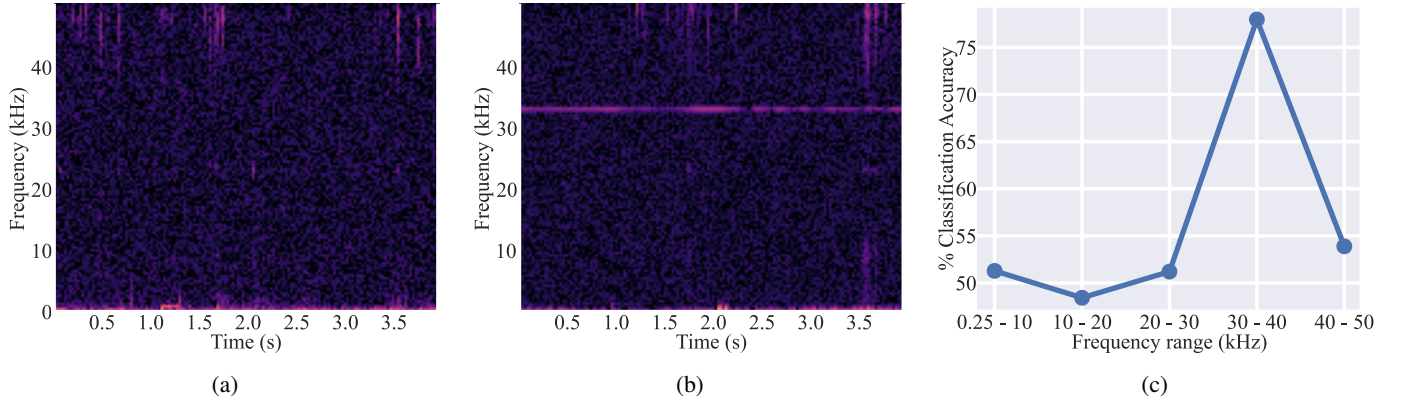
Fig. 4: External 33kHz interference and its effect on the health classification task. Representative time-frequency representations of the audio recordings of a Healthy **(a)** and an Unhealthy **(b)** individual. The constant component interference characterizing the unhealthy samples is visible at around 33kHz. **(c)** The accuracy of the knee health classification as a function of the frequency range demonstrating the effect of external interference on the classification task.

behavior and evaluating the possibility that it is generated by the knee mechanism. Such insights require prior knowledge on the experimental setting and on the biomechanical properties of the knee joint. It is difficult or even impossible to acquire this level of information just from the trained model alone.

Takeaway: Knowledge of the experimental conditions is necessary to attribute signal components to suspected sources.

C. The Relevance of the Sensors

In the third and final case study, we use a custom-made apparatus similar to a knee orthosis to collect acoustic information from the subpatellar medial compartment of the knee joint without hindering the movements of the patient. The apparatus consisted of a silicone-covered frame with embedded electronics to collect and store data from a non-contact acoustic sensor positioned at a constant distance from the area of interest. Institutional review board approval was obtained (BASEC-Nr.: 2020-01031) and all patients signed an informed consent prior to data acquisition. The population consists of 16 subjects undergoing for unilateral total knee replacement surgery: for half of the population, the left leg is the one to be operated, while for the rest half, the right knee. For the sake of consistency with the terminology used in the rest of the manuscript, we will label as *healthy* the non-surgical knees, and *unhealthy* surgical ones. However, it is important to note that we do not have specific information about the health status of the not-to-be-operated legs, and our objective is reduced to distinguish between groups. Patients performed eight unloaded and consecutive flexion-extension movements of the knee at a self-selected comfortable pace while wearing an apparatus on each leg, to record knee sounds. One device was always worn on the left knees, which we will refer to as the left device, while another one on the right ones, the right device. The classification task is thus defined as classifying the legs as to-be or not-to-be operated, depending on their acoustic emissions.

We performed a feature-based analysis similar to the two previous experiments. The acoustical signals were bandpassed in the frequency range 900Hz - 3kHz and then the acoustical features were extracted. We have observed that the family of MFCC features is one of the most relevant ones, so we designed a linear model to classify each knee as surgical or non-surgical based on the MFCC feature of its audio signals. With LOSO cross-validation we achieved a 75% classification accuracy. Neglecting Eq. 3 and 11 we can conclude that acoustic emissions can predict the knee undergoing surgery.

Just like in Sections III-A and III-B, we introduce our causal approach. First, we investigate whether the two devices themselves introduced bias. To this end we pose two questions:

1) Can we infer the device from the same audio features?
2) Does the knee health status classification performance change when conditioning on the device?

Then, we explore the mechanism - structures in the data allowing the separation between healthy and unhealthy knees.

To answer the first question, we train a classifier to infer the device based on the audio features. LOSO validation results in a device classification accuracy of 87.5%. Indeed, the same audio features can differentiate between left and right leg devices.

For the second question we condition on the device side (Figure 5) forming the following subsets from the original dataset:

1) One set consisting only of the left legs, D_L,
2) one only of the right ones, D_R and
3) a third sample as control, D_C, to account for the effect of sample size, since the sample size in each subset, D_L, D_R, is approximately halved compared to the original dataset. D_C is formed by randomly sampling 50% of the samples in the original dataset.

We train and validate (LOSO) a linear classifier from scratch for each of the three subsets: D_L, D_R, D_C. Since D_C is a random sampling of the original dataset, we repeat the experiment for D_C 10.000 times to account for variation introduced between different random splits. A different linear classifier is trained on each iteration.
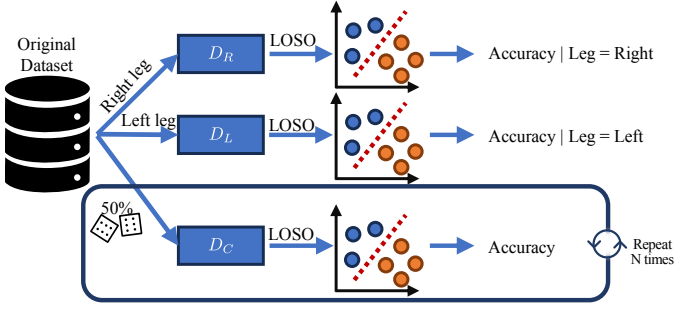
Fig. 5: Illustration of the process to assess the effect of the device on the health classification performance by conditioning on the device.

The results are presented in Figure 6a. We can observe that the effect of subsampling on the accuracy (blue distribution D_C, $67.15\% \pm 7.4\%$) does not explain the significant drop to 50% for D_R (green line).

The dependence of the accuracy on the existence of samples from the left leg is also shown in Figure 6b. Here, we begin with the subset D_R and incrementally add $1, 2, ...$ left leg samples. The left legs to be added are randomly selected and the experiment is repeated 500 times. As a point of reference, we also evaluate the classification accuracy if all legs samples had been drawn randomly but with the same total number of legs. Observe how a large percentage (higher than 40%) of samples from the left leg is necessary for the two distributions to start to converge.

We will now explore how the three structures of the data allow for the classification between healthy and unhealthy knees. For visualization purposes, we will only focus on two of the most informative MFCC features: MFCC8 and MFCC11 (Figure 8) which can achieve 73.75% classification accuracy.

Structure 1: Population The first structure is our population of 16 patients which is divided into two sub-groups:

1) 6 patients with left knee unhealthy and right knee healthy, D_{LU} and
2) 10 patients with left knee healthy and right knee unhealthy, D_{RU}.

There are no patients with both legs healthy or both legs unhealthy, breaking the independence between the samples: let H_{L_i} be the health status of the left leg of patient i and H_{R_i} the health condition of their right leg, then $p(H_{L_i} = Unhealthy) = 1 - p(H_{R_i} = Unhealthy)$.

Structure 2: Device Identification The second data structure is the separation between left and right knees. MFCC 8 differentiates between the left and the right device (Figure 7), MFCC 11 also presents the same separation but the effect is not so strong. As a result, two samples $x_i \in D_L, y_i \in D_R$ from the same patient, either D_{RU} or D_{LU}, tend to be positioned diametrically opposite in the MFCC8-MFCC11 feature map (Figure 8a). In fact, for both subsets the first principal component $(\mathbf{v}_{RU}, \mathbf{v}_{LU})$ distinguishes between the left and right devices, Figure 7b, while it is independent of the leg's health condition.

Structure 3: Distribution shift The third and final data
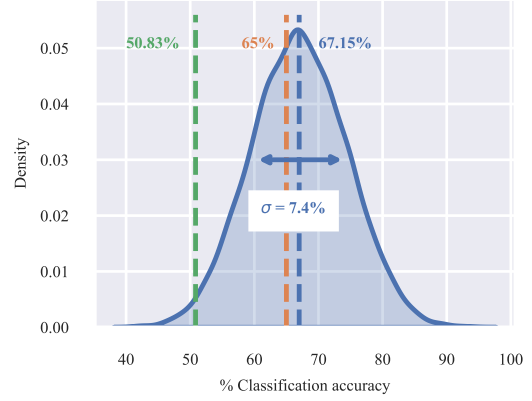
(a)
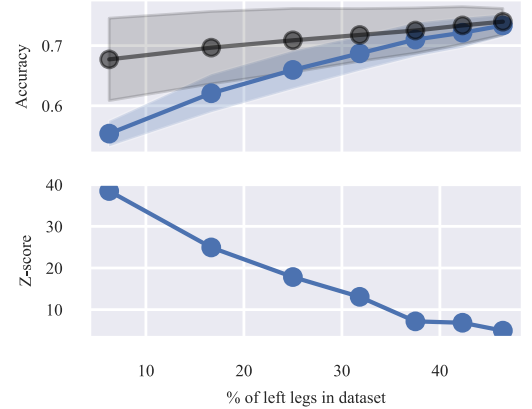
(b)

Fig. 6: Dependence of the classification accuracy on the device. **Top:** Conditioning on the right device the accuracy drops to 50% (green vertical line), a significant reduction that cannot be explained by the reduced sample size (blue distribution). Conditioning on the left device is presented with an orange vertical line. **Bottom:** Classification accuracy is dependent on the percentage of sample of the left leg in the dataset (blue). This dependence cannot be explained by the reduced amount of samples (black).

structure is the distribution shift between D_{LU} and D_{RU}. There is a rotation between $\mathbf{v}_{RU}$ and $\mathbf{v}_{LU}$ (Figure 8a): $\phi(\mathbf{v}_{RU}, \mathbf{v}_{LU}) = cos^{-1}\left(\frac{\mathbf{v}_{RU} \cdot \mathbf{v}_{LU}}{|\mathbf{v}_{RU}| \cdot |\mathbf{v}_{LU}|}\right)$. It is this misalignment, $\phi(\mathbf{v}_{RU}, \mathbf{v}_{LU})$ which allows the classification between healthy and unhealthy legs. Accentuating the shift between the two sub-populations, Figure 8b, results in a better separation and higher classification accuracy (80%). In contrast, aligning the two populations, $\phi = 0$, results in the inability to distinguish between healthy and unhealthy legs (Figure 8c). Interestingly, only a few degrees of distribution shift, $\phi \leq 10°$, can result in a significant inflation of the classification accuracy (Figure 9).

Interpreting the distribution shift, ϕ, is non-trivial. If the effect of knee health on the audio features is considered as prior knowledge, and given the underlying Structures 1 and 2, then ϕ is the mechanism through which the healthy and
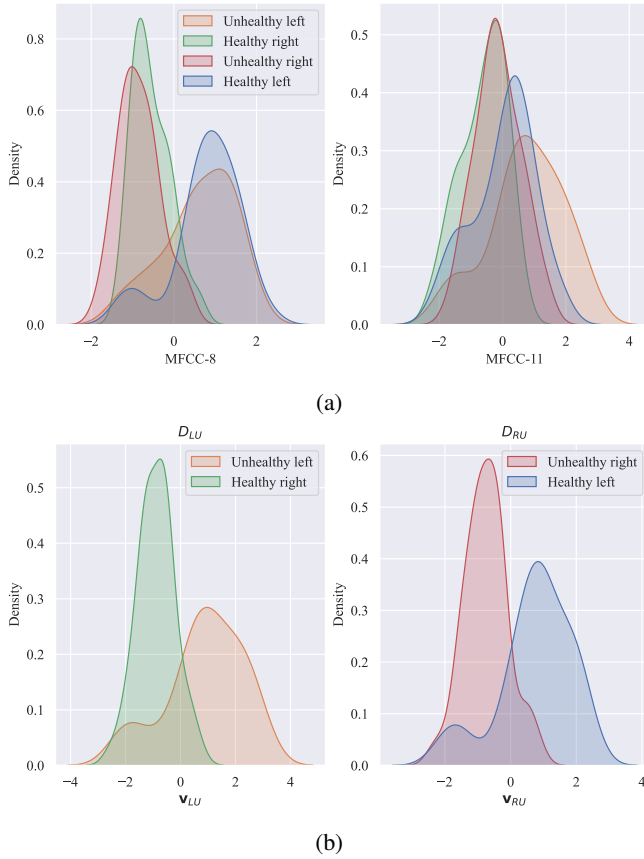
(a)

(b)

Fig. 7: Feature plots indicating separation between left and right leg. **Left:** Leg separation in the MFCC8 and MFCC11. The features do not differentiate between healthy and unhealthy legs: Healthy Right coincides with Unhealthy Right. **Right:** Distributions of features when projected along the first main principal component. Leg side separation is evident while health status is independent. For example, similarly to MFCC8 and MFCC11, Healthy Right overlaps with Unhealthy Right and Healthy Left overlaps with Unhealthy left.

unhealthy knees are separated. Still, this interpretation does not explain the lack of health status separation in the Right device, in contrast to the Left device: the effect of knee degradation should be independent of the device. In any case, in this scenario the results cannot be used as evidence in favor of the Main Hypothesis (Eq. 1), since it has been used as prior information. If the Main Hypothesis is not used as prior information, but rather as a hypothesis we test against, then there's a second possible interpretation of the results, i.e. ϕ is the result of variance between the two subpopulations. To reach a conclusive result further investigation would be needed. Just like in previous examples we can refer to Eq. 11 for possible directions:

1) $p(S_j)$: There is an interference generated by the device itself.
   Potential Solution: Study acoustically the two devices to understand its nature and possibly normalize between the two devices in the physical world.
2) $p(O_i \mid S_j)$: This device-specific interference is being

picked up by the microphones.
   Potential Solution: Understanding how it manifests in audio recordings could potentially help us filter it digitally.
3) $p(Y \mid X, O_i)$: External interference along with the data structures in our dataset allow for the significant inflation of the classification accuracy.
   Potential Solution: We propose three possible solutions:
   - Include patients with both legs healthy and both legs unhealthy
   - Expand study to include a larger population
   - Randomly shuffle the left and right devices between subjects

Our results indicate that the device indeed functioned as a bias source, inflating the performance of the knee health classification. Unlike the previous example, here the external interference observation was indirect, we studied the data structures in the dataset. Still, our extensive feature-based causal investigation was necessary since high accuracy alone was not enough. We demonstrated how data variances, reminiscent of the first example, can explain the separation between healthy and unhealthy samples. The sample size was also limited, sub-sampling half the population results in only 8 subjects, a sample size similar to the 5 subjects of the first example (Section III-A).

Unfortunately, we were unable to directly detect the audio component differentiating the devices despite thorough manual audio-visual inspection, in a similar fashion to the 33kHz component of Section III-B. We were also unable to manually identify audio events or patterns that discriminate between healthy and unhealthy knees. However, we managed to manually distinguish *click* sounds which could potentially be attributed to the knee mechanism.

Similarly to Section III-B all three conditions of Eq. 11 were violated. In the experiment there was external interference ($p(S_j)$) which was captured by our experimental device ($p(O_i \mid S_j)$) and introduced bias into the classification task ($p(Y \mid X, O_i)$). We suspect differences between the left- and right-leg devices; the two devices were prototyped manually which could have introduced dissimilarities between them. For example, the physical interface between the microphone and the skin could present slight differences resulting in variations in the frictions on the skin. We can now extend our conclusion from Section III-B: external interference can be generated by both the environment and the recording device itself.

Takeaway: Condition on device-specific audio properties.

IV. CONCLUSIONS

In this work, we have introduced a comprehensive causal framework to the field of knee vibroarthrometry. Our tools allow us to perform a holistic investigation taking into account various possible sources of information, i.e. the knee mechanism as well as external interference, expanding the current state-of-the-art, which considers only the knee mechanism as the information source.
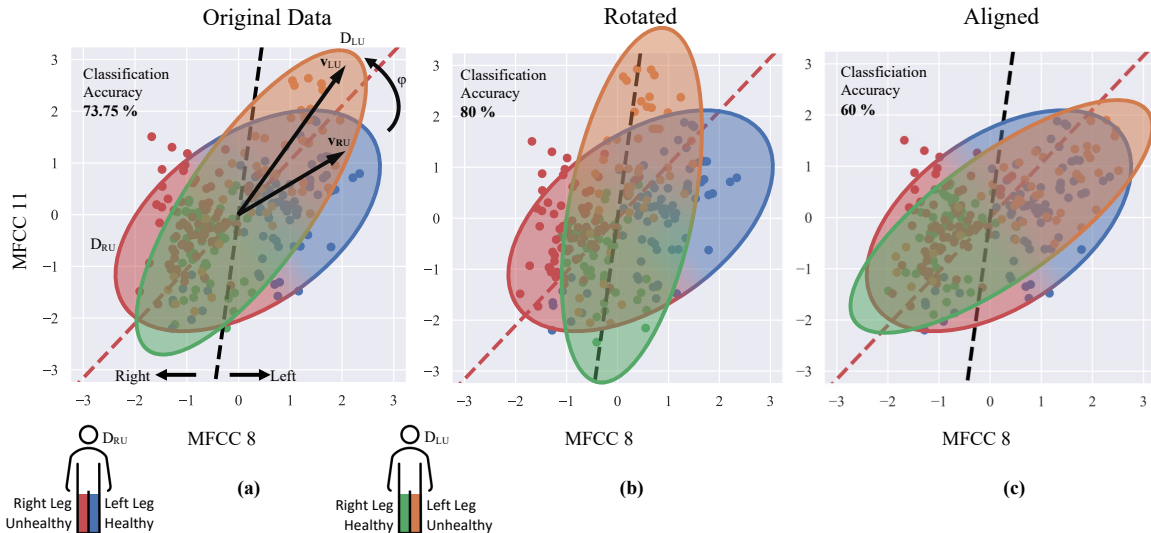
Fig. 8: Normalized MFCC8 - MFCC11 feature map. The colored ellipses visualize the two populations D_{LU} and D_{RU}. **(a)** The original data. **(b)** The samples of D_{LU} have been rotated counter-clockwise to accentuate the separation between the healthy and unhealthy samples. **(c)** D_{LU} has been rotated clockwise so that the first principal components of D_{LU} ($\mathbf{v}_{LU}$) and D_{RU} ($\mathbf{v}_{RU}$) align.
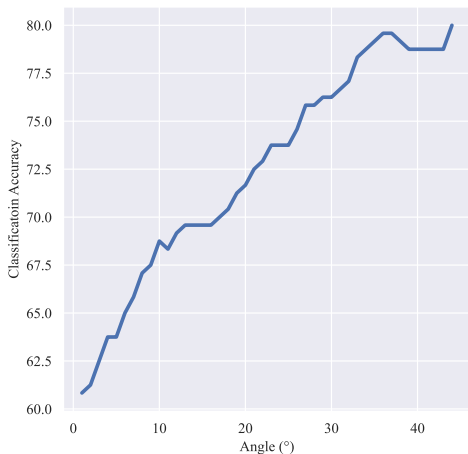

Fig. 9: The ability to distinguish between healthy/unhealthy legs is dependent on the distribution shift between the two sub-populations D_{LU} and D_{RU}. The classification accuracy is sensitive to the shift and just a few degrees of relative angle can inflate it.

We have applied our proposed methodology to three case studies. Our findings challenge the assumption that achieving high classification is sufficient evidence for acoustic knee biomarkers. Across the three experiments, we identified external sources of information that significantly inflated knee health classification performance, leading to a gross overestimation of the ability of the proposed biomarkers to diagnose knee degradation. In two out of three experiments, the proposed biomarkers did not characterize the knee mechanism. In the third one the experimental process was not enough to reach a conclusive interpretation of the result, while external biases were present.

These outcomes underscore the need for a comprehensive causal investigation of audio features to fully claim them as biomarkers. Meticulous audio-visual inspection of the audio recordings plays an important role in our proposed methodology, unveiling external interference and helping attribute observations to sources, as evidenced in the second example. Complementary to manual inspection, a thorough examination of underlying data structures provides insights into the possible external, wr.t. the knee, confounders.

Finally, we identified two categories of external sources that can contribute to the health classification task: environmental and device-specific. Therefore, maintaining consistent environmental conditions across the entire population is paramount. Similarly, a thorough examination of the wearable audio acquisition device is essential to ensure that it does not affect the classification task or, at least, its effect is homogeneous throughout the entire data set.

A remaining challenge to be considered is the labeling of the conditions to enhance the clinical relevance of any study. It is of utmost importance to define labels based on actual medical information (e.g., examination, imaging...) while considering that most clinical scoring is based on symptoms, which have a non-linear and non-continuous relation with disease severity. Thus, different approaches should be explored to assess the impact of specific symptoms on the acoustic emissions, and also the influence of the model properties in the label space. For example, a model assuming a smooth output or a sharp threshold between classes may not be appropriate for representing the clinical definition of the target labels.

REFERENCES

[1] C. N. Teague, S. Hersek, H. Töreyin, M. L. Millard-Stafford, M. L. Jones, G. F. Kogler, M. N. Sawka, and O. T. Inan, "Novel methods for sensing acoustical emissions from the knee for wearable joint health

assessment," *IEEE Transactions on Biomedical Engineering*, vol. 63, no. 8, pp. 1581–1590, 2016.

[2] W. E. Blodgett, "Auscultation of the knee joint," *The Boston Medical and Surgical Journal*, vol. 146, no. 3, pp. 63–66, 1902.

[3] G. F. McCoy, J. D. McCrea, D. E. Beverland, W. G. Kernohan, and R. Mollan, "Vibration arthrography as a diagnostic aid in diseases of the knee. a preliminary report," *The Journal of Bone & Joint Surgery British Volume*, vol. 69, no. 2, pp. 288–293, 1987.

[4] Y.-T. Zhang, C. B. Frank, R. M. Rangayyan, and G. D. Bell, "Mathematical modeling and spectrum analysis of the physiological patello-femoral pulse train produced by slow knee movement," *IEEE transactions on biomedical engineering*, vol. 39, no. 9, pp. 971–979, 1992.

[5] Y. Zhang and R. Rangayyan, "Adaptive cancellation of muscle contraction interference in vibroarthrographic signals," *IEEE transactions on biomedical engineering*, vol. 41, no. 2, pp. 181–191, 1994.

[6] L. Spain, B. Rajoub, D. K. Schlüter, J. C. Waterton, M. Bowes, L.-K. Shark, P. Diggle, and J. Goodacre, "Biomarkers for knee osteoarthritis: New technologies, new paradigms," *International Journal of Clinical Rheumatology*, vol. 10, no. 4, pp. 287–297, 2015.

[7] 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 Transactions on Biomedical Engineering*, vol. 47, no. 6, pp. 773–783, 2000.

[8] D. C. Whittingslow, J. Zia, S. Gharehbaghi, T. Gergely, L. A. Ponder, S. Prahalad, and O. T. Inan, "Knee acoustic emissions as a digital biomarker of disease status in juvenile idiopathic arthritis," *Frontiers in Digital Health*, vol. 2, p. 571839, 2020.

[9] S. Gharehbaghi, D. C. Whittingslow, L. A. Ponder, S. Prahalad, and O. T. Inan, "Acoustic emissions from loaded and unloaded knees to assess joint health in patients with juvenile idiopathic arthritis," *IEEE Journal of Biomedical and Health Informatics*, vol. 25, no. 9, pp. 3618–3626, 2021.

[10] S. C. Abbott and M. D. Cole, "Vibration arthrometry: a critical review," *Critical Reviews™ in Biomedical Engineering*, vol. 41, no. 3, 2013.

[11] J. N. Katz, K. R. Arant, and R. F. Loeser, "Diagnosis and treatment of hip and knee osteoarthritis: a review," *Jama*, vol. 325, no. 6, pp. 568–578, 2021.

[12] D. T. Felson, "Osteoarthritis of the knee," *New England Journal of Medicine*, vol. 354, no. 8, pp. 841–848, 2006.

[13] K. Talari and M. Goyal, "Retrospective studies–utility and caveats," *Journal of the Royal College of Physicians of Edinburgh*, vol. 50, no. 4, pp. 398–402, 2020.

[14] P. Ball, "Is ai leading to a reproducibility crisis in science?" *Nature*, vol. 624, no. 7990, pp. 22–25, 2023.

[15] W. Ge, C. Lueck, H. Suominen, and D. Apthorp, "Has machine learning over-promised in healthcare?: A critical analysis and a proposal for improved evaluation, with evidence from parkinson's disease," *Artificial Intelligence in Medicine*, vol. 139, p. 102524, 2023.

[16] A. J. DeGrave, J. D. Janizek, and S.-I. Lee, "Ai for radiographic covid-19 detection selects shortcuts over signal," *Nature Machine Intelligence*, vol. 3, no. 7, pp. 610–619, 2021.

[17] J. D. Arias-Londoño and J. I. Godino-Llorente, "Analysis of the clever hans effect in covid-19 detection using chest x-ray images and bayesian deep learning," *Biomedical Signal Processing and Control*, vol. 90, p. 105831, 2024.

[18] D. Wallis and I. Buvat, "Clever hans effect found in a widely used brain tumour mri dataset," *Medical Image Analysis*, vol. 77, p. 102368, 2022.

[19] W. R. Hendee and E. R. Ritenour, *Medical imaging physics*. John Wiley & Sons, 2003.

[20] D. C. Whittingslow, H.-K. Jeong, V. G. Ganti, N. J. Kirkpatrick, G. F. Kogler, and O. T. Inan, "Acoustic emissions as a non-invasive biomarker of the structural health of the knee," *Annals of biomedical engineering*, vol. 48, pp. 225–235, 2020.

[21] M. Galar, A. Fernández, E. Barrenechea, H. Bustince, and F. Herrera, "An overview of ensemble methods for binary classifiers in multi-class problems: Experimental study on one-vs-one and one-vs-all schemes," *Pattern Recognition*, vol. 44, no. 8, pp. 1761–1776, 2011.

[22] 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," *Medical engineering & physics*, vol. 31, no. 8, pp. 1013–1022, 2009.

[23] N. Befrui, J. Elsner, A. Flesser, J. Huvanandana, O. Jarrousse, T. N. Le, M. Müller, W. H. Schulze, S. Taing, and S. Weidert, "Vibroarthrography for early detection of knee osteoarthritis using normalized frequency features," *Medical & biological engineering & computing*, vol. 56, pp. 1499–1514, 2018.

[24] J. Peters, D. Janzing, and B. Schölkopf, *Elements of causal inference: foundations and learning algorithms*. The MIT Press, 2017.

[25] H. K. Jeong, D. Whittingslow, and O. T. Inan, "B-value: A potential biomarker for assessing knee-joint health using acoustical emission sensing," *IEEE sensors letters*, vol. 2, no. 4, pp. 1–4, 2018.

[26] S. Hersek, M. B. Pouyan, C. N. Teague, M. N. Sawka, M. L. Millard-Stafford, G. F. Kogler, P. Wolkoff, and O. T. Inan, "Acoustical emission analysis by unsupervised graph mining: A novel biomarker of knee health status," *IEEE Transactions on Biomedical Engineering*, vol. 65, no. 6, pp. 1291–1300, 2017.

[27] G. C. Ozmen, M. Safaei, L. Lan, and O. T. Inan, "A novel accelerometer mounting method for sensing performance improvement in acoustic measurements from the knee," *Journal of Vibration and Acoustics*, vol. 143, no. 3, p. 031006, 2021.

[28] Z. K. Abdul and A. K. Al-Talabani, "Mel frequency cepstral coefficient and its applications: A review," *IEEE Access*, vol. 10, pp. 122136–122158, 2022.

[29] T. Khan, S. Idei, and T. Yoshimura, "Acoustic emission measurement as adaptive biomarker in integrity analysis of knee osteoarthritis," *J. Multimodal User Interface*, vol. 10, pp. 319–324, 2016.

[30] L.-K. Shark, W. Quan, M. A. Bowes, J. C. Waterton, and J. Goodacre, "Discovering associations between acoustic emission and magnetic resonance imaging biomarkers from 10 osteoarthritic knees," *IEEE Transactions on Biomedical Engineering*, vol. 69, no. 11, pp. 3494–3503, 2022.

[31] L. Khokhlova, D.-S. Komaris, S. Tedesco, and B. O'Flynn, "Motion artifact resistant mounting of acoustic emission sensors for knee joint monitoring," in *2021 43rd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*. IEEE, 2021, pp. 7300–7303.

[32] ——, "Test-retest reliability of acoustic emission sensing of the knee during physical tasks," *Sensors*, vol. 22, no. 23, p. 9027, 2022.

[33] L.-S. Ording Muller, P. Humphries, and K. Rosendahl, "The joints in juvenile idiopathic arthritis," *Insights into Imaging*, vol. 6, pp. 275–284, 2015.