A Methodology for Developing Quality of Information Metrics for Body Sensor Design

Italo Armenti, Philip Asare, Juliana Su, John Lach
1Charles L. Brown Department of Electrical and Computer Engineering
2Department of Computer Science
3UVA Center for Wireless Health
University of Virginia, Charlottesville, VA USA 22904
{ia3he, pka6qz, js8dx, jlach}@virginia.edu

ABSTRACT

Body sensors networks (BSNs) are emerging technologies that are enabling long-term, continuous, remote monitoring of physiologic and biokinematic information for various medical applications. Because of the varying computational, storage, and communication capabilities of different components in the BSN, system designers must make design choices that trade off information quality with resource consumption and system battery lifetime. Given these trade-offs, there is the possibility that the information presented to the health practitioner at the end point may deviate from what was originally sensed. In some cases, these deviations may cause a practitioner to make a different decision from what would have been made given the original data. Engineers working on such systems typically resort to traditional measures of data quality like RMSE; however, these metrics have been shown in many cases to not correlate well with the notions of information quality for the particular application. Objective metrics of information distortion and its effects on decision making are therefore necessary to help BSN designers make more informed trade-offs between design constraints and information quality and to help practitioners understand the kind of information being produced by BSNs, on which they have to base decisions. In this paper, we present a general methodology for developing such metrics for various BSN applications, illustrate how this methodology can be applied to a real application through a case study, and discuss issues with developing such metrics.

Categories and Subject Descriptors
J.3 [Life and Medical Sciences]: Health, Medical information systems.

General Terms
Measurement, Performance, Reliability

Keywords
Body sensor networks (BSNs), Quality of Information (QoI),

1. INTRODUCTION

Medical practitioners face important decisions regarding patient care on a daily basis, and the information from various monitoring systems plays a significant role in this decision making process. Body sensor networks (BSNs) are one form of such monitoring systems which enable remote monitoring of patients over longer periods of time than is usually possible in a clinical setting. These benefits, however, come with many challenges. One such challenge is balancing the need to provide the best possible information, which usually requires use of more computation, storage, and communication resources, with the need to provide longer monitoring periods (BSN lifetimes), which is typically achieved by minimizing resource usage.

In a typical BSN used in monitoring applications, data is sensed (and possibly processed) on nodes which then transmit their information to an aggregator (usually a cell phone-like device). Both devices are usually on the patient. There is also a base station (usually in a remote location) where the information from the aggregator is sent to be possibly further processed and then presented to the health practitioner. The practitioner must then make decisions based on information received at the base station. In a single sensor system, the aggregator may be serving only as a way of relaying messages to the base station. In a multi-modal system, the aggregator may be playing a larger role and may also be involved in some of the information processing necessary for the application.

Nodes in a BSN are typically resource-constrained and have limits on the amount and kinds of computation that can be performed, the amount of memory to store data, the amount of data that can be communicated within a specific time, and the amount of energy available to support operation. In addition, the sensors used on the node even for the same physiologic signal may vary in sensing range, sensitivity, sensor noise, and robustness to environmentally-induced artifacts. The aggregator usually has better computational, storage, and communication capabilities than nodes but still with some limits. The base station usually has the most computational and storage capabilities. This variation in capabilities among components of the system necessitates various design choices in order to meet application goals of getting the best possible information to the practitioner while allowing long periods of monitoring. For example, if large amounts of data must be stored over time, then nodes may compress this data in a lossy manner to save memory. Lossy compression may also be done to reduce the amount of data...
for communication since communication typically requires a lot of energy. The form factor requirements for many body sensors limit the battery size and hence the available energy for operation. Depending on the choices made along the way (from node to the base station), the information presented to the practitioner may be distorted.

Figure 1 shows a conceptual view of a typical BSN information flow, processing pipeline and some of the possible sources of information ‘distortion’. The physiological signal $x$ is sensed at the node and is processed and transmitted along the way, resulting in the final signal (or in some cases information extracted from the original signal) $y$. There is the risk that the distortion in $y$ may result in the practitioner making a decision that is different from what he or she would have made based on the original signal, $x$. Some of these distortions may be intentionally introduced ($\varepsilon_i$), for example, in the case of lossy compression. Others may be unintentional ($\varepsilon_u$ and $\varepsilon_{\text{comm}}$), for example, when numerical errors are introduced into computations due to the limitation of computational and communication resources (using integer arithmetic instead of fixed or floating point, using approximations of more powerful algorithms, or when data is lost due to wireless transmission errors). It is therefore important for designers to understand how various choices and the distortions they result in (both intentional and unintentional) impact the medical decisions that are made based on the information received. In particular, what is needed are application-specific metrics that can relate system operating points to resulting distortions and those resulting distortions to an objective measure of their impact on practitioner decisions.

In this paper, we present a general methodology for developing these application-specific metrics, which we term quality of information (QoI) metrics, illustrate how it can be applied in real applications through a case study, and discuss some of the issues involved with developing and applying such metrics.

2. INFORMATION QUALITY METRICS

Traditional ways to measure the quality of a signal are either statistical (root mean square error (RMSE) (1)) and percentage RMS difference (PRD) (2)) or information theoretic (signal-to-noise ratios (SNR) (3)).

$$RMSE(x, \hat{x}) = \sqrt{\frac{\sum_{n=1}^{N}(x(n) - \hat{x}(n))^2}{N}}$$

$$PRD(x, \hat{x}) = \frac{\sum_{n=1}^{N}(x(n) - \hat{x}(n))^2}{\sum_{n=1}^{N}(x(n) - \bar{x})^2} \times 100\%$$

$$SNR(x, \hat{x}) = 10\log\left(\frac{\sum_{n=1}^{N}(x(n) - \bar{x})^2}{\sum_{n=1}^{N}(x(n) - \hat{x}(n))^2}\right)$$

These measures are essentially information-agnostic\(^1\), assigning equal weight to all the data in the signal, and are useful when close-to-perfect signal reconstruction is the goal. However, most medical decisions are based on particular features that can be extracted from the signal. For example, in [6], the authors look at 12 consecutive RR intervals in an ECG signal in order to diagnose atrial fibrillation, a photoplethysmogram from a pulse oximeter is usually processed to extract information like heart rate and blood oxygenation, and EEGs are also processed to determine many features of interest for various applications [11, 3, 10]. Even in non-medical settings, application quality is usually measured in an information-aware manner. For example, the algorithm used in MP3 compression of music files is a lossy compression algorithm that takes into account human psychoacoustics and eliminates portions of the data that would not be heard by humans to begin with. Therefore, it is important to measure the ‘distortion’ of medical information in an information-aware manner as well, considering the features that are important with regards to decision making. It may turn out that distortions that would be considered bad by traditional metrics have little impact on diagnostic features and hence the system designer can take advantage of this knowledge during system design and optimization.

There has been some work in this regard for traditional medical signals. In [12], the authors developed the weighted diagnostic distortion (WDD), a measure\(^2\) of how well ECG compression algorithms preserve those features that are relevant for diagnosis. The authors showed that this measure correlates better with the results from a mean opinion score (MOS) test with three cardiologists when compared to the traditional measures mentioned above. The MOS, however, only measures the cardiologists’ perception of how similar the features between the original and compressed signals are, and subjectively measures the impact of these deviations on diagnosis decisions.

There has been other work since the introduction of the WDD where authors either apply the measure to evaluate their algorithms [1], propose an alternative to the WDD [7], or consider a similar measure and compression techniques for a different physiologic signal [5, 4]. However, most of the work is focused on compression and signal reconstruction, is not necessarily explored in the context of BSNs, and looks at very general measures.

Our work is similar in spirit to that of the WDD and the works based on and inspired by it. However, we believe that good QoI measures must be application-specific and must be applicable to other kinds of processing beyond compression and signal reconstruction. This is because (a) for the same physiologic signal, different features may be important depending on the particular application; and (b) the BSN may be reporting only features extracted to the practitioner instead of information that can be used to reconstruct the original signal.

Despite our insistence on the development of application-\(^1\)By information, we mean those features embedded in the signal in which one may be interested. These features are usually extracted by processing the raw data (signal) in some way.

\(^2\)We use the word measure and metric interchangeably throughout the paper.
specific measures, we believe such measures can be developed using a general methodology. Beyond the development of this methodology, we are also interested in determining these QoI metrics for different applications of interest and developing a framework for applying these metrics in profiling various BSN design points to enable more informed design choices with regards to meeting QoI requirements.

3. METHODOLOGY FOR DETERMINING

QoI METRICS

In this section, we propose a methodology for offline determination of the quality of information delivered by a BSN. In order to develop these measures of information quality, we introduce the notion of ‘truthfulness’ of information. The ‘truthfulness’ of a particular piece of information is a measure of how well it is representative of what was originally sensed (observed) by the system. The more ‘truthful’ the information, the less likely that a wrong decision made by a practitioner is due to ‘bad’ information provided by the system.

Of course, perfect ‘truthfulness’ is not always necessary and there is a tolerance range within which a practitioner may make the same decision even if the data deviates somewhat from what was originally observed. For example, we would call the information provided by the BSN in Figure 2(b) as ‘truthful’ since the only features of interest are the amplitudes ($a_1$ and $a_2$) of and distance ($t_{12}$) between the peaks in the signal in Figure 2(a). So long as these values are within the accepted error margin, we could assign the signal a high quality score.

![Figure 2: Example physiologic signal. (a) Original signal. (b) Reported signal by BSN. $a_1$, $a_2$, and $t_{12}$ are the features of interest.](image)

The above figure also shows the need for feature-aware metrics. Traditional metrics would typically measure signal reconstruction accuracy and would assign the ‘signal’ in Figure 2(b) a bad value, whereas in terms of preserving features that signal is of a high quality. One could argue that if the features are generated in a sequence form, the traditional metrics could still apply (for example, taking the RMSE of extracted RR intervals in an ECG signal). This is not necessarily true. Traditional metrics assume all data points exist in the transformed signal, however in a case like the RR interval extraction example, the algorithm may fail to detect some RR peaks and cause particular points to be omitted. A traditional metric would fail miserably in this case even for a single omission (as the wrong RR values would be compared after the omission). A feature-aware metric would account for omissions and compare the available points in the sequence and penalize omissions more appropriately.

Given the conditions above, the aim of a QoI metric is to assign a ‘truthfulness’ score to a system design point which indicates the likelihood that the information produced by the system at that operating point will (not) mislead the practitioner. A number of things must be known before such a measure can be developed for a particular application:

1. **Expected information**: the information expected from the BSN (the raw signal or features extracted from it)
2. **Diagnostic analysis**: the kind of analysis to be done on the information in order to make a diagnostic decision
3. **System information processing blocks**: the major processing blocks of the BSN which affect the final form of the information

Given this information, one would first evaluate each processing block separately to determine its information preservation properties in relation to design knobs. In addition, the designer must be able to evaluate the impact of the final output of any design point on diagnosis. This evaluation requires what we term a diagnostic distortion measure. Both the raw and diagnostic distortion measures require real data. A lighter weight (and approximate) evaluation, which is necessary when the design space is large, would require what we term a QoI predictor. Developing a QoI predictor requires information from block evaluations and the diagnostic distortion measure values for a subset of possible design points. We elaborate on these evaluations and measures below.

**Block Evaluation.** The block evaluation of a particular processing block evaluates how well the processing block preserves (or extracts) the features that are needed by the next stage in the processing pipeline, given as input the best possible output from the previous processing block. This measure is obtained by comparing the output of a processing block to what would be produced by the same block of an ideal system. Consider the atrial fibrillation detection application in [6]. The block evaluation of the RR extraction block would be how well RR intervals produced from that block match the ground truth RR data provided with the dataset. The input would be the sampled signal at the highest possible resolution of the intended system (even if there is a compression block before the RR extraction block). If there is a compression block before the RR extraction block, then the block evaluation of the compression block would be how well the RR intervals from the compressed signal match the RR intervals in the original signal. This evaluation is tolerance-agnostic as it only provides in absolute terms the ability of the particular processing block to preserve features when compared to the ideal and does not account for deviations that the next block in the pipeline can tolerate. However, it serves as a good first approximation of quality of information provided by specific blocks.

**Diagnostic Distortion Measure.** The diagnostic distortion measure requires care practitioner input. Here we assume that, given the right information, care practitioners interested in the particular application have an objective analysis technique that maps the results of the analysis to a particular diagnosis. Table 1 summarizes such a well-known technique for hypertension based on systolic (during heart beat) and diastolic (between heart beats) blood pressure measurements.

When a reasonable objective analysis technique is unavailable, a pseudo-objective technique can be produced by surveying care practitioners. In such a survey, care practitioners should be asked to make a diagnosis based on information typically provided by a system used in the particular application (without their knowledge of which ones are the original
and which ones are ‘distorted’). Each care practitioner must also indicate which features played a significant role in their final decision. The diagnoses based on the original information can be considered ground truth, and the set of features that played a significant role in those results can be considered the more important set of features. One can then measure the deviations of the ‘distorted’ from the original based on the survey information.

Given this analysis technique and mapping to a diagnosis, one can then pick various design points, and compare the diagnosis based on their output to the diagnosis based on the output of the ideal system. An illustration of such a comparison is shown in Figure 3.

<table>
<thead>
<tr>
<th>Condition</th>
<th>Systolic (mm Hg)</th>
<th>Diastolic (mm Hg)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Normal</td>
<td>&lt; 115</td>
<td>&lt; 75</td>
</tr>
<tr>
<td>Prehypertension</td>
<td>120 to 139</td>
<td>80 to 89</td>
</tr>
<tr>
<td>Stage 1 Hypertension</td>
<td>140 to 159</td>
<td>90 to 99</td>
</tr>
<tr>
<td>Stage 2 Hypertension</td>
<td>&gt; 160</td>
<td>&gt; 100</td>
</tr>
</tbody>
</table>

Figure 3: Illustration of comparison of diagnosis from ideal system out to diagnosis from various system design point outputs

The diagnosis index is the output of the objective analysis of the practitioner. In this illustration, we assume that there are two possible diagnoses: a particular condition is present or absent. There is a threshold value beyond which the practitioner concludes that the condition is present. In practice, the index may be more indicative of the likelihood of the presence or absence of the condition, but we consider the absolute case for simplicity. $I_0$ is the index from analysis of the output of the ideal system based on a single piece of data. $D_i$ is the index from analysis of the output of design point $i$ of the system on the same piece of data. For the purposes of this illustration the increasing index $i$ indicates a decreasing amount of resources for meeting application goals. One thing to note is that it may not be immediately obvious what the relationship between design points and their deviations from the ideal system are. As shown in the figure, $D_1$ deviates more than $D_2$ even though based on our design point definition $D_1$ is a superior design point compared to $D_2$ in that it provides more resources for the application. This evaluation of design points and their deviation from the ideal system must be done on many data samples to get a sense of how particular design points deviate in general from the ideal system.

Once this evaluation has been done for many sample points, one can then start to map specific design points to particular quality values based on their deviations from the ideal system. One such mapping is shown in Figure 4. Here, $D_i'$ represents the output of analysis of deviations of the system with design point $i$ from the ideal over a number of data points. Some measures that can be considered in this analysis are the mean and standard deviation of the deviations from the ideal over a number of data points. The diagnostic distortion measure is a function that takes in the outputs $D_i'$ and maps them to quality of information values.

The ideal system, $I_0'$, is always assigned the maximum possible quality. If we assume that the deviations in Figure 3 are representative of the general case, then $D_0'$ is assigned a quality close to maximum because even though its diagnosis index is in general different from $I_0'$, it produces a diagnosis similar to that of the ideal system each time. $D_2'$ is assigned a value much lower than $D_0'$ because its deviations put it closer to the diagnosis threshold. $D_1'$ and $D_3'$ are of course assigned bad quality numbers because their deviations result in a different diagnosis from that based on the output of the ideal system.

**QoI Predictor.** Of course, in complex designs, one cannot evaluate all possible design points over large datasets. Tackling this problem is analogous to tackling the testing problem in other fields like electronic design where test set sizes can be prohibitive. In our case, the solution is to combine the block evaluation and diagnostic distortion measure to run a few initial ‘tests’ in order to identify a rough analytic relationship between appropriate properties of the processing blocks (which make up the design point) and a diagnostic distortion value. This analytic relationship is the QoI predictor.

With this predictor, one can test a wider range of design points much easier than going through the full diagnostic distortion evaluation. Based on the output of the QoI predictor, one can then select a number of ‘promising’ design points to evaluate using the full diagnostic distortion evaluation. Going back to our RR extraction example where data is compressed before processing, the QoI predictor there would take in as input the compression parameters and the parameters of the RR extraction algorithm and produce as output the quality of information produced by that combination of system design parameters. Note that, usually, the block evaluations consider blocks in isolation. The QoI predictor accounts for how blocks affect each other when connected. In our example, the compression will affect the ability to identify RR peaks correctly and may also affect the actual values of the correctly identified RR intervals, and the QoI predictor must account for this.

4. **EXPLORING QoI METRICS IN A REAL APPLICATION: GAIT ANALYSIS**

In this section, we illustrate how our methodology can be applied in a real application through a case study based on gait analysis. The results are mainly for illustrative purposes as the aim is not to rigorously explore design points or de-
velop an effective solution for this application. We selected five data sets from 30 elderly residents of an assisted living facility who were monitored continuously for three hours each. The BSN consists of three inertial 6 degrees-of-freedom sensors (one wrist, one ankle, and the sacrum) which continuously log patient data. The practitioners were interested in identifying the segments of the data when patients were walking for further analysis. Although the practitioners have not identified the particular gait feature of interest, we assume, for the purposes of illustration, that they are interested in determining the cadence of the patients when they are walking in order to detect abnormal cadence.

Following our methodology, we determine the three pieces of information needed to evaluate our system for this application: (a) the practitioners expect the raw signals from walking segments only; (b) the practitioners are interested in extracting cadence in order to detect abnormalities in this gait attribute\(^3\); (c) the system must perform (i) classification to identify walking segments and (ii) compression to manage the data size within the system. With these pieces of information identified, we proceed to perform the various evaluations to determine the QoI of a number of system design points.

Note that we only illustrate the block evaluations and the development of the diagnostic distortion measure because these are essential for any QoI evaluation, and when the design space is reasonably small, design points to be evaluated using the diagnostic distortion measure can be selected by looking at data from the block evaluations. Also, developing the QoI predictor is a complex process that deserves its own separate treatment, and we plan to address that process properly in future work.

4.1 Block Evaluations

Here we evaluate the various blocks involved in information processing within our system (described below).

**Basic System Design.** The classification of data segments into walking and non-walking segments is done using data from the accelerometer on the sacrum. The sacrum was chosen because it provided the best classification results in previous experiments compared to other locations for our particular classifier. The classifier computes the standard deviation of the vector magnitude (assuming two or more axes) of the data from different axes over two-second epochs with a one-second overlap sliding window. It then uses thresholding and window-to-window comparison to classify each epoch as walking, not walking or a transition between the two possible states. The cadence is evaluated based on the Z-axis data from the gyroscope of the left ankle. For the block evaluations, we assume that the data is compressed before it is classified (at the base station). A block diagram of this processing configuration is shown in Figure 5. We explore alternative processing configurations (e.g. on node classification and classification before compression illustrated by the block diagram in Figure 6) when looking at the diagnostic distortion measure.

**Classification.** The main function of our system is to classify the data into walking and non-walking segments. In order to evaluate this block in our system, we need to determine a measure of truthfulness of the classifier. Given that our event of interest (whether the patient is walking) is currently happening, the classifier can produce one of two outputs: \(h\), indicating that the event happening (true positive), with probability \(P(h|h) = x_1\) and \(h\) indicating that the event is not happening (false negative), with probability \(P(h|h) = x_2\). Given that the event is not happening, the classifier output could be \(h\), with probability \(P(h|h) = x'_1\) and \(h\) with probability \(P(h|h) = x'_2\). Also, we define \(x_1 + x_2 = 1\) and \(x'_1 + x'_2 = 1\).

Looking at the definitions carefully, one can see that a measure of truthfulness is a function of \(x_1\) and \(x'_2\) since both measure how often the system is telling the truth about the event (and both are related to the other two directly). In addition, one will notice that \(x_1 = \frac{TP}{TN+FP}\) is the sensitivity of the classifier, and that \(x'_2 = \frac{TN}{TN+FP}\) is the specificity. Hence, we can deduce that a general measure of ‘truthfulness’ of a binary classifier is a function both its sensitivity and its specificity and only those two. The ideal system must have \(x_1 = x'_2 = 1\). We did not vary any knobs in the classifier, however, with this measure we can evaluate the effect of other design knobs on classification. We term the value of the sensitivity and specificity on the 0 to 1 scale the classification power.

**Sensor Choice.** We evaluated the effect of the choice of the number of axes of the sacrum accelerometer on classification power. Fewer axes requires fewer resources to acquire and process data. The ability to use fewer axes would therefore allow for longer monitoring periods due to the reduced resource consumption. The results (in Figure 7) show that using two out of the three axes generally does not affect classification power.

**Compression.** In order to manage the large amount of data, reduce the amount of communication needed to get data to the practitioner, and increase the lifetime of the BSN and hence monitoring periods, the data may be adaptively compressed using techniques like those explored in [2]. This has two effects. First, it may affect the classification results, and cause walking blocks to be discarded and non-walking blocks...
to be introduced into the data presented to the practitioners for analysis. Second, the blocks for data analysis may be ‘distorted’ in ways that affect the analysis results. Here, we mainly consider the ability of the compression algorithms to preserve the features used in classification (which in this case is the standard deviation of the two epochs). The second effect is accounted for in the QoI evaluation using the diagnostic distortion measure.

We considered two algorithms: (1) an implementation of adaptive differential pulse-code modulation (ADPCM) [8] and (2) decimation with shift quantization (RateRes). Both algorithms use 16-bit integer operations since they are designed to run on microcontrollers. The output of the ideal block here is the features from the uncompressed data. Our evaluation metric ($q_{\text{comp}}$) was the mean ($\mu$) and standard deviation ($\sigma$) of the absolute difference between features from the ideal and from compression over the dataset (4).

$$q_{\text{comp}} = [\mu(|\hat{g}(x) - g(x)|), \sigma(|\hat{g}(x) - g(x)|)]$$  \hspace{1cm} (4)

Here, $\hat{g}(\cdot)$ is the standard deviation over the two-second epochs of the particular data. We chose these metrics since they are analogous to accuracy ($\mu$) and precision ($\sigma$). We could have considered other statistical measures as well. A good compression algorithm has small values for $\mu$ and $\sigma$ since these represent small deviation from the ideal. The result of this evaluation (using the whole data set) for the two algorithms are summarized in Table 2.

### Table 2: Block Evaluation Results for Compression Algorithms

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>$\mu$</th>
<th>$\sigma$</th>
</tr>
</thead>
<tbody>
<tr>
<td>ADPCM</td>
<td>0.0212</td>
<td>0.055</td>
</tr>
<tr>
<td>RateRes</td>
<td>1.5183</td>
<td>0.1904</td>
</tr>
</tbody>
</table>

We also considered the effect of the algorithms on classification power. In a more complex design such an evaluation may not be possible because of the large design space; however, the evaluation here helps us confirm what the block evaluation for the compression block tells us (i.e. ADPCM will affect classification less because it preserves the features for classification better than RateRes). The results of this evaluation are shown in Figure 8.

### 4.2 Diagnostic Distortion Measure

Since the diagnosis was based on the cadence, we computed the average cadence value from the output of the system for each point in the data set and compared to that based on the ground truth data (representing the ideal sys-

![Figure 7: Effect of sensor choice at the sacrum on classification power](image1)

Figure 7: Effect of sensor choice at the sacrum on classification power

![Figure 8: Effect of compression on classification power](image2)

Figure 8: Effect of compression on classification power

tem, $I_0$). The choice of system design points are shown in Table 3.

### Table 3: System design points

<table>
<thead>
<tr>
<th>Design point</th>
<th>processing path</th>
</tr>
</thead>
<tbody>
<tr>
<td>$I_0$</td>
<td>perfect classification → highest resolution data</td>
</tr>
<tr>
<td>$D_0$</td>
<td>perfect classification → ADPCM compressed data</td>
</tr>
<tr>
<td>$D_1$</td>
<td>perfect classification → RateRes compressed data</td>
</tr>
<tr>
<td>$D_2$</td>
<td>base station classifier → high resolution data</td>
</tr>
<tr>
<td>$D_3$</td>
<td>good classification → ADPCM compressed data</td>
</tr>
<tr>
<td>$D_4$</td>
<td>ADPCM classification → ADPCM compressed data</td>
</tr>
<tr>
<td>$D_5$</td>
<td>good classification → RateRes compressed data</td>
</tr>
<tr>
<td>$D_6$</td>
<td>RateRes classification → RateRes compressed data</td>
</tr>
</tbody>
</table>

Table Notes: **Perfect classification** refers to the ground truth walking segments, **highest resolution data** refers to the non-compressed data, **base station classifier** refers to classification done at the base station on non-compressed data, **good classification** refers to classification done on node based on uncompressed data (implying that only what are identified as walking segments are compressed). **ADPCM classification** refers to classification done at the base station based on data compressed with ADPCM, and the same applies to **RateRes classification**.

For our diagnostic distortion measure, we computed again the average ($\mu$) and standard deviation ($\sigma$) of the absolute difference between the average cadence values of the system design point and the ideal system over the whole data set. Note that $\mu$ and $\sigma$ here are applied to a different type of data than what was considered in the compression block evaluation. We also counted the number of times ($V$) the diagnosis based on the value from the system design point differed from that based on the value from the ideal system over the data set. The results from this evaluation are shown in Table 4. Again, note that we could have chosen other statistical measures beyond $\mu$ and $\sigma$.

We combined the three different measures into a single metric $Q_{\text{gait}}$ shown in Equation 5 in order to map the design points to particular QoI values. The $\omega$ are weighting factors.

$$Q_{\text{gait}} = \max\{0, \omega_\mu e^{-\mu} + \omega_\sigma e^{-\sigma} - \omega V\}$$  \hspace{1cm} (5)

This equation is one of many possible forms that the diagnostic distortion (QoI) measure can take. In particular, we selected this equation so that $Q_{\text{gait}}$ lies between 0 and 1. The $e^{-\mu}$ and $e^{-\sigma}$ terms both evaluate to 1 when $\mu$ and $\sigma$ differ significantly from their respective averages, while the $V$ term ensures that the overall value increases with the number of errors.
We computed the Pearson Product-Moment Correlations \( r \) between the different measures of quality and the observed changes in classification power and the results are summarized in Table 5.

### Table 5: Correlation between measures of compression quality and effect on classification

<table>
<thead>
<tr>
<th>Quality Measure</th>
<th>( r ) (sensitivity)</th>
<th>( r ) (specificity)</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \mu(\cdot) )</td>
<td>0.59</td>
<td>0.90</td>
</tr>
<tr>
<td>( \sigma(\cdot) )</td>
<td>0.86</td>
<td>0.86</td>
</tr>
<tr>
<td>( F )</td>
<td>-0.36</td>
<td>0.80</td>
</tr>
<tr>
<td>normalized SNR</td>
<td>0.78</td>
<td>0.82</td>
</tr>
</tbody>
</table>

Our feature-aware block evaluation measure is a vector containing \( \mu(\cdot) \) and \( \sigma(\cdot) \) derived from (4), \( F \) is the RMSE-based measure (6), and normalized SNR is the normalized signal-to-noise ratio (to reduce the dynamic range). Based on our block evaluation measure definition we expect a positive correlation between our measure and changes in classification power if a correlation exist. This is because, for our measure, smaller values indicate higher qualities, and smaller changes in classification power means that the compressed data has little effect on classification. We expect the traditional measures \( F \) and normalized SNR to show a negative correlation because higher numbers there indicate higher qualities. In general our measures show moderate to strong correlations with the observed changes in classification power. The RMSE-based measure also shows moderate to strong correlations, however, in each case our measures show a stronger correlation compared to the RMSE-based measure. We acknowledge that more investigations are necessary before we can make any conclusions, but the current results show our feature-aware block evaluation measure to be better than the other measures considered.

### 5. DISCUSSION

Our preliminary results from the above case study show that feature-aware metrics are most useful for determining the impact of information produced by a system at different design points on diagnosis. Such metrics are easier to develop for traditional physiologic signals since the features of interest for various diagnoses have been thoroughly investigated over the years. For emerging sensing modalities like inertial sensing, there is still much work to be done to understand features that can be extracted from such signals (which can not be easily observed) and their relation to various gait attributes which are important for certain diagnoses. Our methodology for developing metrics should work well in this case as well (as shown by the application to the gait case study), and could even aid in the process of identifying useful features.

One main issue in developing such metrics is in the choice of features on which to base the metrics. The analysis done by the practitioner may involve processing the extracted features in ways that produce intermediate ‘features’ which are then used to develop the final diagnostic feature used in making decisions. One could use these intermediate features or the features passed to the analysis stage, and more work needs to be done for various applications to determine which choice of features factor well into developing quality metrics.

### 6. CONCLUSION

In addition to the evaluation of the QoI for this application, we briefly explored how our feature-aware block evaluation for compression using \( q_{\text{comp}} \) (4) compared to the traditional metrics in terms of ‘predicting’ the effects of compression on classification power since the signal would be reconstructed in order to perform classification. The traditional normalized SNR, an RMSE-based metric from Barth et al. [2] show in Equation (6).

\[
F = 2 \frac{\text{RMSE}(x, \hat{x})}{\sigma} \tag{6}
\]

The equation is a data fidelity measure designed to produce a number between 0 and 1. The factor \( b \) is set to 400 so that the 50% fidelity mark is when the RMSE is 400 [2].
We presented a methodology for developing QoI metrics and illustrated how it can be used in applications through a case study. However, much work still remains. We must still explore how this methodology can be effectively incorporated into the system design process to inform design choices. The design points considered in our gait application, for example, though reasonable, were not determined through a rigorous process accounting for other application requirements. In practice, many aspects of the application in addition to the QoI requirement would factor into system design choices. We are currently looking at a number of applications where objective measures for diagnosis are well-defined as a starting point and developing more in-depth case studies for these applications than was done here.

Acknowledgements
The authors wish to thank Dr. Amy Papadopoulos and our other collaborators at AFrame Digital for their role in collecting the de-identified gait data used in this paper. The authors would also like to thank Shanshan Chen for providing the algorithm for extracting cadence from the inertial data. This work is supported in part by the National Science Foundation under grants ECCS-0901686 and CNS-1035771 and by the National Institute of Aging under SBIR award 1R43AG039176-01A1 with AFrame Digital.

7. REFERENCES


