# Human Activity Recognition Literature Review

[Introduction](#_7d507xfrstzw)

[Evolution of the Activity Recognition Pipeline](#_8eiuxwburx67)

[Objectives](#_clxcoxkozvot)

[Overview of Wearable Sensors](#_4248lkdvqe3w)

[Design Challenges](#_8i2k2yhg7n9g)

[References](#_53qhlgrx9ezs)

## **Introduction**

Previously marketed towards health-concerned consumers as fitness trackers, wearables have become a common tool in research and clinical studies. Wearables, which entail any device that can monitor mechanical and physiological states over time, have seen substantial growth since the 2010s when The Fitbit and Apple Watch--the most popular wearables for consumers--were released (Pew Research, 2020). According to multiple surveys, approximately one in five Americans currently own wearable devices (Gallup, 2019; Pew Research, 2020). These wearable devices, and the data generated from individuals, has led to a large increase in growth in the human activity recognition (HAR) field.

**What is Human Activity Recognition?**

Human activity recognition is the automatic recognition of physical activities through wearable devices that monitor mechanical and physiological states over time (Dunn, 2018). The information that these wearables provide has an astounding variety of applications.

**Wearable Healthcare Applications**

First, utilizing wearables for healthcare enables long-term and longitudinal care outside healthcare facilities. The portable nature of both the wearables themselves and concurrent analytics platforms allow this technology to be used in rural or low-resource areas. (Dunn, 2018). For example, wearables have provided continuous remote monitoring of cardiovascular diseases that could greatly benefit users in rural areas with low access to healthcare facilities (Dilraj, 2015). Hospitals and clinics have greatly benefited from wearable sensors as well. Wearable fitness trackers can be used in an ICU to detect heart arrhythmias without bulky equipment (Kroll, 2017).

Wearables have also shown promise in assisting in the diagnosis of neurodegenerative diseases (ND) such as ALS, Huntington’s, and Parkinson’s. A common feature of these diseases is that they manifest in gait abnormalities due to degeneration of the cerebellum and/or spinal cord. For these individuals, cerebellar ataxia is generally unnoticeable until later in life as the ND progresses. Research has shown that, with the use of wearables, these abnormal gait patterns can be detected earlier and distinguished from each other, leading to earlier and more accurate diagnosis and treatment of these diseases (Zhao et al., 2018).

Similarly, wearables have made way for inexpensive, continuous long-term care that can provide preventative strategies or management for chronic illnesses. (Dunn, 2018). Novel ambient assisted living systems such as LiveNet and Verity that consist of a wearable accelerometer, skin thermometer, and machine learning-based platform, can track falls in the elderly population (Winkley, 2012; Sung, 2005). The Empatica E4 wearable was used to collect blood oxygen levels, skin temperature, heart rate, and electrodermal activity to detect migraine attacks (Kosimaki, 2017). Wearables can assist in the diagnosis of HAR predictive models that take wearables data and also have to offer personalized care. A precision medicine platform powered by a wearable accelerometer, gyroscope, and magnetometer presented by Qiu is helping healthcare providers counsel patients that suffer from gait (Qiu, 2015).

Finally, wearables can give a healthcare provider a deeper insight into patient behavior, or allow a user to manage his or her own behaviors better. A study showed that data from wearable sensors that tracked electrodermal activity and skin temperature was better at classifying the stress level of students than actual behavior features such as study and screen-time patterns (Sano, 2018). A drowsy driver detection algorithm, powered by a feed-forward neural network and a wearable biosensor that collects posture, heart rate, and breathing rate data can predict drowsy and awake states (Warwick, 2015). Long-term behavioral tracking can also help both wearables users and their healthcare providers to understand which behaviors need to change in managing or preventing chronic illnesses.

## **Evolution of the Activity Recognition Pipeline**

Most HAR applications to healthcare rely on correctly classifying health-related behavior, such as physical activity, from sensor data. The behavioral data is then used to inform treatment. Thus, the mechanism that translates sensor data to behavior is incredibly important.

Machine learning (ML), including deep learning (DL), techniques are obvious candidates for this mechanism. ML and DL have been used extensively for classification purposes, and more specifically within the HAR field. Naturally, there have been extensive research programs devoted to using these machine learning techniques to perform behavioral classification of sensor data.

When machine learning is used for HAR, development generally follows what Bulling, Blanke, and Schiele (2014) term the “Activity Recognition Chain” (ARC).

**Overview of the ARC**

The ARC generally follows the proceeding steps (Bulling et al., 2014). After data acquisition, the sensor data is preprocessed. The kind of preprocessing technique performed is dependent on the type of sensor data. Next, data is segmented into behaviorally relevant epochs. This process reduces the total amount of data the model has to process since activity-irrelevant information is discarded, though segmentation may not be perfect when humans have to estimate when the activity happens. Whether or not a deep learning model is selected will determine if feature engineering is required as a next step. Deep learning methods automatically discover representations of outputs as functions of inputs, or features (Bengio, 2009). These representations can then be used to infer outputs from new instances of those inputs. If a traditional machine learning method (non-deep learning) is selected, then one will have to engineer features of the data. The model then takes instances of the hand-created features as inputs and attempts to find relationships between them and behavioral classifications. The end goal for both models is the same: To correctly classify a high percentage of unseen data. Various assessments, such as confusion matrices, are used to estimate model performance. Other techniques, like ensemble classifiers, can then improve model performance.

The following subsections provide in-depth detail of each step in the ARC and how choices within it can impact model performance and use.

**Data Acquisition**

Data acquisition architecture for HAR systems involves wearable sensors attached to the human body which communicate with an integration device (cell phone, laptop, etc.) which preprocesses sensor data and can send data to an application server for monitoring, visualization, and/or analysis. This data can be stored locally on the integration device or remotely on a server (Lara, 2013).

The choice of what kind(s) of sensor data to collect may impact model performance if some kinds of sensor data are more relevant to the desired activity than others. Additionally, models that are created that must take certain sensor inputs cannot generalize to use by others who do not have that kind of sensor data. Even when using the same sensor input(s), there are vast quality differences between data collected by one wearable or another. Differences still remain when the quality is similar, but collected by different kinds of wearables. For example, in Bent, et. al, measures of mean error of heart rate varied significantly between and within the consumer-grade and research-grade wearable sensors (2020). These differences may impact model performance. Data quality may also be sacrificed to make the design of the wearable optimal for its users. Some of these design features include visual aesthetics, weight, energy optimization, tightness/comfort, and privacy (Mukhopadhyay, 2010).

**Preprocessing**

Preprocessing may affect model performance, and the effect may not always be an improvement. For example, Lee, Cho, and Yoon (2017) demonstrate when total variation (TV) minimization is used to reduce noise in triaxial accelerometry data that model performance improves when using a Random Forest, but overall model performance becomes worse after TV minimization when using a Convolutional Neural Network (CNN).

HAR data is generally segmented into different temporal periods of interest which are likely to contain activities. This process is referred to as “spotting” or activity detection. While human-based segmentation may work for training data, for test data it is error-prone, inconsistent, and due to the amount of data present, it is not feasible for a large number of datasets. Commonly used computationally-based segmentation techniques for classification models are sliding window, energy-based segmentation, rest-position segmentation, and the use of external context or sensors. (Bulling et al., 2014). The sliding window approach uses a defined set window time length over the temporal data to extract segments. This approach disregards all information about the structure of the underlying data. Energy-based segmentation sets a threshold on the energy based on signals from the sensors to define activities. The one caveat with this method is that it makes the assumption that rest periods occur between activities, which enables segmentation borders to be produced (Bulling et al., 2014). While using a sliding-window approach appears to be the most popular method to segment activities in the HAR research space, an algorithm known as “greedy Gaussian segmentation” (GGS), which maximizes the likelihood of the data for a fixed number of segments, has been shown to perform just as well if not better. GGS assumes, that for each segment, that the mean and covariance are independent between segments and constants (Li et al., 2019).

**Feature Extraction**

Feature extraction has extensive implications for model performance and generalizability. Features are needed in order to distinguish instances of the data from one another so that they can be classified as relating to a specific behavior (Lara, 2013). There are two main techniques for feature extraction: hand-crafting features and automatic feature extraction. Manually creating features is time-intensive and requires domain knowledge of what the signals represent. While they are interpretable by humans, these features may be too specific to the activity they represent (Nweke et al., 2018) or data they come from (source). Automatic feature learning is accomplished using DL methods. These methods simultaneously extract features and build models (Wang et al., 2019), making less work for the human, but more for the machine. Automatically generated features have less generalization problems, the reasoning for which is explained in the following paragraph.

The nature of the features themselves is reflected by the architectures that use them. Deep architectures produce features that collectively reflect the multiple levels of abstraction of what they come to represent (Bengio, 2009), while manually created features only capture one aspect of the data. If the object of classification itself (activity in HAR applications) can be represented through multiple levels of abstraction, then features that capture these levels may better represent the object as a whole (Yang, 2015; Bengio, 2009). This reasoning could very well explain why features generated from deep architectures are more generalizable than handmade features. By capturing the essence of an object in its multiple abstractions, new instances of the object which have the same essence are easily classified and are not confused with other, similar objects. The essence is an understanding of the variations within an object throughout its hierarchy such that when variations occur, the object that has variations is still understood as a kind of that object (Bengio, 2009).

The previous reasoning is directly applicable to HAR. Human activities are made up of smaller movements, which in turn may be also made of smaller movements, creating a hierarchy of representation for the original activity (Yang, 2015). Just like a sentence can be broken down into phrases, or words, or letters, activities can be broken down into smaller and smaller movements. This hierarchical structure provides motivation for using deep learning techniques to produce features that will capture the hierarchy, and thus the activity. Capturing the essence of the activity addresses the interclass similarity and intraclass variability problems in HAR. Better understanding of an activity and its variations helps to make it distinct from other activities while not dividing it up into separate activities based on variations in how it’s performed. These problems are brought up by Bulling et al. (2014) and are thoroughly explained in a later subsection (Activity Classification Challenges).

The inadequacy of manually creating features for HAR should now be apparent. Engineered features have been likened to just a part of one layer in a deep architecture (Yang, 2015). Creating a vast amount of these features and organizing them in a way that captures the appropriate hierarchical representations seems like an impossible task, and is what motivates the use of deep learning in many HAR applications (Bengio, 2009), including our own. Additionally, attempting to hand make more complex features to compensate for this problem may inadvertently lead to overfitting (source?).

**Deep Learning Architecture**

Deep learning is not a catch-all term; there are several kinds of deep learning models that each have their own uses. Some of the prevalent, basic models include the multilayer perceptron (MLP), CNNs, and RNNs (including LSTMs and Gated Recurrent Units, or GRUs).

Many of the top-performing HAR models utilize convolutional neural network(s) (CNNs) for automatic feature extraction. CNNs are preferred for feature extraction compared to other traditional machine learning and deep learning methods because it can:

1. Learn the signal motifs of each activity independent of the others, building a strong foundation so that the model can generalize to less common or new activities. In other words, CNNs are domain-independent: the fact that the low-level motifs represent the basic building blocks of each activity means that the model can be adapted to new datasets (Yang, 2015; Singh, 2020);
2. Create features that show local salience of signals from each sensor or time-interval, which can then be built upon either hierarchically or longitudinally to get global features of each sensor or time interval. These feature vectors can be extracted at any level (Ma, 2019; Yao, 2017).

A study by Kim, et. al shows that CNNs outperform traditional machine learning methods, such as SVM, and feed-forward neural networks, in HAR classification on the same dataset because traditional ML methods require more hands-on feature extraction or engineering. Interestingly, singular RNNs (recurrent neural networks) outperform singular CNNs, because the RNN architecture is more suited to sequential data than CNNs. Thus, many of the top-performing models include some hybrid CNN-RNN structure to ensure automatic feature selection works best for temporal data (Kim, 2020).

RNNs are able to make use of sequential information by using “memory.” They perform the same task on every element of a sequence. Using their “memory,” RNNs are able to use what calculations have been performed previously to influence their output. This is done by passing the previous hidden state, to the next step of the sequence. In this regard, the hidden state acts as the “memory.” Bi-directional RNNs take this concept further and in addition to past data, use predicted future data to influence the output. Since RNNs have a short-term memory, long short-term memory (LSTM) and gated recurrent units (GRUs) can be used to extend this memory further, in addition to helping to solve the vanishing gradient problem common in RNNs. LSTM and GRUs are able to effectively capture important information while disregarding irrelevant information from past data. The LSTM uses four gates, the input gate, the input modulation gate, the output gate, and the forget gate. The GRU uses two gates, the forget gate and the update gate (Fu, 2016).

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Methods** | **Description** | **Strengths** | **Weaknesses** | **Recent HAR Applications** |
| Convolutional Neural Network (CNN) | Deep neural network with interconnected structure inspired by biological visual cortex | Widely implemented in deep learning with a lot of training strategies proposed. Automatically learn features from raw sensor data. Moreover, CNN is invariant to sensor data orientation and change in activity details. | Require large dataset and high number of hyper-parameter tuning to achieve optimal features. Maybe difficult to support effective on-board recognition of complex activity details. Different sensor modalities come with different domains, and treating sensors the same without distinction may degrade performance of classification. Not all timesteps contribute equally to the activity recognition task. | Predict relationship between exercises and sleep patterns, automatic pain recognition during strenuous sports activities, energy expenditure estimation and tracking of personal activities. |
| Recurrent Neural Network (RNN) | Neural network for modelling sequential time series data. Incorporate temporal layer to learn complex changes in data | Used to model time dependencies in data | Difficult to train and suffer from vanishing or exploding gradients. In the case of LSTM, requires too many parameter updates. Large parameter updates are challenging for real-time activity predictions. | Model temporal patterns in activity of daily living (ADL), progressive detection of activity levels, fall and heart failures in elderly. |
| Long Short Term Memory (LSTM) | Similar control flow as a RNN, differences are the operations within the LSTM’s cells, uses four gates, the input gate, the input modulation gate, the output gate, and the forget gate | Information from earlier time steps can make its way to later time steps, reducing effects of short-term memory | Require more memory to train. Dropout is much harder to implement. Sensitive to different random weight initializations. | Capture the spatial and temporal features from time series data from multiple activity recognition sensors |
| Gated Recurrent Units (GRUs) | Similar to LTSM but GRU’s only have two gates which are the forget gate and the update | Information from earlier steps can make its way to later time steps, reducing the effects of short-term memory. | The GRU is more computationally efficient than LTSM due to less gates but has similar performance. | Capture the spatial and temporal features from time series data from multiple activity recognition sensors |

(Nweke, 2018).

**Hybrid Models**

Current HAR architecture focuses on deep learning due to the complexity of the time-based and frequency-based features that multimodal sensor data produces. Deep learning is especially promising to improve performance and discover features encoded in human motion dynamics. The following hybrid models have interesting applications in multimodal HAR:

1. *DeepConvLSTM* (Ordóñez, 2016)

DeepConvLSTM is a deep learning framework composed of convolutional and LSTM recurrent layers, capable of automatically learning feature representations and modeling temporal dependencies between their activation. This framework can be applied to different sensor modalities individually and can even fuse them to improve performance. It combines the potential of using deep CNNs to learn features in time series and LSTMs which are suitable to learn temporal dynamics in sensor signals. The main difference between DeepConvLSTM and the baseline CNN is the dense layers structure. In the case of DeepConvLSTM, the units of these layers are LSTM recurrent cells, and in the case of the baseline model, the units are non-recurrent and fully connected. Signals coming from the wearable sensors are processed by four convolutional layers, which serve to learn features from the data. Two dense layers perform a non-linear transformation, yielding a classification outcome with a softmax logistic regression output layer.

1. *AttnSense*(Ma, 2019)

AttnSense is an attention-based deep neural network model, which focuses on weighted averaging of a series of input vectors, where the attention weights represent the relative importance of the corresponding input. This is especially applicable in HAR because different sensor modalities have different domains: for example, gyroscopes may be better in distinguishing directional features while accelerometers may be better in distinguishing between magnitude-based features. Also, not all timesteps contribute equally to the activity recognition task, so AttnSense considers the temporal dependencies in HAR. AttnSense consists of (1) an individual convolutional subnet for each sensor to extract modality-specific features, (2) an attention-fusion subnet that considers the relative importance of each sensor modality to fuse together modality-specific features, (3) an attention-based Gated Recurrent Units (GRU) subnet that extracts the importance of different timesteps and fuse together the hidden state, and (4) an output layer that uses softmax logistic regression output to obtain probabilities for classification.

1. *DeepSense*(Yao, 2017)

DeepSense is the first architecture that can model temporal relationships and fuse multimodal sensor inputs. The latter will be especially important, as our project has more sensor inputs than the average HAR model, and it adds to our ability to explore interactions within local sensor inputs. DeepSense uses a combination of convolutional and recurrent neural networks. (1) Multimodal data is split by time interval and type of sensor and is fed into an individual CNN to learn the interactions of different sensors within each interval and overall longitudinal features of each sensor input. (2) The new feature vectors that represent the latter are processed through an RNN with Gated Recurrent Units (GRU) to learn historical features. The DeepSense framework can be modeled in two ways to address noise differently: regression and classification. For the classification framework, the CNN extracts local sensor features and the RNN extracts temporal features.

**Activity Classification Challenges**

Using identical models on the same data to classify different behaviors can have dramatically different results. There are a few relevant design challenges that we aim to avoid in the creation of our model, as outlined in the Bulling, et. al paper:

1. *Definition and Diversity of Physical Activities*

Deciding and defining which activities will be included in the model informs much of how a HAR model will perform. Trying to balance a tradeoff to have both high accuracy and a high level of generalizability will compound this problem. Not only does our model have to recognize activities from different sensors and sources, but the requirements for each activity must balance the tradeoff between intraclass variability and interclass similarity. This issue will be at the forefront of our considerations when tuning the model.

1. *Intraclass Variability*

Since the same activity can be performed in different ways by different people, some HAR models may classify the same activity as different activities incorrectly. Thus, there is a tradeoff between having a more detailed list of features (classified activities) that may discriminate too much against different ways people perform actions and having a more general list of features that is more flexible classifying the same action in different ways. Intraclass variability could be fixed by tuning hyperparameters properly.

1. *Interclass Similarity*

This phenomenon is the opposite of intraclass variability, in which a HAR model classifies different tasks that have very similar sensor readings as the same task, such as eating and brushing teeth. A remedy to interclass similarity is incorporating more sources of data. For example, a single accelerometer on the wrist may not be able to differentiate between standing and sitting, but a second accelerometer on the ankle would give more insight into the true activity. Our multimodal approach could reduce an interclass similarity problem (Bulling et al. 2014).

1. *Null Class Problem*

In multichannel time-series HAR, datasets are often imbalanced towards the “NULL” class. Given the imbalance between relevant and irrelevant data, activities of interest can be easily confused with activities of interest can easily be confused with activities that have similar patterns but are irrelevant to the application. The NULL class can be identified by thresholding on either the raw feature values or the confidence scores calculated by the classifier.

The data in Bent, et. al specify and assign labels to four types of activities: rest, typing, deep breathing, and activity. Activity refers to five minutes of walking (Bent, et. al, 2020).

**Model Validation**

Model validation, which is to gauge model performance and tune hyperparameters, uses a subset of the data. Numerous previous studies have used leave-one-out cross-validation (LOOCV) for unimodal models. Due to the number of features present in a multimodal model, in conjugation with the size of the data, LOOCV will be computationally expensive. Depending on the aggregation level chosen for the relevant data, which infers the size of the overall data, K-fold validation may be suitable for the multimodal models. However, observations from individual subjects are likely to be correlated due to underlying environmental, biological, and demographics factors. This, in combination with the temporal correlation of an individual, has been shown to overestimation of classification performance (Dehghani et al., 2018). Hold-out validation is the least computationally expensive method but is also subject to the correlative and temporal components of observations from a single individual. While LOOCV is computationally expensive, it does aid in avoiding this overestimation in model performance. Since our data only uses 51 individuals, LOOCV is feasible for the amount of data we will be working with, especially based on the level of aggregation chosen (Bent et al., 2020).

**Performance Evaluation**

In supervised machine learning, meaning that ground truth labels have been provided, the measures of precision, recall, and F1 are simple and effective measurements for the accuracy of the model. Even complex hybrid neural network architectures such as AttnSense (Ma, 2019), DeepSense (Yao, 2017), and DeepConvLSTM (Ordonez, 2016; Singh, 2020) all use these measurements to validate their model (Hammerla, 2016). Precision and recall measure the percent of classified true positives out of all classified positives and the percent of classified true positives out of all true positives, respectively. The most commonly used metrics for multi-classes are F1 score, Average Accuracy, and Log-loss. The F1 score is a composite of precision and recall that essentially balances this tradeoff: is the model too sensitive and is successful at classifying all true positives but has too many false positives, or is the model too specific and can preventing false positives but fails to classify all true positives as positive? Another common measure of efficacy is average accuracy, which is the proportion of how many correct classifications the model made.

Now that we have covered the ARC and the variations on the choices we can make throughout the steps, including what lacks current research has, it is now appropriate to introduce our own model. The specific gaps in the research will be explicitly stated in this section and be in line with the examples that were given in the ARC section.

## **Objectives**

Many HAR models only use one or two kinds of mechanical sensor data as inputs to infer behaviors (Dernbach, 2012; Kwapisz, 2011; Zeng, 2014). Our model is the first to incorporate several kinds of physiological data as well as one kind of mechanical sensor data. The model uses Blood Volume Pulse (physiological), Electrodermal Activity (physiological), Skin Temperature (physiological), and 3-axis accelerometry (mechanical).

With our novel approach to using both mechanical and physiological data for activity recognition, we could possibly provide more detailed insight into user behavior and habits. Thus, our aim to make our model flexible enough to work well with other HAR datasets could help in providing better care for chronic diseases that often require major lifestyle changes for patients and monitoring vulnerable patients over long periods of time such as elderly patients or patients in intensive care. Because our model is multimodal, including a range of different types of physiological and mechanical data can aid in the effort of making care more personalized and/or effective. Lastly, using elements of multi-attribute classification, which is the inclusion contextual data such as age and gender to further tailor HAR models to each user, could improve the model’s accuracy for new datasets that include demographic data (Lara, 2013).

In sum, the goal of the Human Activity Recognition Team is to create a predictive model that:

1. Takes in multimodal data from mechanical sensors (such as accelerometers) **and physiological** sensors (such as electrodermal sensors and pulse oximeters).
2. Classifies human activity (Rest, Deep Breathing, Walking, Typing) at high accuracy and precision, while being **generalizable and adaptable to other HAR datasets.**

## **Overview of Wearable Sensors**

A plethora of wearable biosensors exist on the market and are easily accessible by consumers and researchers. Two of the most popular consumer-grade wearables are the Apple Watch and the Fitbit (Dooley et al., 2017). These devices contain multiple sensors allowing individuals to monitor features such as their heart rate and blood pressure. Due to their popularity, data gathered from all of these sensors prove invaluable for researchers. However, most of the data is considered personal and not freely given out by the parent corporation. Surveys have shown that a large proportion of those who wear wearables do not feel comfortable sharing their wearable data (Perez & Sherali, 2018). For research purposes, investigators tend to prefer sensors that have higher sampling rates and contain more sensors. Research wearables generally can be broken down into three major subcategories: mechanical, physiological, and biochemical.

**Mechanical Sensors**

Mechanical sensors detect mechanical deformations that occur due to various stimuli (Lamkin-Kennard & Popovic, 2019). The most common sensor, present in almost all devices, is the accelerometer, which measures acceleration force along two or three perpendicular planes.

Gyroscopes are similar to accelerometers but are able to measure the rotation of objects in motion. Applications of these two types of mechanical sensors include fall detection and gait analysis. When location is relevant in a study, GPS devices are used to monitor a user’s location over time, although there can be issues with signal and detecting movement on a more granular level. Pressure sensors can be used for a variety of applications such as pulse and respiration monitoring.

**Physiological Sensors**

Physiological sensors are used to monitor physiological signals such as heart rate, electrodermal activity, skin temperature, and blood oxygen saturation in a non-invasive manner. Photoplethysmographic (PPG) sensors work by using LED lights to detect changes in the blood. Generally, green LEDs are used for measuring the amount of oxygenated blood while infrared diodes are used for measuring blood flow and pressure (Tamura, et al., 2014). Electrodermal sensors measure electrical properties of the skin in response to sweat (Freedman et al., 1994). Skin conductance (SC), or electrodermal activity (EDA), is a useful physiological metric for measuring emotional arousal and has been used largely in quantifying mental and emotional states longitudinally with wearables. Eccrine sweat glands, which are located all over the human body, are influenced by emotional arousal thus allowing for emotion recognition (Martinez-Rodrigo et al., 2015). Finally, skin temperature is measured through an optical infrared thermometer (Garbarino, 2014). In conjugation with other features derived from sensors, skin temperature has been used to predict core body temperature, thermal comfort sensation, stress in students, and ulcer development in feet of diabetic patients (Sano, 2018; Armstrong et al., 2007).

**Biochemical Sensors**

Biochemical sensors test for chemicals for medical diagnosis or patient monitoring. These sensors can be both invasive and non-invasive. Examples of these kinds of biochemical sensors include glucose, alcohol, and electrolyte sensors. Many of these sensors work by detecting biomarkers in blood, sweat, tears, and respiratory gases (Li & Wen, 2020). Glucose sensors, generally used by diabetic patients, are probably the most well known biochemical sensors. Generally, these sensors are invasive, requiring the prick of a finger for a blood sample. However, in the past two decades, non-invasive versions of these glucose sensors have been created for consumer use (Vashist, 2012).

The placement of sensor(s) has shown to affect data quality, and in turn, model performance as well. Single accelerometers, especially on the wrist, have trouble differentiating between standing and sitting, as the torso is in the same position. So, single accelerometers are most effective on the waist or torso (Yang, 2010). Furthermore, while accelerometers on the wrist were more effective than other placements in showing very low to medium level activities, such as lying down, eating, and walking, wrist accelerometers were not optimal in showing high level and transitional activities, such as running and getting up (Atallah, 2011). However, waist and chest-worn wearables are not as convenient and attractive as wrist wearables (the most popular form of consumer wearable), prompting the addition of more and different types of sensors in wearables, such as gyroscopes, pressure sensors, magnetometers, and physiological sensors. The addition of a gyroscope to a triaxial accelerometer helps in discerning between sitting and standing, and wrist gyroscopes can pick up small differences in low-level activities such as typing and eating (Parkka, 2007). Also, the inclusion of physiological sensors such as the photoplethysmographic (PPG) sensor, skin conductance sensors, and skin thermometers can inform the model about the body’s physiological state.

PPG sensors are currently the standard for wearable heart rate, blood oximetry, and blood volume pressure sensors. In addition to performing multiple measurements, PPG sensors are low-cost and fairly reliable (Ghamari, 2018). Wearable wrist PPG sensors are suitable in consumer-grade wearables and tend to perform well in almost all locations on the wrist, but are sometimes prone to somewhat poor data quality due to motion artifacts, and may not be as effective on females’ smaller wrists (Lee, 2016). Some other factors that affect wrist PPG efficacy include material of the sensor, blood vessel pressure, motion artifacts due to adjacent muscle movements, and displacement of the sensor due to physical activity (Ghamari, 2018).

Wearable wrist EDA sensors show promising results in giving accurate readings. Of the non-obtrusive locations a wearable sensor could be, the wrist was one of the most highly responsive parts physiologically for skin conductance (Dooren, 2012). Wrist wearable EDA sensors have shown above 93% correlation with baseline wrist and finger electrode measurements (Poh, 2010).

**Sensors Summary**

To summarize, a multimodal approach to collecting data may help in solving issues of data quality. A paper on an older model of the multimodal wearable sensor that our data was collected from showed high accuracy in the readings from the PPG sensor, EDA sensor, accelerometer, and skin temperature sensor (Garbarino, 2014).

To train our multimodal model, Empatica E4 data from the Bent et al., 2020 study will be used. This research-grade wearable contains a PPG, triaxial accelerometer, electrodermal sensor, and a skin thermometer which uses an optical infrared thermometer. The sampling rate of these sensors is 64 Hz, 32 Hz, 4 Hz, and 4 Hz, respectively (Bent et al., 2020).

## **Design Challenges**

There are a few relevant design challenges that we aim to avoid in the creation of our model, as outlined in the Lara paper:

Because our model will be put into a pipeline for public use, it is extremely important to optimize both runtime and processing power to provide optimal user experience, no matter the dataset used. Some properties that we must consider are model accuracy, system latency, and processing power. A possible solution to these design tradeoffs would be to incorporate a neural network model that processes and fuses data from each sensor, such as the AttnSense (Ma, 2019) and DeepConv LSTM Models (Ordonez, 2016).

Class imbalance tends to be a large problem in HAR. In datasets from continuous, long-term monitoring, a majority of the time is spent in resting positions such as sitting or sleeping, while more “complex” activities such as brushing teeth happen only a few times per day. Thus, a class imbalance between less and more complex activities, and a predictive model may favor the larger class even when that assumption is incorrect. Some possible solutions to this problem would be to either oversample the smaller class (Bulling, Blanke, & Schiele, 2014) or use SMOTE methods to artificially create data from the less-occurring class. A paper by Nguyen proposes BLL SMOTE, which is designed to improve performance in neural network architectures for HAR. More research needs to be done to see if BLL SMOTE is effective for a temporal data-driven recurrent neural network (2018).

## **References**

Armstrong, D. G., Holtz-Neiderer, K., Wendel, C., Mohler, M. J., Kimbriel, H. R., & Lavery, L. A. (2007). Skin Temperature Monitoring Reduces the Risk for Diabetic Foot Ulceration in High-risk Patients. *The American Journal of Medicine*, *120*(12), 1042–1046. doi: 10.1016/j.amjmed.2007.06.028

Atallah, L., Lo, B., King, R., & Yang, G.-Z. (2011). Sensor Positioning for Activity Recognition Using Wearable Accelerometers. *IEEE Transactions on Biomedical Circuits and Systems*, *5*(4), 320–329. doi: 10.1109/tbcas.2011.2160540

Bent, B., Goldstein, B., Kibbe, W., & Dunn, J. (2020). Investigating Sources of Inaccuracy in Wearable Optical Heart Rate Sensors (Preprint). doi: 10.2196/preprints.18126

Bulling, A., Blanke, U., & Schiele, B. (2014). A tutorial on human activity recognition using body-worn inertial sensors. *ACM Computing Surveys*, *46*(3), 1–33. doi: 10.1145/2499621

Cheng, X., Yang, B., Tan, K., Isaksson, E., Li, L., Hedman, A., … Li, H. (2019). A Contactless Measuring Method of Skin Temperature based on the Skin Sensitivity Index and Deep Learning. *Applied Sciences*, *9*(7), 1375. doi: 10.3390/app9071375

Dehghani, A., Glatard, T., & Shihab, E. (2019). Subject Cross Validation in Human Activity Recognition. *ArXiv, abs/1904.02666*.

Dernbach, S., Das, B., Krishnan, N. C., Thomas, B. L., & Cook, D. J. (2012). Simple and Complex Activity Recognition through Smart Phones. *2012 Eighth International Conference on Intelligent Environments*. doi: 10.1109/ie.2012.39

Dilraj, N., Rakesh, K., Krishnan, R., & Ramesh, M. (2015). A Low Cost Remote Cardiac Monitoring Framework for Rural Regions. *Proceedings of the 5th EAI International Conference on Wireless Mobile Communication and Healthcare - "Transforming Healthcare through Innovations in Mobile and Wireless Technologies"*. doi: 10.4108/eai.14-10-2015.2261648

Dooley, E., Golaszewski, N., & Bartholomew, J. (2017). Estimating Accuracy at Exercise Intensities: A Comparative Study of Self-Monitoring Heart Rate and Physical Activity Wearable Devices. *JMIR Mhealth And Uhealth*, *5*(3), e34. doi: 10.2196/mhealth.7043

Dooren, M. V., Vries, J. (G.-J. D., & Janssen, J. H. (2012). Emotional sweating across the body: Comparing 16 different skin conductance measurement locations. *Physiology & Behavior*, *106*(2), 298–304. doi: 10.1016/j.physbeh.2012.01.020

Dunn, J., Runge, R., & Snyder, M. (2018). Wearables and the medical revolution. *Personalized Medicine*, *15*(5), 429–448. doi: 10.2217/pme-2018-0044

Freedman, L. W., Scerbo, A. S., Dawson, M. E., Raine, A., McClure, W. O., & Venables, P. H. (1994). The relationship of sweat gland count to electrodermal activity. *Psychophysiology*, *31*(2), 196–200. [https://doi.org/10.1111/j.1469-8986.1994.tb01040.](https://doi.org/10.1111/j.1469-8986.1994.tb01040.x)x

R. Fu, Z. Zhang and L. Li, "Using LSTM and GRU neural network methods for traffic flow prediction," 2016 31st Youth Academic Annual Conference of Chinese Association of Automation (YAC), Wuhan, 2016, pp. 324-328, doi: 10.1109/YAC.2016.7804912.

Gallup, I. (2020). One in Five U.S. Adults Use Health Apps, Wearable Trackers. Retrieved 29 May 2020, from https://news.gallup.com/poll/269096/one-five-adults-health-apps-wearable-trackers.aspx

Garbarino, M., Lai, M., Tognetti, S., Picard, R., & Bender, D. (2014). Empatica E3 - A wearable wireless multi-sensor device for real-time computerized biofeedback and data acquisition. *Proceedings of the 4th International Conference on Wireless Mobile Communication and Healthcare - "Transforming Healthcare through Innovations in Mobile and Wireless Technologies"*. doi: 10.4108/icst.mobihealth.2014.257418

Ghamari, M. (2018). A review on wearable photoplethysmography sensors and their potential future applications in health care. *International Journal of Biosensors & Bioelectronics*, *4*(4). doi: 10.15406/ijbsbe.2018.04.00125

Koskimäki, H., Mönttinen, H., Siirtola, P., Huttunen, H.-L., Halonen, R., & Röning, J. (2017). Early detection of migraine attacks based on wearable sensors. *Proceedings of the 2017 ACM International Joint Conference on Pervasive and Ubiquitous Computing and Proceedings of the 2017 ACM International Symposium on Wearable Computers*. doi: 10.1145/3123024.3124434

Kim, E. (2020). Interpretable and Accurate Convolutional Neural Networks for Human Activity Recognition. *IEEE Transactions on Industrial Informatics*, 1–1. doi: 10.1109/tii.2020.2972628

Kroll, R. R., Mckenzie, E. D., Boyd, J. G., Sheth, P., Howes, D., Wood, M., & Maslove, D. M. (2017). Use of wearable devices for post-discharge monitoring of ICU patients: a feasibility study. *Journal of Intensive Care*, *5*(1). doi: 10.1186/s40560-017-0261-9

Kwapisz, J. R., Weiss, G. M., & Moore, S. A. (2011). Activity recognition using cell phone accelerometers. *ACM SIGKDD Explorations Newsletter*, *12*(2), 74–82. doi: 10.1145/1964897.1964918

Lara, O. D., & Labrador, M. A. (2013). A Survey on Human Activity Recognition using Wearable Sensors. *IEEE Communications Surveys & Tutorials*, *15*(3), 1192–1209. doi: 10.1109/surv.2012.110112.00192

Lamkin-Kennard, K., & Popovic, M. (2019). Sensors: Natural and Synthetic Sensors. *Biomechatronics*, 81-107. doi: 10.1016/b978-0-12-812939-5.00004-5

Lee, S.-M., Cho, H., & Yoon, S. M. (2017). Statistical noise reduction for robust human activity recognition. *2017 IEEE International Conference on Multisensor Fusion and Integration for Intelligent Systems (MFI)*. doi: 10.1109/mfi.2017.8170442

Lee, S., Shin, H., & Hahm, C. (2016). Effective PPG sensor placement for reflected red and green light, and infrared wristband-type photoplethysmography. *2016 18th International Conference on Advanced Communication Technology (ICACT)*. doi: 10.1109/icact.2016.7423470

Li, G., & Wen, D. (2020). Wearable biochemical sensors for human health monitoring: sensing materials and manufacturing technologies. *Journal Of Materials Chemistry B*, *8*(16), 3423-3436. doi: 10.1039/c9tb02474c

Li, K., Habre, R., Deng, H., Urman, R., Morrison, J., & Gilliland, F. et al. (2019). Applying Multivariate Segmentation Methods to Human Activity Recognition From Wearable Sensors’ Data. *JMIR Mhealth And Uhealth*, *7*(2), e11201. doi: 10.2196/11201

Ma, H., Li, W., Zhang, X., Gao, S., & Lu, S. (2019). AttnSense: Multi-level Attention Mechanism For Multimodal Human Activity Recognition. *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence*. doi: 10.24963/ijcai.2019/431

Martínez-Rodrigo, A., Zangróniz, R., Pastor, J., & Sokolova, M. (2015). Arousal level classification of the aging adult from electro-dermal activity: From hardware development to software architecture. *Pervasive And Mobile Computing*, *34*, 46-59. doi: 10.1016/j.pmcj.2016.04.006

Mukhopadhyay, S. C. (2015). Wearable Sensors for Human Activity Monitoring: A Review. *IEEE Sensors Journal*, *15*(3), 1321–1330. doi: 10.1109/jsen.2014.2370945

Nguyen, Ky Trung, François Portet, Catherine Garbay. Dealing with Imbalanced data sets for Human Activity Recognition using Mobile Phone sensors. 3rd International Workshop on Smart Sensing Systems, Jun 2018, Rome, Italy. Ffhal-01950472f

Nweke, H. F., Teh, Y. W., Al-Garadi, M. A., & Alo, U. R. (2018). Deep learning algorithms for human activity recognition using mobile and wearable sensor networks: State of the art and research challenges. *Expert Systems with Applications*, *105*, 233–261. doi: 10.1016/j.eswa.2018.03.056

Ordóñez, F., & Roggen, D. (2016). Deep Convolutional and LSTM Recurrent Neural Networks for Multimodal Wearable Activity Recognition. *Sensors*, *16*(1), 115. doi: 10.3390/s16010115

Parkka, J., Ermes, M., Antila, K., Gils, M. V., Manttari, A., & Nieminen, H. (2007). Estimating Intensity of Physical Activity: A Comparison of Wearable Accelerometer and Gyro Sensors and 3 Sensor Locations. *2007 29th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*. doi: 10.1109/iembs.2007.4352588

Perez, A., & Zeadally, S. (2018). Privacy Issues and Solutions for Consumer Wearables. *IT Professional*, *20*(4), 46-56. doi: 10.1109/mitp.2017.265105905

Piwek, L., Ellis, D. A., Andrews, S., & Joinson, A. (2016). The Rise of Consumer Health Wearables: Promises and Barriers. *PLOS Medicine*, *13*(2). doi: 10.1371/journal.pmed.1001953

Poh, M.-Z., Swenson, N. C., & Picard, R. W. (2010). A Wearable Sensor for Unobtrusive, Long-Term Assessment of Electrodermal Activity. *IEEE Transactions on Biomedical Engineering*, *57*(5), 1243–1252. doi: 10.1109/tbme.2009.2038487

Qiu, S., Wang, Z., Zhao, H., Liu, L., & Jiang, Y. (2018). Using Body-Worn Sensors for Preliminary Rehabilitation Assessment in Stroke Victims With Gait Impairment. *IEEE Access*, *6*, 31249–31258. doi: 10.1109/access.2018.2816816

Tamura, T., Maeda, Y., Sekine, M., & Yoshida, M. (2014). Wearable Photoplethysmographic Sensors—Past and Present. *Electronics*, *3*(2), 282–302. doi: 10.3390/electronics3020282

Sano, A., Taylor, S., Mchill, A. W., Phillips, A. J., Barger, L. K., Klerman, E., & Picard, R. (2018). Identifying Objective Physiological Markers and Modifiable Behaviors for Self-Reported Stress and Mental Health Status Using Wearable Sensors and Mobile Phones: Observational Study. *Journal of Medical Internet Research*, *20*(6). doi: 10.2196/jmir.9410

Singh, S. P., Lay-Ekuakille, A., Gangwar, D., Sharma, M. K., & Gupta, S. (2020). Deep ConvLSTM with self-attention for human activity decoding using wearables. *ArXiv, abs/2005.00698*.

Sung, M., Marci, C., & Pentland, A. (2005). Wearable feedback systems for rehabilitation. *Journal of NeuroEngineering and Rehabilitation*, *2*(1), 17. doi: 10.1186/1743-0003-2-17

Vashist, S. (2012). Non-invasive glucose monitoring technology in diabetes management: A review. *Analytica Chimica Acta*, *750*, 16-27. doi: 10.1016/j.aca.2012.03.043

Vogels, E. A. (2020, January 9). About one-in-five Americans use a smart watch or fitness tracker. Retrieved from https://www.pewresearch.org/fact-tank/2020/01/09/about-one-in-five-americans-use-a-smart-watch-or-fitness-tracker/

Warwick, B., Symons, N., Chen, X., & Xiong, K. (2015). Detecting Driver Drowsiness Using Wireless Wearables. *2015 IEEE 12th International Conference on Mobile Ad Hoc and Sensor Systems*. doi: 10.1109/mass.2015.22

Winkley, J., Jiang, P., & Jiang, W. (2012). Verity: an ambient assisted living platform. *IEEE Transactions on Consumer Electronics*, *58*(2), 364–373. doi: 10.1109/tce.2012.6227435

Yang, C.-C., & Hsu, Y.-L. (2010). A Review of Accelerometry-Based Wearable Motion Detectors for Physical Activity Monitoring. *Sensors*, *10*(8), 7772–7788. doi: 10.3390/s100807772

Yang, J. B., Nguyen, M. N., San, P. P., Xiao, L. L., & Krishnaswamy, S. (2015). Deep Convolutional Neural Networks On Multichannel Time Series For Human Activity Recognition. *Proceedings of the Twenty-Fourth International Joint Conference on Artificial Intelligence (IJCAI 2015) 3995*. Retrieved from https://www.aaai.org/ocs/index.php/IJCAI/IJCAI15/paper/view/10710/11297

Yao, S., Hu, S., Zhao, Y., Zhang, A., & Abdelzaher, T. (2017). DeepSense. *Proceedings of the 26th International Conference on World Wide Web*. doi: 10.1145/3038912.3052577

Zeng, M., Nguyen, L. T., Yu, B., Mengshoel, O. J., Zhu, J., Wu, P., & Zhang, J. (2014). Convolutional Neural Networks for Human Activity Recognition using Mobile Sensors. Proceedings of the 6th International Conference on Mobile Computing, Applications and Services. doi: 10.4108/icst.mobicase.2014.257786

Zhao, A., Dong, J., Li, J., Qi, L., & Yu, H. (2018). LSTM for diagnosis of neurodegenerative diseases using gait data. *Ninth International Conference on Graphic and Image Processing (ICGIP 2017)*. doi: 10.1117/12.2305277