Discriminative Information Added by Wearable Sensors for Early Screening – a Case Study on Diabetic Peripheral Neuropathy

Abstract — Wearable inertial sensors have demonstrated their potential to screen for various neuropathies and neurological disorders. Most such research has been based on classification algorithms that differentiate the control group from the pathological group, using biomarkers extracted from wearable data as predictors. However, such methods often lack quantitative evaluation of how much information provided by the wearable biomarkers contributes to the overall prediction. Despite promising results from internal cross validation, their utility in clinical practice remains unclear.

In this paper, we highlight in a case study – early screening for diabetic peripheral neuropathy (DPN) – evaluation methods for quantifying the contribution of wearable inertial sensors. Using a quick-to-deploy wearable sensor system, we collected 106 in-hospital diabetic patients’ gait data and developed logistic regression models to predict the risk of a diabetic patient having DPN. Adopting various metrics, we evaluated the discriminative information added by gait biomarkers and how much it improved screening. The results show that the proposed wearable system added useful information significantly ($p < 3 \times 10^{-4}$) to the existing clinical standards, and boosted the C-index significantly ($p < 0.02$) from 0.75 to 0.84, surpassing the current survey-based screening methods used in clinics.

I. INTRODUCTION

Wearable technologies have been revolutionizing clinical research by supplying objective, quantitative measures (“biomarkers”) with rich, dynamic information regarding an individual’s functional or physiological state [1]. When demonstrating the utility of these technologies in clinical settings, one typical approach relies on classifying health outcomes with discriminative algorithms, such as support-vector machines and random forests, and applying leave-one-subject-out cross-validation or repeated k-fold cross-validation to evaluate the methods with metrics such as accuracy, sensitivity, and precision. Given particular datasets and classifiers, biomarkers producing high validation accuracy are considered useful for clinical adoption. However, as the contribution of the biomarkers is rarely evaluated against existing clinical methods, their true utility in clinics is open to question.

Other challenging issues are also associated with this practice. One is the class imbalance problem during the training phase [2] [3]. As the population with a disease is often a small percentage of the general population, the sample obtained from an observational study is likely to have imbalanced classes as well. In addition, if the sample size of the dataset is small, the evaluation results may have a high variance too. Overall, these issues challenge the consistent evaluation of biomarkers extracted from wearable sensor data under a machine-learning framework, potentially impeding the reproducibility of wearable sensor-based research.

In this paper, we aim to address these issues in a case study – screening for diabetic peripheral neuropathy (DPN). Early screening for DPN is a critical part of diabetes management, as undetected DPN may lead to amputation. DPN is often associated with degradation in gait, such as slower gait, shorter steps, and more limited knee and ankle mobility. In our case study, we used an ear-worn inertial sensor [4] to quantitatively measure gait differences in DPN with different degrees of severity. Ultimately, we aim to incorporate wearable sensor-enabled gait analysis into clinical settings, and test whether wearable-enabled data can provide additional information when screening diabetes mellitus (DM) patients. Our contribution lies in the following:

• Proposing a strapless, quick-to-deploy wearable gait analysis system to enable early screening for DPN
• Evaluating the proposed system on 106 DM patients
• Adopting novel strategies to evaluate the added information from wearable sensors relative to existing clinical methods

II. DATA COLLECTION METHODS

A. Participants

We recruited 133 in-hospital patients, who had been diagnosed with DM according to the WHO 2006 criteria [5], from 2014 to 2016 at Shanghai Ninth People’s Hospital, affiliated to Shanghai Jiao Tong University School of Medicine in China.

This work is partially supported by NIH UL1TR002649
The study protocol was approved by the Ethics Committee of Shanghai Ninth People’s Hospital. 27 patients who did not have confirmatory diagnoses or complete gait data were excluded from the analysis. The remaining 106 participants (aged 38 to 83, 52 male and 54 female), of whom 30 were confirmed as DPN positive (28.3% of the 106 subjects) and 76 as DPN negative, were included in the analysis.

B. Standard Clinical Screening and Diagnosis

To identify the patients with DPN, the physicians also obtained the 106 participants’ Michigan neuropathy screening instrument (MNSI) survey data [6], including both the history (MNSI History) and records of physical examination of motor and sensory functions performed on the non-dominant upper and lower limbs (MNSI PE), and nerve conduction studies (NCS) [7]. The results from the NCS were used as the ground truth diagnosis.

C. Proposed System

To collect the gait data, an ear-worn inertial sensor (e-AR) consisting of a 3-axis accelerometer was used, as shown in Figure 1. When fitted on the ear during walking, the e-AR sensor captures motion patterns caused by movement in lateral, forward and vertical directions using a 3-axis accelerometer, at a sampling rate of 100Hz. During the experiment, the raw accelerometer data were wirelessly streamed to a Windows 8 tablet in real-time. During their in-hospital stay, participants were asked to wear the e-AR sensor and walk as fast as possible for 10 meters for three rounds. After the experiment, the de-identified data were sent back to Imperial College London for post data analysis. From the raw accelerometer data, gait patterns and critical gait events can be extrapolated. Details of the data analysis are presented in Section III.

III. DATA ANALYSIS

A. Gait Feature Extraction from e-AR Sensors

To extract gait features from the fast 10-meter walk data, we first segmented gait cycles using the vertical acceleration. Using a customized peak detection algorithm to identify heel-strike events, we delineated each gait cycle from a heel-strike event to the instant before the next heel-strike event (shown in Figure 2). Then with each gait cycle, critical gait events (e.g. heel-strike, toe-off) were detected, and descriptive measures (e.g. variance, skewness, and kurtosis) of the gait cycle signals were obtained. Other features include the amplitudes sampled at the heel-strike (HS) and toe-off events (TO); the range of gait signals per gait cycle; the maximum amplitudes of three axes of acceleration; cadence per minute; average step length and gait speed during the 10-meter walk. Some of these gait features were selected to enter the predictive model for estimating the risk of having DPN.

B. Predictive Modeling

To develop a predictive model for screening, we resorted to a simple logistic regression model which is captured in Equation (1):

$$\text{Prob}\{Y = 1|X\} = \frac{1 + \exp(-X\beta)}{1}$$

Equation (1) returns the conditional probability that a DM patient has DPN (labeled as “1”) given the observation $X$ – in our case the gait predictors and the scored items from the MNSI surveys. The binary decision whether a given patient has DPN can be obtained by thresholding the conditional probability. This direct probabilistic model has the following advantages:

- The “class imbalance” problem caused by the natural prevalence of DPN can be handled directly by the probabilistic model, as it directly estimates the probability of DPN prevalence as a function of the predictors $X$;
- As DPN progresses over the years, DPN symptoms may be a continuous spectrum that does not have a clear enough cutoff to be dichotomized as DPN positive or negative;
- Furthermore, the onus of diagnostic decisions should be on the medical experts acting on the risk probability obtained from the logistic regression model, instead of on the algorithm.

Five categories of logistic regression model were fitted to the data using various predictors: 1) Model A uses only the total score from the MNSI History; 2) Model B uses only the total score from the MNSI PE; 3) Model C uses both the MNSI History and the MNSI PE; 4) Model D uses only gait features from the e-AR sensor; 5) Model E combines the MNSI history and the gait features from the e-AR sensor. Age and gender were used in all five models. Models A, B, and C can be considered to be the baseline models used in current clinical settings. Model D captures the discriminative capability of the e-AR sensor alone. Model E tests whether
gait features obtained from the e-AR sensor will enhance existing clinical tests. To select predictive gait features, we tested the group mean difference between the positive and negative groups for each feature. Features showing significant differences ($\alpha=0.05$) were selected for Models D and E. In our evaluation, we focused on testing whether Model E significantly adds information to the baseline models.

C. Evaluation

Our goal is to test whether using wearable sensors is useful and quantify the additional information they provide, rather than validating whether our models can be adopted for clinical screening at this stage. Thus, we did not validate our model using independent samples reserved from model development. Instead, we focused on comparing the existing predictive models with new models incorporating the gait biomarkers.

**Discriminative Ability of Models** The area under the receiver operating characteristic (ROC) curve (AUROC or C-index) is often used as a metric to evaluate overall accuracy. We also used this metric to evaluate all five models. To test whether Model E significantly improved on Model A in terms of discriminative ability, we used DeLong’s test to compare the AUROC.

**Added Information from the Sensor-based Gait Test** As clinical tests exist to screen for DPN, new tests enabled by wearable sensors must provide additional information in order to be considered for clinical use [8]. The standard test for added information is a likelihood ratio (LR) test for the nested models [9]. A significantly increased LR test stat (also known as LR $\chi^2$) means the new biomarkers improve the fit of the model in a statistically significant sense, suggesting that the added information provided by the new test (the gait test, in our case) has advantages. The new information added ($I_A$) to the existing method can be quantified as [10]:

$$I_A = 1 - LR_{existing}/LR_{new},$$

which is the fraction of the information added to the model by using the gait features, based on the ratio of the LR of the existing predictive model ($LR_{existing}$) to the new model involving the biomarkers ($LR_{new}$). Other metrics, such as net reclassification index (NRI) and integrated discriminative index (IDI) were also proposed to test the effect of the additional information [11] [12]. We applied these evaluation metrics to determine whether Model E improved on Model A.

**Probability Calibration and Prediction Error** Since an overfitted model is likely to underestimate the risk of low-risk patients whilst overestimating the risk of high-risk patients [13], a risk prediction model must be calibrated to avoid overfitting. To check for overfitting, we plotted the predicted probability against the observed probability and quantified how close they were, using the Brier score. This is the mean squared error (MSE) between the true probabilities and the estimated probabilities from the probabilistic model described in Equation (1) [9]. The smaller the Brier score, the better the model.

**Bootstrapping:** To compensate for overfitting, we adopted a bootstrapping procedure as follows: first, a random sample was drawn from the original dataset $X$ and $Y$ of $N$ samples with a replacement. Then a logistic regression model was derived from the bootstrapped dataset and applied to the original dataset, from which common evaluation metrics such as accuracy can be obtained. The difference between the evaluation metrics from the original dataset and the bootstrapped dataset is called “optimism”. Repeating this process 200 times, we can obtain the average optimism and correct the original evaluation metrics by subtracting optimism from them. This bootstrapping process along with the predictive modeling were achieved using the “rms” package in R [9].

IV. RESULTS

In all five models, being male implied a significantly higher risk of developing DPN. However, age at the time of the study did not significantly affect DPN risk. After testing for the group mean difference of each gait feature, we found the most discriminative features came from lateral acceleration. Three features showed significant differences between groups — the skewness of lateral acceleration (feature 1), the maximum amplitude of lateral acceleration (feature 2), and the range normalized maximum amplitude of lateral acceleration (feature 3). These features revealed that gait patterns in DPN positive groups, when walking as fast as possible, tend to be less extreme and exhibit less sway in the lateral direction. To our surprise, although not significant when testing for group mean difference, gait features such as step length and speed tended to be higher in the positive group. We suspect this is due to the height impact on DPN [14], which confounds the assessment of gait degradation. Since features 2 and 3 capture similar aspects of gait patterns, we picked features 1 and 3 to be incorporated into Models D and E. Further investigation will be required to confirm these findings.

Enhanced by predictors provided by the e-AR sensor, Model E showed better overall performance (Table I). Compared to the best baseline model (Model A), gait features provided by wearable sensors added a significant amount of information to Model E ($p<3e-4$, Table II), quantified ($I_A$) as contributing 37% information to Model E. Other evaluation metrics, NRI and IDI, also demonstrated that Model E added significant information to Model A. Moreover, despite adding two more predictors from the e-AR sensor, Model E was well-calibrated (Figure 3) and showed a better Brier score than all baseline models.

In terms of discriminative capability, Model E is by far the best model with the largest AUROC and Model B is the worst. Compared to MNSI PE, MNSI History shows better discriminative ability, and combining the two surveys (Model C) does not improve the discriminative capability. Indeed, when it comes to evaluating physical symptoms, using only two gait features from the e-AR sensor (Model D) is better than the existing clinical physical examination survey (i.e., MNSI PE in Model B), although Model D has similar performance to models using the MNSI History score (Models A and C). Most
sensors can provide useful additional information for screening among diabetic patients.

![Fig. 3: Probability calibration curve for Model E. The predicted curves are very close to the calibration line, indicating that Model E did not overfit.](image)

![Fig. 4: ROCs for all models. Overall, Model E (green) has the largest area under the ROC, demonstrating the best discriminability.](image)

importantly, when combining physical information captured by the sensors with DM history captured by the MNSI History score, the discriminative measure, C-index, is boosted from 0.75 to 0.84, which is a significant improvement ($p < 0.02$).

The discriminative ability of all five models is also illustrated in Figure 4.

### TABLE I: Model performance

<table>
<thead>
<tr>
<th>Model</th>
<th>C-index</th>
<th>Adjusted C</th>
<th>Brier</th>
<th>Adjusted B</th>
<th>LR Stat</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model A</td>
<td>0.77</td>
<td>0.75</td>
<td>0.14</td>
<td>0.16</td>
<td>27.5</td>
</tr>
<tr>
<td>Model B</td>
<td>0.71</td>
<td>0.68</td>
<td>0.18</td>
<td>0.19</td>
<td>12.9</td>
</tr>
<tr>
<td>Model C</td>
<td>0.76</td>
<td>0.74</td>
<td>0.14</td>
<td>0.16</td>
<td>27.6</td>
</tr>
<tr>
<td>Model D</td>
<td>0.77</td>
<td>0.74</td>
<td>0.16</td>
<td>0.18</td>
<td>23.4</td>
</tr>
<tr>
<td>Model E</td>
<td>0.87</td>
<td>0.84</td>
<td>0.12</td>
<td>0.14</td>
<td>43.9</td>
</tr>
</tbody>
</table>

Adjusted C and B are the C-index and Brier score after correcting for “optimism” using the bootstrapping procedure.

### TABLE II: Model comparison for added information

<table>
<thead>
<tr>
<th>Model</th>
<th>C-index</th>
<th>LR Stat</th>
<th>NRI</th>
<th>IDI</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model A</td>
<td>0.77</td>
<td>27.5</td>
<td>N/A</td>
<td>N/A</td>
</tr>
<tr>
<td>Model E</td>
<td>0.87</td>
<td>43.9</td>
<td>0.502</td>
<td>0.117</td>
</tr>
</tbody>
</table>

Overall, our results show that gait data collected by e-AR sensors can provide useful additional information for screening DM patients for neuropathy, highlighting the capability of wearable sensors to capture a patient’s functional state better than survey-based physical examinations. Since surveys probing DM history (MNSI History) can be done in-home remotely, the additional, better information provided by the wearable sensors suggests that our system has the potential to be used jointly with the history survey for in-home DPN screening among diabetic patients.

### V. CONCLUSIONS

In this paper we have explored methods for evaluating the additional information provided by wearable sensors in early screening for disease. Enabled by a quickly deployable wearable system, we collected gait data from 106 diabetic patients in China, to test whether the ear-worn sensor could provide additional information in early screening for diabetic neuropathy. Using logistic regression models and rigorous evaluation methods, we showed that the gait data collected by wearable sensors do indeed provide extra information that can be combined with the existing clinically adopted surveys, enhancing the discriminative capability of the current clinical methods. Our easy-to-deploy system has the potential to be used for in-home DPN screening among diabetic patients.

### REFERENCES