Université de Montréal

Investigating Intra and Inter-subject Performance with Deep Learning for Gait on Irregular Surfaces

par

Guillaume Lam

Département d'informatique et de recherche opérationnelle
Faculté des arts et des sciences

Mémoire présenté en vue de l'obtention du grade de
Maîtrise en sciences (M.Sc.)
en Informatique avec Option Intelligence Artificielle

31 janvier 2024

Université de Montréal

Faculté des arts et des sciences

Ce mémoire intitulé

Investigating Intra and Inter-subject Performance with Deep Learning for Gait on Irregular Surfaces

présenté par

Guillaume Lam

a été évalué par un jury composé des personnes suivantes :

Aaron Courville
(président-rapporteur)

Philippe C. Dixon
(directeur de recherche)

Irina Rish
(codirecteur)

Sébastien Lemieux
(membre du jury)

Résumé

La médecine personnalisée promet des soins adaptés à chaque patient. Cependant, l'apprentissage automatique appliqué à cette fin nécessite beaucoup d'améliorations. L'évaluation des modèles est une étape cruciale qui nécessite du travail pour amener à un niveau acceptable pour son utilisation avec des participants. Actuellement, les performances sur les ensembles de données biomédicales sont évaluées à l'aide d'un découpage intra-sujet ou inter-sujet. Le premier se concentre sur l'évaluation des participants présents à la fois dans les ensembles d'entraînement et de test. Ce dernier sépare les participants pour chaque ensemble. Ces termes sont respectivement synonymes de fractionnement aléatoire et par sujet. Deux méthodes principales se présentent comme des solutions pour obtenir des performances de franctionnement aléatoires lors d'entraînement de méthodes par sujet, calibration et sans calibration. Alors que la calibration se concentre sur l'entraînement d'un petit sous-ensemble de participant non vues, les méthodes sans calibration visent à modifier l'architecture du modèle ou les traitements préliminaire pour contourner la nécessité du sous-ensemble. Ce mémoire étudiera la calibration non paramétrique pour ses propriétés d'indépendance de la modalité. L'article présenté détaillera cette enquête pour combler l'écart de performance sur un ensemble de données d'essais de marche sur des surfaces irrégulières. Nous déterminons que quelques cycles (1-2) de marche sont suffisants pour calibrer les modèles pour des performances adéquates (F1 : +90%). Avec accès à des essais de cycle de marche supplémentaires (+10), le modèle a atteint à peu près les mêmes performances qu'un modèle formé à l'aide d'une approche de fractionnement aléatoire (F1 : 95-100%). Suivant les objectifs de la médecine personnalisée, des voies de recherche supplémentaires sont décrites, telles qu'une méthode alternative de distribution de modèles qui s'adapte aux étapes de recherche tout en réduisant les coûts de calcul pour les développeurs de modèles. Nous constatons que l'étalonnage est une méthode valable pour surmonter l'écart de performance. Les résultats correspondent aux découvertes précédentes utilisant l'étalonnage pour obtenir des performances robustes.

Mots clés : Médecine personnalisée, Apprentissage Machine, Apprentissage Profond, Évaluation Intra/Inter-Sujet, Fractionnement Aléatoire/Par sujet

Abstract

Personalized medicine promises care tailored to each patient; however, machine learning applied to this end needs much improvement. Evaluation of models is a crucial step which necessitates attention when utilized with participants. Currently, performance on biomedical datasets is evaluated using either intra-subject or inter-subject splitting. The former focuses on the evaluation of participants present in both training and testing sets. The latter separates participants for each set. These terms are synonymous with random-wise and subject-wise splitting, respectively. Two main methods present themselves as solutions to achieving random-wise performance while training on a subject-wise dataset split, calibration and calibration-free methods. While calibration focuses on training a small subset of unseen data trials, calibration-free methods aim to alter model architecture or pre-processing steps to bypass the necessity of training data points. This thesis investigates non-parametric calibration for its modality-agnostic properties. The article presented details this investigation at bridging the performance gap on a dataset of gait trials on irregular surfaces. We determine few (1-2) gait cycles are sufficient to calibrate models for adequate performance (F1:+90%). With access to additional gait cycle trials, the model achieved nearly the same performance as a model trained using a random-split approach (F1:95-100%). Following the goals of personalized medicine, additional research paths are outlined, such as an alternative model distribution method which fits with research steps while reducing computational costs for model developers. We find that calibration is a valid method to overcome the performance gap. The presented results correspond with previous findings by using calibration to achieve robust performance.

Key Words: Personalized medicine, Machine Learning, Deep Learning, Intra/Inter-subject evaluation, Random/Subject-wise split

Table des matières

Table des figures

Liste des sigles et des abréviations

AI	Intelligence Artificiel, de l'anglais Artificial Intelligence
ML	Apprentissage machine, de l'anglais Machine Learning
DL	Apprentissage profond, de l'anglais Deep Learning
CL	Apprentissage Continu, de l'anglais Continual Learning

Chapter 1

Introduction

At the time of writing this thesis, machine learning (ML) in clinical use cases is very much in its infancy. According to Weissler et al. (2021), the frequency of publications relating to ML and clinical research started taking off in 2012. In 2018, just under 100 publications were submitted, doubling in 2019. As programming becomes ever more present in the lives of the population, the power of ML becomes accessible to a more significant number of people. As researchers in all fields see the potential of ML, its usage and subsequent research are only expected to grow.

Adopting ML models in clinical settings will benefit clinicians and patients (Weissler et al. 2021; Deo 2015). Faster diagnosis and treatment would benefit the patients. Off-site systems will help cases where a patient's movement is hindered. Rehabilitation could be performed in the comfort of the patient's home without compromising the quality of supervision. Continuous monitoring in critical times would allow quicker responses for patients needing medical attention. Additionally, help from computing systems would unburden clinicians, allowing for better management of their attention toward patients in immediate need.

However, medicine aided by ML is not without its counterpart of potential danger. Since models potentially impact the diagnosis and treatment of real people, mistakes are extremely costly; several challenges stand in the way of greater adoption (c.f. Watson and al. 2019). Explainable artificial intelligence is a required field for clinical doctors, mainly to have more confidence in the systems deciding patients' future (Obermeyer and Emanuel 2016). Barriers to the adoption of these systems are not to be underestimated. If the friction for usage is too great, researchers and clinicians will favour traditional methods. Focus on making these systems intuitive, easy to adopt, and filling the needs of the platform's users will be crucial (Patel et al. 2008). Most importantly, as mistakes are bound to happen, researchers must put safety nets around the models to catch errors caused for multiple reasons.

The future of AI-augmented medicine is inspiring for its various possibilities. With much work to bring ML to an acceptable level of rigour for clinical use, this work attempts to

advance progress. For assurance and trust in these models, evaluation is needed. Indeed, two modes of assessment present themselves in a repeated-measures study of participants. Recorded trials might be split across both training and testing sets. This method is commonly used in ML theory. In the biomechanics literature using ML methods, it is referred to as random-wise or record-wise splitting or intra-subject evaluation.

Additionally, the participants could be subject to splitting, impacting the distribution of trials. The article presented in this thesis, submitted to the Journal of Biomechanics (Accepted as of April 2023), will resolve the performance discrepancy on a dataset pairing inertial measurement unit measuring gait on various irregular surfaces.

While models trained via random-wise splits perform better on a test set, ML users may resolve the performance gap by calibrating from new individuals. Additionally, this presents itself as the more realistic scenario in the face of model development in research. Integrating model distribution requires less work, allowing public access to the research model development workflow. While feasible, using the former method with the intent of public access, i.e. new participants, requires a computing burden and setup of software infrastructure. Researchers or clinicians utilizing ML methods might need these programming skills. The additional calibration step becomes a step at another level than the model's initial training. It thus becomes embedded in the life cycle of model distribution to users. We may formulate additional life cycle diagrams with hypothesized benefits by reasoning about the process graphically and modelling for diminishing calibration validity.

1.1. Motivation

This section presents the motivation for this work. Broadly, the next step in the medical revolution is available in tandem with ML models, which can run on multiple available devices to people. Following its predecessor, evidence-based medicine, which looks for treatments by tracking the success rates in various populations, personalized medicine aims to tailor treatments for each user. This paradigm fits perfectly with the current availability of devices which can run powerful predictive models.

1.1.1. Problem setting: Irregular Surface Classification

The problem setting chosen for this thesis is the classification of various surfaces based on the collected inertial measurement data from inertial measurement units (IMUs) attached at multiple points on the lower half of the participants. This particular task was chosen as it is essential for numerous reasons.

Throughout an average day, humans walk on highly heterogeneous surfaces. The surfaces vary in complexity (e.g. concrete vs grass/gravel) and incline (e.g. flat vs stairs). The adaptive walking patterns we employ for each surface differ significantly (Grant et al. 2022).

These gait patterns are thus critical biomarkers. Biomechanical clinicians can leverage this complex data to tailor treatment. Furthermore, gait on a subset of surfaces might yield tell-tale signs of potential walking habits, which might lead to future injuries (Moy et al. 2013). Thus, stratifying a patient's daily step count per surface is precious information for an individual's health.

Each person is a highly complex individual, and this gait changes throughout one's lifetime (Sutherland et al. 1980). To accurately characterize individuals (e.g. for their step count), investigation of personalized modelling is crucial.

1.1.2. Personalized Medicine

The newest framework in medicine was preceded by evidence-based medicine. Evidence-based medicine is a systematic approach to medical practice that incorporates the best available evidence from research studies, clinical experience, and patient values when making decisions about diagnosis and treatment. It is a medical practice combining clinical expertise with the best evidence from clinical studies and patient values. This approach is used to ensure that medical decisions are based on the most reliable evidence available and help healthcare practitioners provide the safest and most effective care for their patients. Evidence-based medicine is used in healthcare settings to ensure that treatments and interventions are as effective and efficient as possible.

Several drawbacks are associated with evidence-based medicine. Firstly, evidence-based medicine is limited by the available evidence. Unfortunately, not all medical treatments have been studied extensively, making it challenging to conclude their effectiveness (Möller 2022). Secondly, evidence-based medicine is expensive. Studies must be conducted, and the cost of doing so is often passed on to the patient through higher insurance premiums or out-of-pocket expenses (Straus and McAlister 2000). Next, the data used in evidence-based medicine often needs to be more reliable and complete. This can lead to incorrect conclusions, and the wrong treatment is prescribed (James 2013). Lastly, evidence-based medicine can be seen as an attempt to replace clinical judgment with a set of rules. While evidence-based medicine has benefits, it can also be seen as limiting a physician's discretion and flexibility (White and Taylor 2002).

Personalized medicine aims to decrease these cons while offering different pros. It is a form of healthcare that provides personalized treatments and therapies to individuals based on their unique genetic makeup and health history. This approach to healthcare allows for tailoring treatments to the needs of the individual patient rather than relying on a one-size-fits-all approach (Hamburg and Collins 2010). Personalized medicine is a growing field that is making strides to improve the quality and efficiency of healthcare. Its use of genetic and genomic data allows researchers to understand the underlying causes of disease better and

to develop more effective treatments with fewer side effects. By utilizing this approach, it is hoped that patients will have better outcomes and improved quality of life.

Utilizing this different paradigm comes with various benefits. Firstly, personalized medicine can help doctors tailor treatments to an individual's genetic makeup, improving treatment outcomes (Offit 2011). Secondly, by providing the proper treatment for the right person, personalized medicine can help reduce inefficiencies in the healthcare system (Ahmed et al. 2020). Additionally, by better understanding a person's genetic profile and risk factors, personalized medicine can provide predictive power to help doctors make more informed decisions (Drake, Cimpean, and Torrey 2022). Next, by better targeting treatments and reducing inefficiencies, personalized medicine can help save money and lower care costs (Jakka and Rossbach 2013). Finally, by assisting people in receiving the right treatments, personalized medicine can improve many patient's quality of life (Phillips et al. 2014).

The adoption of personalized medicine also comes with its share of drawbacks. First, personalized medicine raises ethical questions about using its information (Chadwick and O'Connor 2013). Second, despite the promises of personalized medicine, its results' accuracy is only sometimes reliable (Volm and Efferth 2015). Last, while personalized medicine is still in its early stages, there is a lack of evidence to show that it is more effective than conventional treatments (Meckley and Neumann 2010; Ginsburg and Kuderer 2012; Conti et al. 2010).

Hopefully, this chapter has clarified that the focus on individualized and calibrated models fits the vision of personalized medicine. The usage of ML for personalized medicine-oriented goals is a fruitful path.

1.2. Objectives

The following objectives accompany this thesis. Personalized medicine still needs further research to marry with ML properly. The first aim of this thesis is to investigate the evaluation of ML models faced with biomarker datasets. Due to different priors, several steps in the model development pipeline must be adapted. The second aim is to bridge the performance gap between intra-subject and inter-subject evaluation. This gap, if not known, leads ML utilizers to overestimate the performance of their models when encountering new participants.

This thesis may guide clinicians and researchers to utilize ML for purposes of personalized medicine appropriately.

1.3. Literature Review

This chapter presents various concepts researchers may use to solve the gap between random and subject-wise splitting. As it is a multi-faceted problem, different methods have been devised to resolve the disparity in performance between the two splitting techniques.

To obtain high performance on prediction, two main methods exist. In supervised learning, the first method observes combinations of inputs and outputs while optimizing parameters. Exposing a model to these data points reduces the error in prediction for the various categories.

It is, however, unrealistic to expect a single training phase for most deep learning applications. Suppose new data has symmetrical properties, e.g. repeated-measure studies, label similarities, existing label subsets, etc., to the previously seen data. In that case, the optimization process will require fewer data points while achieving similar performance on previous axes of symmetry, e.g. per participant performance, per label performance, etc. (Charoenphakdee, Lee, and Sugiyama 2021). Calibration thus allows us to gather data to train a model they deem helpful. After, this model can be distributed and repurposed for relatively similar use cases while leveraging all previously seen data points.

While one avenue calibrates models, the other focuses on calibration-free methods. These methods generally avoid the need to gather additional data. This allows a model to be distributed while necessitating no additional training for strong prediction of new data. The general approach of these methods is to embed the symmetries of encountered data with architectural priors or preprocessing steps. By having a particular model architecture or preprocessing steps, e.g. abdomen imaging (Tomi-Tricot et al. 2019), blood pressure (Kachuee et al. 2015), segment positioning (Yang et al. 2022), the labels are predictable for all individuals, seen or unseen.

The rest of this chapter presents the individualized calibration methodology used in the article of chapter 2. The following section introduces some recent calibration-free methods to gain a contrasting view.

1.3.1. Calibration Methods

Recent advances in machine learning have enabled the development of individualized calibration techniques. Training the model on an individual's data and testing it on new data is adjusted to fit the individual data better and then tested on the latest data to determine its performance. This process is repeated until the model's performance meets the desired accuracy. The method of individual calibration is beneficial for optimizing the performance of an ML model for an individual, as it allows the model to fit the individual's data better and thus improve its performance (Bol and Hacker 2012).

This section reviews individualized calibration, including both parametric and non-parametric approaches. We discuss different techniques' potential advantages and drawbacks and possible future research directions.

Parametric approaches for individual calibration involve fitting a model with theorized or known parameters to an individual's data. Researchers can use techniques, including least squares regression, generalized linear models, and Bayesian methods.

They generally require fewer data for each individual (Dawkins, Srinivasan, and Whalley 2001) but can be computationally expensive (Tolson and Shoemaker 2007; Huot et al. 2019). The main advantage of these methods lies in the high explainability of these methods. Since these methods' parameters and their relations are explicitly provided, understanding these techniques is generally more accessible. Unfortunately, an underlying theory and prior are needed to filter the many parameters present.

Non-parametric approaches for individual calibration involve using models where the direct relation between the learned problem space and the target is not necessarily known. As such, more variables can be provided to the model in the hope that the model will highlight predictive from non-predictive parameters. Examples of such models include decision trees, random forests, and neural networks.

These models are usually more accurate than parametric models but require more data for each individual. Additionally, they are challenging to interpret as the models generally have an increased complexity(Q. Zhang, Wu, and Zhu 2018).

Calibration is viable for multiple data types of biomedical literature. Transfer learning may be used for brain-computer interface decoding models (Khazem et al. 2021; Vidaurre et al. 2011). Khazem et al. (2021) successfully reduced calibration set size by selecting the minimal dataset needed for training a priori. Calibration depends on the data type to inform the number of trials. Cano et al. (2022) apply calibration to cardiovascular data to obtained a 30% increase in F1-score utilizing one trial. Lehmler et al. (2021) gain 35% in F1-score using five gait cycles of electromyography data.

Individual calibration can improve existing calibration techniques, increase the accuracy of calibration techniques, and reduce the computational cost of existing techniques. Potential applications of individual calibration cover areas such as healthcare.

1.3.2. Calibration-Free Methods

An individual calibration-free summary is a machine learning model that does not require users to adjust each data point for accuracy manually. This model assumes that unique data points should be treated equally, without assumptions about handling each data point. Instead, the model uses algorithms to automatically analyze and process each data point to create an overall summary. This model offers an efficient way to process large datasets and can be used to generate summary statistics and visualizations (Yongle Luo et al. 2021).

Han et al. (2020) utilize raw heart rate signals processed by convolutional neural networks in combination with body mass index information to estimate blood pressure and hypertension class with no initial calibration. Kwon et al. (2019) analyze electroencephalography data to develop and subject-independent method with better results than subject-dependent models.

Among other fields, individual calibration-free models have been used in marketing (Campuzano et al. 2020), finance (Idili et al. 2019), and healthcare (Mirshekari, P. Zhang, and Noh 2017). This model can be used in marketing to identify customer segments and develop targeted campaigns. In finance, the model can be used to analyze financial data and detect patterns in the market. In healthcare, clinicians can use this model to identify trends in patient health and treatment outcomes.

1.4. Sections

Firstly, chapter 2 presents the article. The methodology presented in the article demonstrates the correct evaluation in light of participants yet to be seen. Secondly, chapter 3 presents an in-depth discussion, which will dive deeper into different hypothesized methodologies in line with the personalized medicine vision. Finally, chapter 4 provides concluding remarks related to the thesis and suggests future research directions.

Chapter 2

Article

Currently, most utilized ML models follow the basic and traditional pipeline. Participants gather data according to a protocol to acquire information of interest. The data is then labelled as a training target. Various models are then deployed and trained on the features and labels of this dataset. Some data is kept for training, while the rest is saved for validation and testing. This splitting follows the random-wise approach, whereby trials are divided randomly across both sets regardless of participant association. The model with the best performance on the test dataset is kept and optimized. Finally, the deep learning model is distributed and used to predict the possible label for new participants given their gathered features.

The previously mentioned pipeline has advantages. Models generated are generally robust as the training and testing are done in a broader distribution. For a community not centred around machine learning and computer science, this single powerful model is beneficial and practical as the single model is distributed.

Unfortunately, disadvantages are present. Firstly, unlike a food classifier, the rate of new participants' data increases substantially. Thus, these models need to be re-trained at an alarming rate, which is unsustainable for laboratories that need access to immense computing devices. Additionally, transporting this amount of data is a task in its own right. Unique to the clinical setting, there is also the possibility that new and rare diagnoses will appear. Secondly, this evaluation framework is unlike the settings customarily encountered. It is most likely that new participants are continuously being exposed to the model. Evaluation and assessment of models should be done with this use case in mind. As such, models generated from the pipeline are implicitly expected to predict equal performance on the testing data and the new participants. Shah et al. (2022) showcase this overestimation of performance when indicating surface types. Shah et al. contrast the random-wise split with the subject-wise split. This splitting scheme reserves some participants for the test set exclusively. This contrast exposes the gap in performance of models trained with a random

vs subject-wise split (F1-scores of 0.96 vs 0.78, respectively). Therefore, most prediction models encountering datasets sampled from participants lead deep learning users to believe their model will perform better than expected.

A method to recover random-wise performance on a subject-wise split-trained model is to expose it to samples of the unseen participant. This is known as calibration. As stated previously, this technique is understood and used in various biomedical domains. The calibration dynamics are, however, very varied and dependent on the dataset.

In the presented article, calibration will be observed using a biomedical dataset. More precisely, a biomechanical dataset is employed. While this does not cover all use cases, it is a valid template for various datasets which may be encountered in clinical settings. As a stepping stone, finding the minimum calibration will allow for extrapolation to training schemes that can balance use cases ranging from single calibration to continuous calibration (chapter 3).

The following section of the chapter will be the article as a part of the thesis. The article presents a non-parameterized individualized calibration solution to the gap problem applied to a biomechanical gait dataset when travelling on irregular surfaces. This article has been accepted in the Journal of Biomechanics.

Estimating individual minimum calibration for deep-learning with predictive performance recovery: an example case of gait surface classification from wearable sensor gait data

2.1. Abstract

Clinical datasets often comprise multiple data points or trials sampled from a single participant. When these datasets are used to train machine learning models, the method used to extract train and test sets must be carefully chosen. Using the standard machine learning approach (random-wise split), different trials from the same participant may appear in both training and test sets. This has led to schemes capable of segregating data points from a same participant into a single set (subject-wise split). Past investigations have demonstrated that models trained in this manner underperform compared to those trained using random-split schemes. Additional training of models via a small subset of trials, known as calibration, bridges the gap in performance across split schemes; however, the amount of calibration trials required to achieve strong model performance is unclear. Thus, this study aims to investigate the relationship between calibration training set size and prediction accuracy on the calibration test set. A database of 30 young, healthy adults performing multiple walking trials across nine different surfaces while fit with inertial measurement unit sensors on the lower limbs was used to develop a deep-learning classifier. For subject-wise trained models, calibration on a single gait cycle per surface yielded a 70% increase in F1-score, the harmonic mean of precision and recall, while 10 gait cycles per surface were sufficient to match the performance of a random-wise trained model. Code to generate calibration curves may be found at (https://github.com/GuillaumeLam/PaCalC).

2.2. Introduction

Deep learning has proven helpful in numerous areas. Naturally, practical issues arise in clinical settings due to the novelty of applying machine learning tools to these fields (Miotto et al. 2018; Tobore et al. 2019; Wang et al. 2022; Zemouri, Zerhouni, and Racoceanu 2019). Handling repeated trials from participants is one such issue.

Generally, training and testing sets are derived for training a model and assessing its generalizability on unseen data points. As such, data for testing must be kept isolated from the training set to avoid contaminating the model with information that should not be available ("data leakage").

When facing a context where multiple data points or trials may be associated with a single participant, how to split the dataset remains to be determined. For example, in a

gait study, participants may be asked to perform walking tasks numerous times. Should different trials from the same participant be present in the training and testing datasets, known as a "random-wise"/"record-wise" (intra-subject) split? Or, should all trials from a single participant appear only in either subset, respecting the principle of unseen data points, known as a "subject-wise" (inter-subject) split?

This issue is an ongoing discussion (Saeb et al. 2017) and a debated one (Cao 2022; Little et al. 2017). As shown by Shah et al. (2022), in the context of surface identification from gait data across multiple participants and trials, evaluation using a random-wise approach led to an over-estimation of the predictive power of models compared to a subject-wise split (F1-scores of 0.96 vs 0.78, respectively). Thus, models trained with some of a participant's data outperformed those completely naive to the participant. The high evaluation performance of random-wise trained models may lead to overly optimistic assumptions that the model could achieve the same performance when deployed on new, unseen participants.

Training a model on a small sample of data from a new participant, commonly called calibration or transfer learning, may bridge the gap, i.e., achieve an acceptable model performance without overfitting to participants. Calibration is generally understood as training on a primary dataset and then re-training on a more specific/different dataset to maximize performance. These techniques are applied to a variety of data in the biomedical literature. Khazem et al. (2021) and Vidaurre et al. (2011) used transfer learning to brain-computer interface decoding models. Furthermore, Khazem et al. (2021) successfully reduced calibration training set size by selecting the minimal dataset needed for training a priori; however, the number of calibration trials appears to depend on the data type. Cano et al. (2022) obtained a 30% increase in F1-score utilizing one trial of cardiovascular data; while Lehmler et al. (2021) achieved an increased of 35% in F1-score using five gait cycles of electromyography data.

The actual behaviour of calibration and the impact of the number of calibration training trials remains unknown. Thus, this paper aims to investigate the relationship between the number of calibration trials and the prediction accuracy of the corresponding calibrated deep learning model in a classification study based on biomechanical data. We hypothesized that a model's performance would increase with the number of calibration trials, eventually achieving the same performance as a model trained using a random-wise splitting approach. This research could inform on calibration training set sizes and behaviour for models based on data with multiple trials per participant.

2.3. Methods

2.3.1. Dataset

The current research used the public database of Y. Luo et al. (2020). This database was gathered from 30 young, healthy adults (15 females; 23.5 ± 4.2 years) as they performed walking trials across 9 outdoor surfaces while fit with inertial measurement unit (IMU) sensors (Xsens Awinda, Enschede, The Netherlands). The sensors collected acceleration, angular velocity, and magnetometer data at 100 Hz. They were located at the wrist, fifth lumbar vertebra, anterior thigh (bilaterally), and anterior shank (bilaterally) (Fig. 2.1). The outdoor surface types were: flat-even, slope-up, slope-down, stairs-up, stairs-down, cobblestone, grass, banked-left, and banked-right (Fig. 2.2). Further information on the dataset is available in Y. Luo et al. (2020). In line with the results of Shah et al. (2022), the wrist sensor, which did not provide a meaningful contribution to model performance, was not used herein. As such, our supervised machine learning problem setting classified the surface type based on gait data gathered from the 5 IMUs.

2.3.2. Data Processing

Raw IMU data were passed through a 4th-order Butterworth low pass filter (6 Hz cut-off) to smooth the signal. Following the validated approach of McGrath et al. (2012), gait cycles were delineated using an adaptive gyroscope-based algorithm, which utilizes the angular velocity about the y-axis to calculate robust gait events. All found gait cycles were then normalized to 101-time points.

The processed data were divided into training and testing sets using random and subject-wise approaches. In random-wise splitting, 10% of all gait cycles appeared in the test set, regardless of the associated participant. In subject-wise splitting, gait cycles were separated such that all gait trials from $n = 3$ participants of the total 30 participants (10%) appeared in the test set (Fig. 2.3).

The attributes used from the original dataset include three-dimensional acceleration, angular velocity, and magnetometer data. Based on the work of Shah et al. (2022), an additional plane was generated from these three planes' magnitude. These attributes were fed into a linear discriminant analysis (LDA) for feature extraction. Including the magnitude input, the passing of pre-processed data to LDA resulted in 32 features per sensor. Applying this process to acceleration, angular velocity, and magnetometer data resulted in 96 extracted features. The LDA feature extraction was performed using scikit-learn (Pedregosa and al. 2011) with the single value decomposition solver, no shrinkage, and a 0.0001 solver threshold. More details on the feature extraction process can be found in Shah et al. (2022).

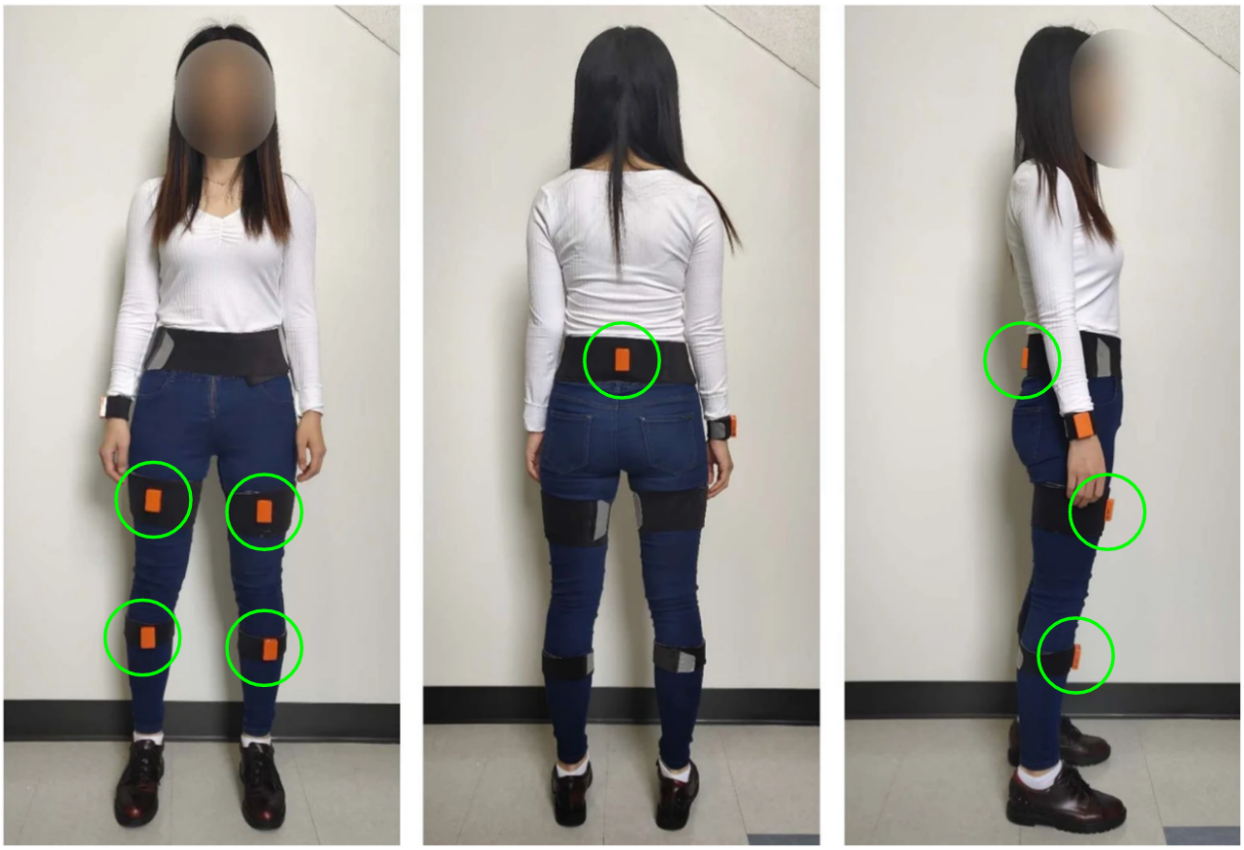

Figure 2.1 – Sensor locations from the Y. Luo et al. (2020) database. Sensors used in the present study and the surface classification work of Shah et al. (2022) shown with green circles: fifth lumbar vertebra (lower back), anterior thigh (bilaterally), and anterior shank (bilaterally).

The first deep learning surface classification model followed the feedforward neural network (FFNN) architecture presented by Shah et al. (2022). Developed in Keras (Chollet et al. 2015), the network comprised three hidden layers: a layer of 606 units, 303 units, and 606 units. All of these layers had a ReLU activation function. The final output layer contained 9 units representing each label and a softmax activation function. The model was trained with an Adam optimizer (learning rate of 0.001) and a categorical cross-entropy loss function. During training set optimization, forward and backward passes were looped over 50 epochs with batch sizes of 512 data points (gait cycles) for random and subject-wise splits. Furthermore, a convolutional neural network (CNN) was also employed to assess the robustness of results to model architecture; details of the model can be found in the supplementary material (Fig. 2.6). Overall, the CNN had substantially more parameters than the FFNN (1,536,817 and 665,094, respectively). Thus, a stronger focus was brought to the model performing better given their parameter count.

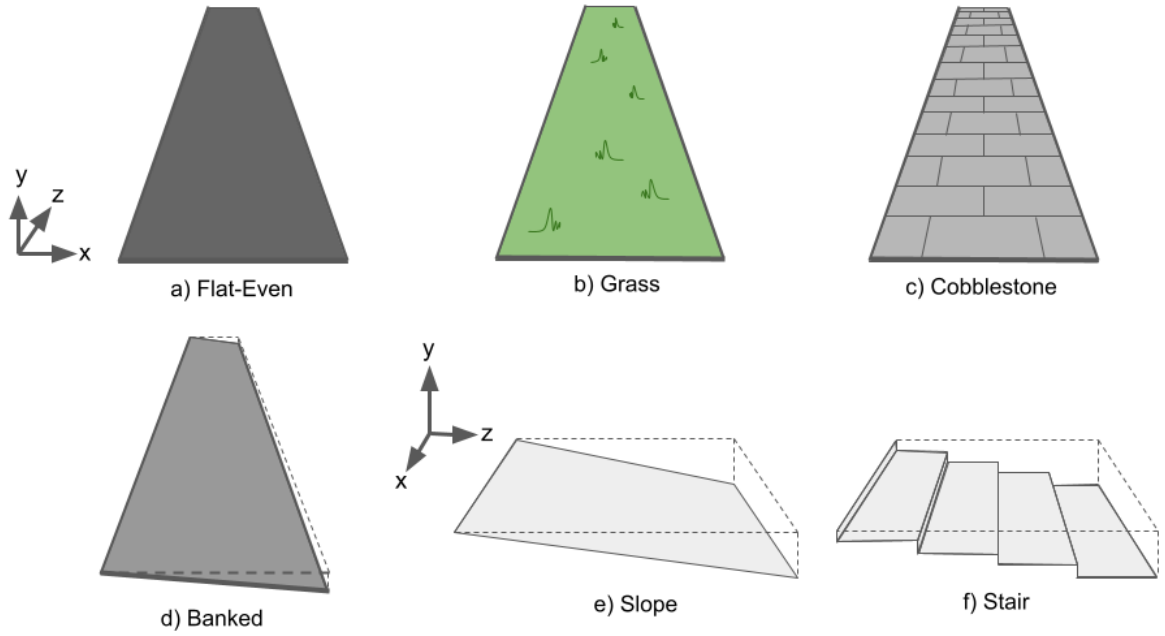

Figure 2.2 – The irregular surface dataset labels from the Y. Luo et al. (2020) database. Outdoor surfaces types: a) flat-even, b) grass, c) cobblestone, d) banked (right/left) e) slope (down/up), f) stairs (down/up).

Two versions of the dataset were generated following each splitting scheme. The training with random-wise split produced the model referred to here as the "random-wise prior" model. This model was evaluated, and the performance was reported in red (Fig. 2.4, 2.5). Similarly, the "subject-wise prior" model was evaluated with its corresponding testing set. The performance was reported in green (Fig. 2.4, 2.5). From the subject-wise test set, two sets were derived: the calibration training set (C_{Tr}) and the calibration test set (C_{Te}) (Fig. 2.3 b.). The calibration test set was fixed to 50%. The calibration training set was initially empty. The subsequent evaluation by the subject-wise prior model yielded the baseline performance. Single gait cycles per label were added to the calibration training set with evaluation in between to generate the calibration curve shown in blue (Fig. 2.4, 2.5). A fourteen-fold validation was performed across the subject-wise split train and test set (Fig. 2.3 a.). This cross-fold evaluated robustness and heterogeneity by observing calibration across different participants. Thus, calibration was conducted with a batch size of 1. Model performance was evaluated using the F1-score. An early stop was added to stop the calibration of models that obtained an F1-score of 1.0 for seven consecutive evaluations. Finally, the calibration models were reset between participants to not influence the calibration curves.

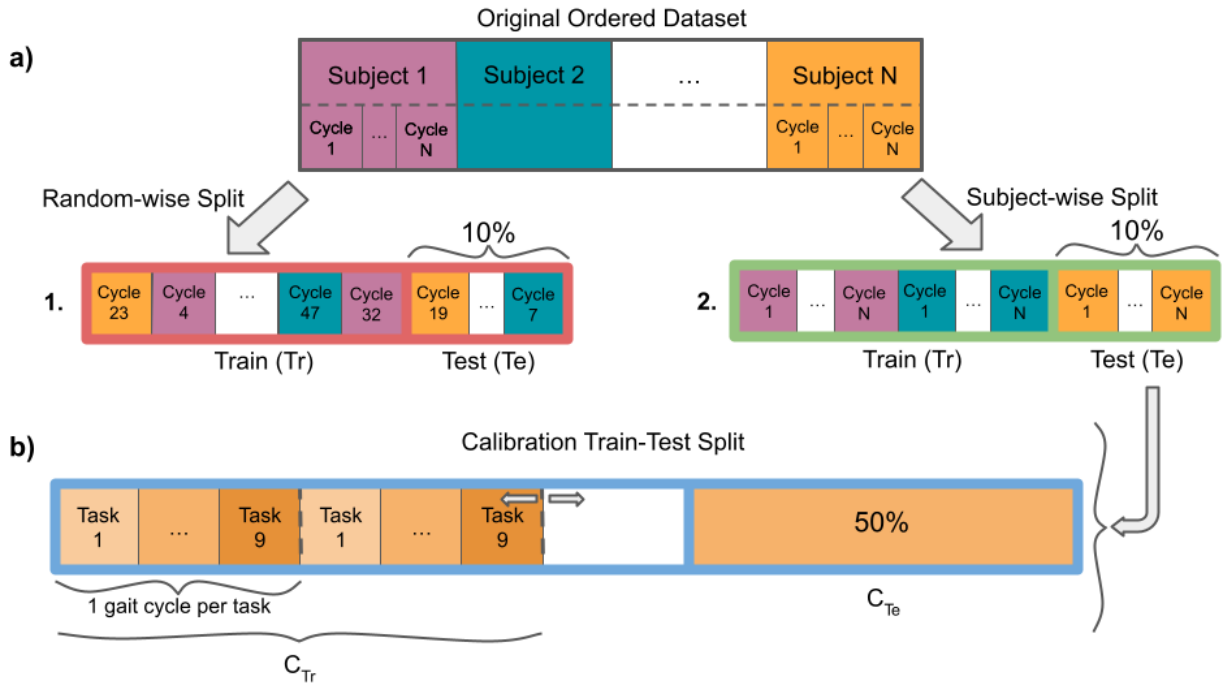

Figure 2.3 – a) Data splitting schemes: 1. Random-wise: splits total gait cycles uniformly 2. Subject-wise: splits gait cycles such that some participants only appear in the test set b) Derivation of calibration sets from the test set of a subject-wise split. The calibration test set (C_{Te}) is set to 50% of a participant's gait cycles per label. The calibration train set (C_{Tr}) grows progressively larger with the addition of balanced gait cycles per task.

2.4. Results

The calibrated models showed full recapture of performance, as measured by the F1-score while accessing 1-2 gait cycles of previously unseen participants (Fig. 2.4). Calibration with a training set of size $\geqslant 10$ was equivalent to training on a random-wise split based on the F1-score. A calibration training set size of 1 per label recovered an F1-score of $95 \pm 3\%$ (FFNN) while a calibration size of 10 per label obtained a virtually perfect F1-score.

A discrepancy was noted between the performance of a model trained solely on the subject-wise training set and the calibrated model with no training. As observed, the F1-score of the FFNN before calibration was $80 \pm 14\%$ while the calibrated FFNN trained on no gait cycles scored $19 \pm 12\%$.

Most surface types followed the trend of near 1.0 F1-score recovery with exposure to at least 10 gait cycles (Fig. 2.5). The surfaces cobblestone, flat-even, slope-up, and stair-down all converged to an F1-score of 1.0 predictably; however, banked-left, banked-right, grass, and slope-down did so with a varying number of gait trials. In contrast, stair-up never quite recovered an F1-score of 1.0 for some participants, as demonstrated by the high standard deviation.

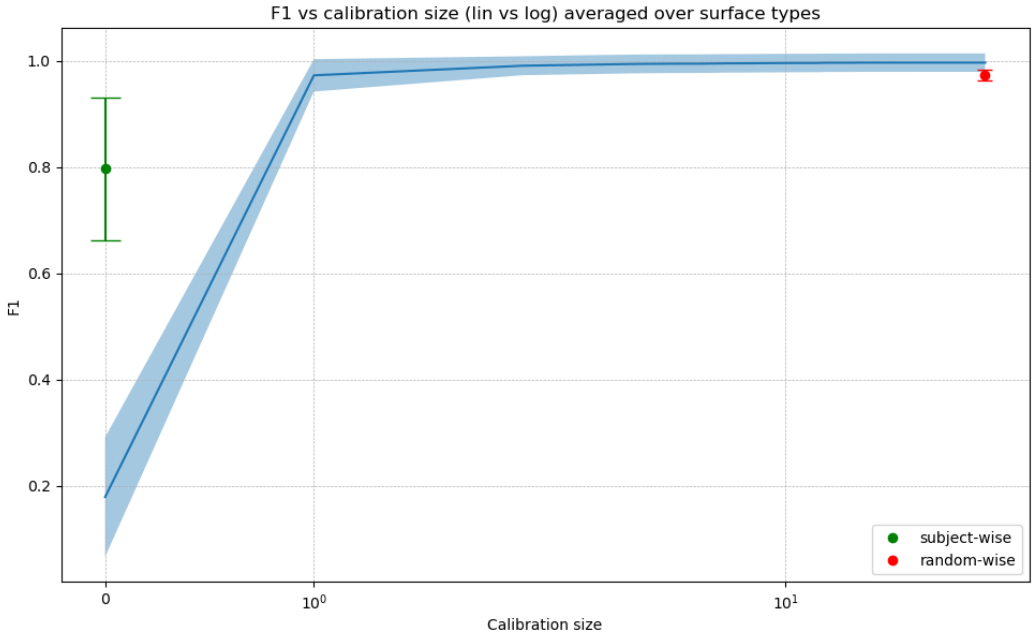

Figure 2.4 – Effect of calibration train set (C_{tr}) size (log scale) on the F1-score. Subject-wise mean performance with standard deviation is shown as a green dot and error bars, respectively. Random-wise mean performance with standard deviation is shown as a red dot and error bars, respectively. The mean performance of the calibration is show in dark blue with 1 standard deviation shown in light blue.

Furthermore, the discrepancy of the models before calibration was different across surface types. The calibrated FFNN with 0 gait trials scored an F1-score of $0\pm1\%$ for slope-up/down, grass, and flat-even. An F1-score of $5 \pm 1\%$ was initially obtained for stair-up. F1-scores of $25 \pm 17\%$, and $25 \pm 5\%$ were obtained for stair-down and cobblestone, respectively. Finally, the banked surfaces achieved approximately 0.5 F1-score with no training. Scores of $55\pm50\%$ and $48 \pm 40\%$ were achieved for the banked surfaces, respectively.

Overall, the CNN achieved similar results as the FFNN with greater variance and obtaining an F1-score of $5 \pm 6\%$ with no training on the cobblestone surface. Figures for the results of the CNN on the irregular surface task may be found in the supplementary material (Fig. 2.7, 2.8).

2.5. Discussion

This project assessed the impact of the number of calibration trials on the performance of a deep-learning surface classifier based on gait data from inertial sensors. The calibrated models solved the task in general, given a single trial per label. With access to additional gait cycle trials, the models achieved nearly the same performance as a model trained using

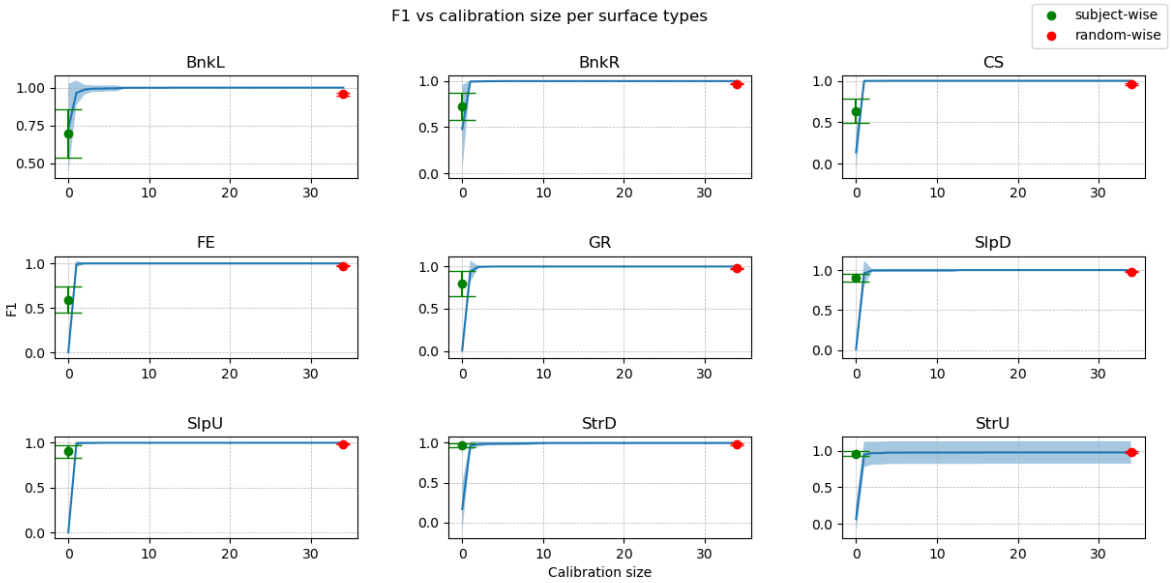

Figure 2.5 – Effect of calibration train set (C_{tr}) size on F1-score for each surface: a) banked-left, b) banked-right, c) cobblestone, d) flat-even, e) grass, f) slope-down, g) slope-up, h) stairs-down, i) stairs-up.

a random-wise split approach. These results illustrate the impact of calibration on model performance.

Our results confirmed our hypothesis. Overall, an increase in the number of calibration trials improved model performance. Ten gait cycles per surface were sufficient for mastery of surface type prediction in the current data set for both tested models. We also generated calibration curves per label to inform on the models' needed calibration trials and the relative accuracy return of each additional trial.

Calibration is a practical approach to solving the performance gap of a subject-wise split, compared to random, for deep learning models exposed to repeated measures of human subject-based biomedical datasets. Further supported by the findings of Shah et al. (2022), as some bio-patterns of a label are potentially more variable across patients than across all the chosen labels, calibration is an integral and necessary step for some datasets. This paper highlights that a calibrated subject-wise prior deep learning model can match random-wise prior deep learning model performance while needing exposure to only a fraction of the total number of a participant's trials.

The present research is mainly limited by the range of biomedical datasets, types of deep learning models, and machine learning settings tested: a single data set, two deep learning architectures, and a single supervised classification setting. This limitation has led to a very narrow search of the range of calibration dynamics. Nonetheless, the small number of trials required for gains in performance observed in our present work is consistent with similar

literature (Cano et al. 2022; Khazem et al. 2021; Lehmler et al. 2021; Vidaurre et al. 2011). Our methodology can also be transferred to different classification settings Bird et al. 2021; Ching et al. 2018. or applied to regression problems (Caywood et al. 2017; Davis et al. 2017).

Future investigations may focus on the behaviour or properties of the calibration curve. For example, a model could require smaller amounts of calibration to achieve similar performance results, increasing the sample efficiency. Additional investigations might target high performance on many participants as current deep learning models are susceptible to "catastrophic forgetting", where the model fails tasks previously known (Kirkpatrick et al. 2017). The machine learning field, continual learning, aims at solving this phenomenon. Finally, the frequency of calibration of an individual to maintain high performance needs to be explored. An individual's gait may change over time enough not to be recognized, necessitating a re-calibration.

Deep learning models can be calibrated for strong prediction on participants not known during training (previously unseen participants). Based on our results of calibration dynamics, a minimum calibration training set size greatly benefited the performance of deep learning in this context. By observing the calibration dynamics, researchers are subject to learn about the interactions between their datasets and their models. Developers of real-world applications should be aware of the benefits of model calibration.

2.6. Code availability

Python functions are supplied on Github to train the models presented herein and generate participant calibration curves (PaCalC) (https://github.com/GuillaumeLam/PaCalC). This code could be adapted to help users discern the number of calibration trials required for their given application.

2.7. Acknowledgements

We acknowledge the support of the Natural Sciences and Engineering Research Council of Canada (NSERC), [funding reference number RGPIN-2022-04217] and the Fonds de recherche du Québec–Santé (FRQS) Bourses de chercheurs-boursiers et chercheuses-boursières Junior 1 (Volet Santé-technologie).

2.8. Supplementary Material: Figures

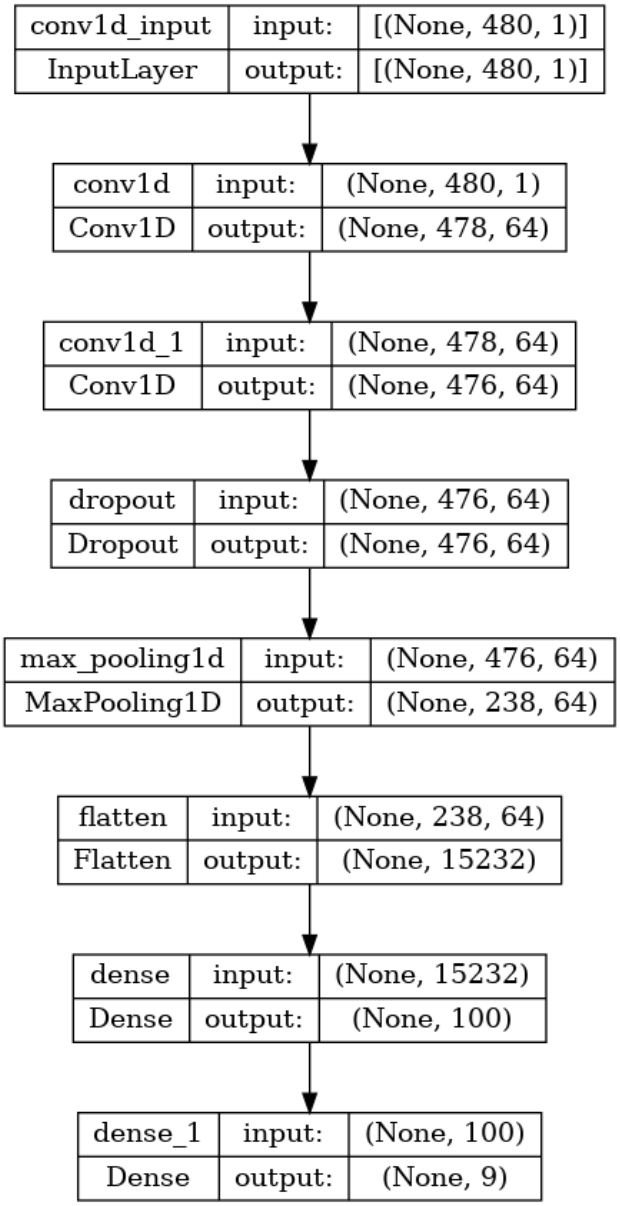

Figure 2.6 – Architecture of the convolutional neural network (CNN) used for the classification of irregular surfaces. The first two Conv1D and the second to last Dense layer used a ReLU activation function. The final Dense layer used a softmax activation to convert the inputs into normalize probabilites for all labels. Similar to the feedforward neural network (FFNN) used in this paper, the CNN model was trained with an Adam optimizer with a learning rate of 0.001 and a categorical cross-entropy loss function

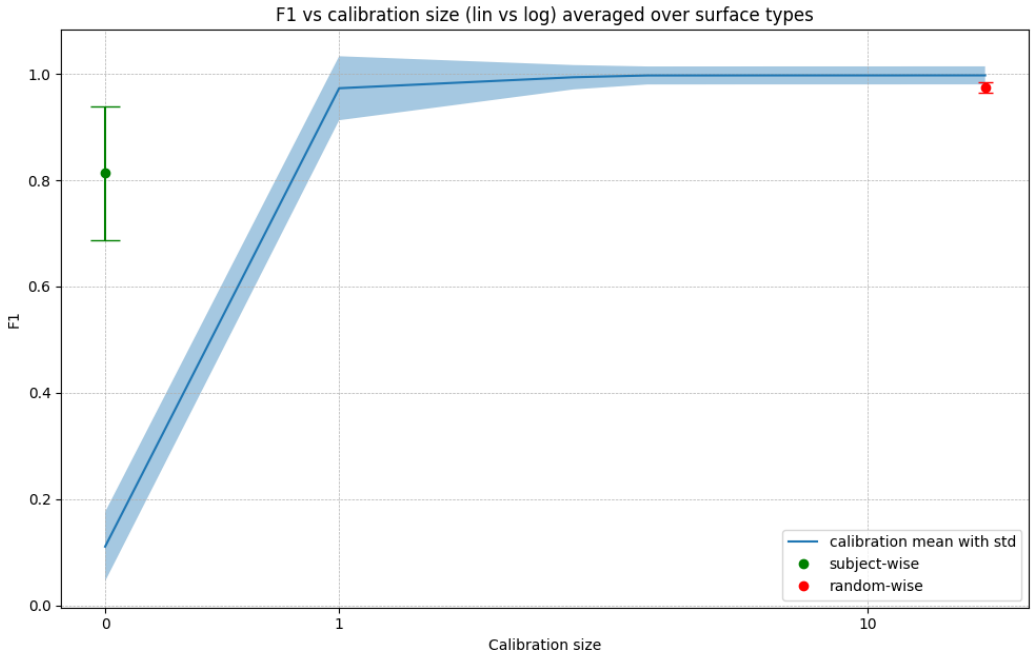

Figure 2.7 – Effect of calibration train set (C_{tr}) size (log scale) on the F1-score for the convolutional neural network architecture.

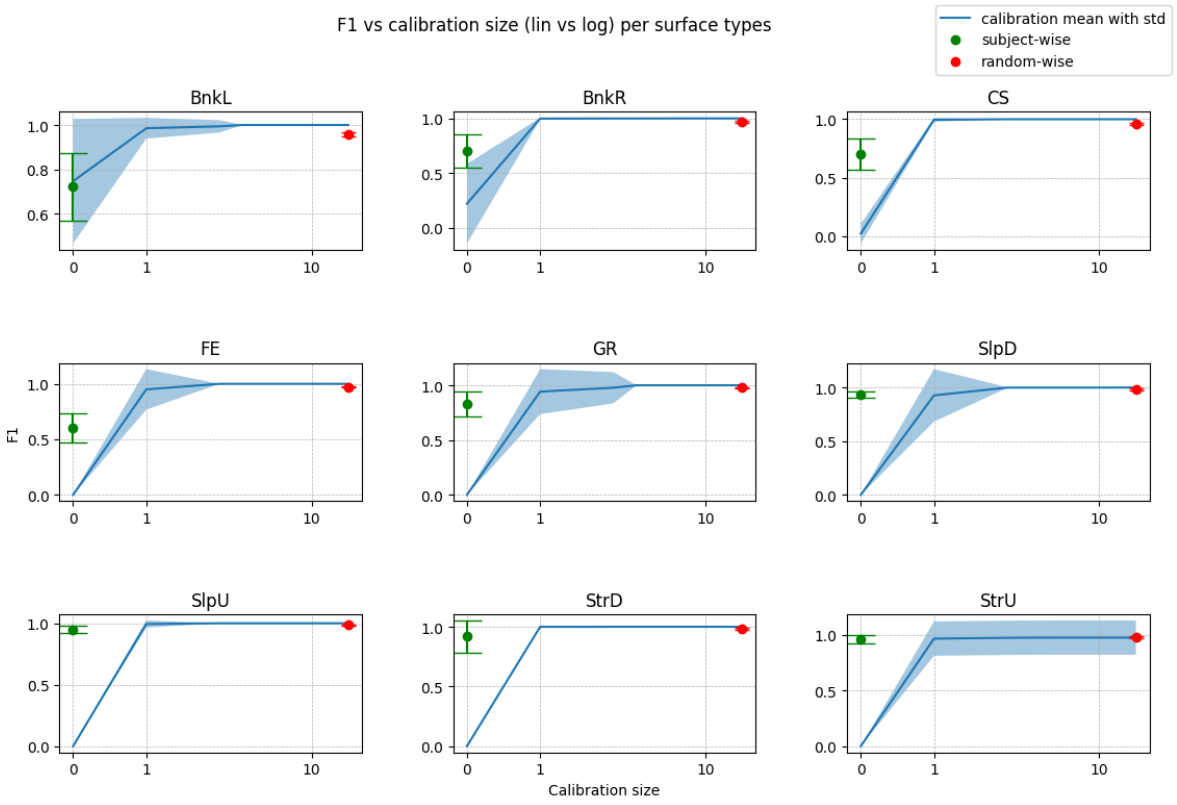

Figure 2.8 – Effect of calibration train set (C_{tr}) size on F1-score for each surface for the convolutional neural network architecture: a) banked-left, b) banked-right, c) cobblestone, d) flat-even, e) grass, f) slope-down, g) slope-up, h) stairs-down, i) stairs-up.

Chapter 3

Discussion

3.1. Significance of Work

Personalized medicine, in conjunction with machine learning (ML), has the potential to help people. Work will be needed to improve it to an acceptable level for usage with less strict supervision. To contribute to this endeavour, this project investigated the evaluation of models facing the assessment of unseen people. This project thus aimed to assess the impact of the number of calibration trials on the performance of a deep learning surface classifier based on gait data from body-worn inertial measurement units (IMUs).

The article "Estimating individual minimum calibration for deep-learning with predictive performance recovery: an example case of gait surface classification from wearable sensor gait data," presented in Chapter 2, demonstrates how prediction discrepancy between inter- and intra- subject performance could be overcome with calibration. The minimum calibration per label and subject is calculated while recovering intra-subject optimization level performance. Minimum calibration informs on the complexity of mastery of label prediction while giving a range for future collection.

The calibrated model generally solves the task, given a single trial per label. From the utilized dataset (Yongle Luo et al. 2021), it was determined that few (1-2) gait cycles are sufficient to calibrate models for adequate performance (F1:+90%). With access to additional gait cycle trials, the model achieves nearly the same performance as a model trained using a random-split approach. Overall, ten gait cycles allow for mastery of the task (F1:95-100%). These results illustrate the impact of calibration on model performance. Additionally, we find that some labels require more calibration than others. Most labels are initially predicted with a much lower F1-score than would be led to believe with inter-subject prediction F1-score; calibration might be much more helpful than initially thought. In most calibration curves, the F1 performance of the subject-wise prior model does not match the performance of models with a calibration training set size of 0. Due to a lack of time, a theory has yet to

be postulated. The code was thoroughly inspected for data leakage, and none was present. Further research and investigation will be needed to solve this odd behaviour.

Our results confirm our hypothesis. Overall, an increase in the number of calibration trials improved model performance. Ten gait cycles per surface are sufficient for mastery of surface type prediction for the irregular surface dataset. We also generated calibration curves per label to inform the model's needed calibration trials and the relative accuracy return of each additional gait cycle.

3.2. Limitations

The main limitation of this project is the breadth of combinations of model types, datasets, and problem settings. A single public dataset (Yongle Luo et al. 2021) was investigated. Furthermore, two deep learning model architectures were tested (Shah et al. 2022). Additionally, only classification problems were observed. While few calibration dynamics were explored, our findings agree with those of the literature and are transferable to other settings. Overall, a few trials can provide significant gains in performance.

3.3. Future Directions

The present work has focused on a specific use case and dataset. Multiple improvement paths are current as we try to follow the vision of personalized medicine. The final section of this chapter focuses on relevant future directions worth investigating. The model distribution methodology will present an alternative ML model distribution schema. This schema has multiple benefits, including added simplicity and reduced computing burden for the model distributor. Next, schemes are devised to deal with exceptional cases where calibration may or may not last forever. A spectrum ranging from single calibration to continuous calibration is shown.

3.3.1. Model Distribution Methodology

As many models are evaluated with an intra-subject scale, an implication of a single model is suggested. This leads to following the distribution of the monolith prediction model. These methodologies solve the gap problems by exposing sufficient data points from the initial round of training. Since most models are not made widely available as they are more proof of concept, the monolith models are not a genuine issue. When these models are desired to interact with participants, the traditional model distribution incurs multiple disadvantages.

The first step is to use the data gathered to train the model. Next, when new participants are presented, the latest data is added to the sum of collected data. By training on this more extensive set, new patients are well predicted by the model. It can then be supplied

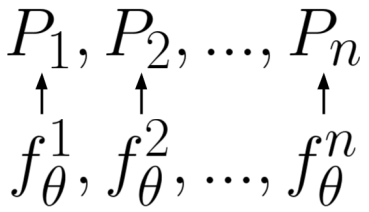

Figure 3.1 – Current model distribution schema. Monolith model f_θ is trained, distributed, and made available to a set of users P. Additional data from new user P_x is sent back for the model to train upon, generating f_{θ_x}.

to the participants to expect more of their collected information. This schema is presented in figure 3.1, where P_x are participants with their associated trials to calibrate the models f_{θ_x}. Several negative traits are present in this architecture. Firstly, there is a heavy computational burden on the side of the suppliers of the predictive model. If the base data set is represented by P_m, where there are m participants, the model f_{θ_1} will need to be trained on $P_{new} = P_m \cup P_1$. The model constantly retrains on previously computed points as more participants are added. For smaller teams, this computation burden might be too significant. Secondly, this architecture necessitates the management of new participant data. In cases where sensitive data is concerned, security becomes an additional requirement. Lastly, a software infrastructure is needed to support the distribution and gathering of data. The researchers or clinicians utilizing ML methods might not possess the required programming skills.

3.3.2. Alternative Participant Model Distribution

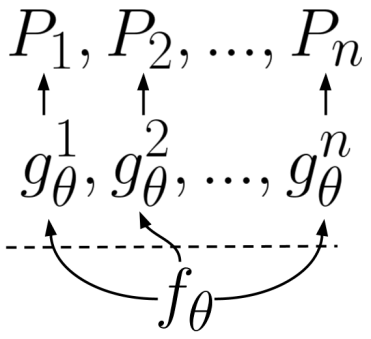

Figure 3.2 – Alternative model distribution schema. Base model f_θ is duplicated as model g_θ. The duplicated model is calibrated for participant P_x to generate an individualized model g_θ^x.

To mitigate the drawback previously stated, an alternative distribution scheme is proposed. The significant difference lies in duplicating and calibrating the base model to each new individual. Additionally, the calibration computation is offloaded to the individuals. As seen in figure 3.2, calibrating the base model f_θ for participant P_n results in a personalized model g_θ^n.

This schema comes with advantages. Firstly, since each model is calibrated and not exposed to increasing amounts of data, these models are not prone to catastrophic forgetting. This phenomenon, where ML models forget previously known tasks, is known to be present unless CL methods are implemented. Secondly, the individualization of models respects the vision of personalized medicine. Thirdly, information about new participants is kept private by default. Information should be relayed back to the group that created the model. Individuals may benefit from the model without giving away potential critical details. Lastly, the computational burden would be significantly reduced for the creators of the base model. Since the calibration of g_θ would be computed on the user's side, the only centralized computation would be for training the base model f_θ.

Overall, this model distribution presents itself as the more realistic scenario in the face of model development in research. Integrating model distribution requires less work, allowing public access to the research model development workflow.

3.3.3. Calibration vs Transfer Learning

A note should be made about the similarities between calibration and transfer learning. In most cases, including the one presented in this thesis, transfer learning can be employed instead of calibration for additional benefits.

Transfer learning is an ML technique in which a model trained on one task is modified and used to perform a second task. The idea behind this approach is to use the knowledge gained from one task to improve the performance of a different task. Transfer learning can save time and computational resources when training a new model and enhance the performance of existing models. This technique is handy for tasks requiring large amounts of data, such as natural language processing or image recognition.

Calibration is, therefore, a basic technique of transfer learning. In contrast, transfer learning would allow for high performance on two tasks. The similarity distance would determine the calibration change to match the previous performance. Additionally, training solely on the last layers of transfer learning could easily be employed in the calibrations of figure 3.2 for datasets where the early representations need to be strongly conserved.

3.3.4. Calibration Frequency

As the article discusses, the calibration frequency must be examined. An individual's gait may change over time. If this change were too significant, the personalized model would need re-calibration. This idea applies to all bio-markers collected. Of the possible values, a spectrum is possible.

At one extreme, a single calibration would be necessary. Otherwise, one calibration would be sufficient for the participant's lifetime.

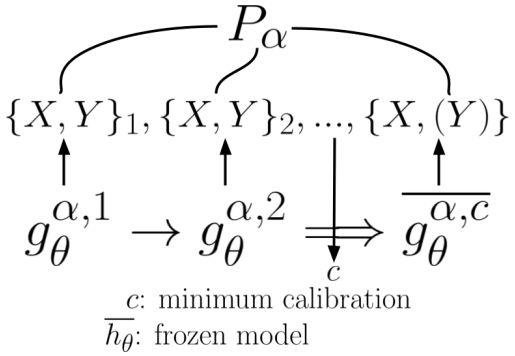

Figure 3.3 – Single calibration model life cycle. For the participant P_α, the duplicated base model g_θ^α is generated. Exposure to the first trial X,Y_1, results in partially calibrated $g_\theta^{\alpha,1}$. When c trials have been observed, the model $g_\theta^{\alpha,c}$ is adequately calibrated. The weights can be frozen to retain performance, yielding $\overline{g_\theta^{\alpha,c}}$. After which, labels of trials (i.e. Y) become optional.

As figure 3.3 shows, the duplicated model is calibrated for the specific participant. After the full calibration, the model may be frozen to preserve accuracy. Since it has been determined that a single calibration is required, all predictions after exposure to the minimum calibration trials will match previous performance. This situation allows for a twofold advantage. Since the minimum calibration has been determined, computing costs have been reduced. Additionally, since the model is frozen afterwards, compressing the model for quick prediction on edge computing becomes highly accessible.

At the other extreme, continuous calibration would be required. In this situation, continual learning (CL) should be employed. This allows access to all benefits available to the CL framework.

Since the CL framework explicitly deals with situations of continuous exposure to new data points, it is perfectly adapted to deal with this situation. Exposure to m new data points will result in the base model potentially updating m times. Figure 3.4 shows that the model is categorized as $\gamma_\theta^{\beta,m}$ after m trials.

Lastly, a balance may be struck. The timed calibration would allow a model's calibration to be valid for certain predictions or times. After which, calibration would be necessary to generate the calibrated model.

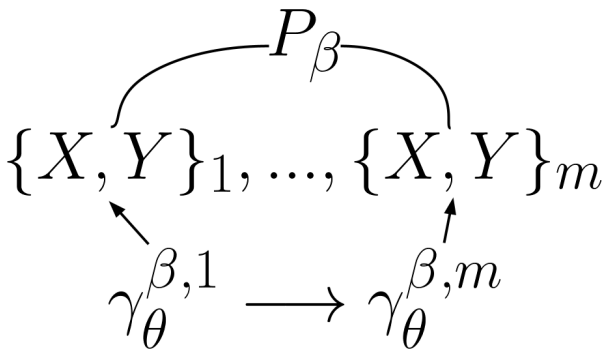

Figure 3.4 – Continuous calibration model life cycle. Given a continual learner model γ_θ, the model will continuously adapt to the participant P_β. After exposure to m trials, the model can be expressed as $\gamma_\theta^{\beta,m}$.

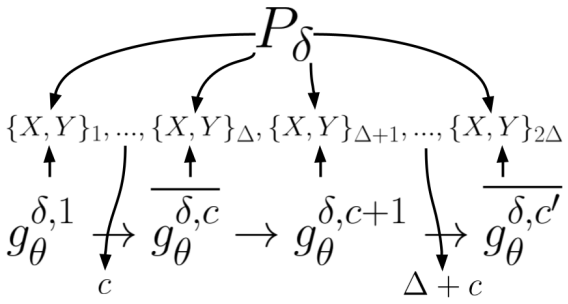

Figure 3.5 – Timed calibration model life cycle. Similar to figure 3.3, the model is calibrated and frozen after the necessary c trials, generating $\overline{g_\theta^{\delta,c}}$. With a threshold Δ specified, calibration will be triggered after Δ trials.

As shown in figure 3.5, a threshold variable Δ is specified to trigger a re-calibration. The previously frozen model $\overline{g_\theta^{\delta,c}}$ is unfrozen. This allows centering to the new reference data, yielding $g_\theta^{\delta,c+1}$ and eventually the re-calibrated model $\overline{g_\theta^{\delta,c'}}$. This cycle will be repeated ad infinitum. By adding a timed validity to calibration, we allow for more flexibility and support for more cases of biomarker data. Since people tend to change over time, multiple healthcare datasets are susceptible to benefit from this model control. Unfortunately, the re-calibration raises the computational cost. While this step might not be costly, the newer model must be pushed again to edge devices.

Chapter 4

Conclusion

The new recent phase of modern medicine, personalized medicine, is exciting. While many implementation and safety challenges are unsolved, their achievement would mark an important milestone for the medical care of patients. Currently, powerful mobile devices allow deep learning models to be accessible outside laboratory settings and in the real world. This situation aligns with personalized medicine, which aims for tailored treatment for its users. The evaluation of said models to participants must be verified to tailor models appropriately. When dealing with users' biomedical data, performance evaluation might be split into inter- vs intra- subject. The presented article investigates matching intra- level performance by initially inter-subject training. This is synonymous with matching random-wise split performance with initial training on a subject-wise split of training and testing datasets. These deep learning models are calibrated to achieve robust predictions on participants unknown during training. Our results of calibration dynamics demonstrate a minimum calibration set size which significantly benefits the performance of deep learning. Observing these dynamics will inform researchers about their datasets' and models' interactions. Developers will benefit from integrating model calibration into real-world applications.

Bibliography

Ahmed, Zeeshan et al. (2020). "Artificial intelligence with multi-functional machine learning platform development for better healthcare and precision medicine". In: *Database* 2020.

Bird, Jordan J et al. (2021). "Synthetic Biological Signals Machine-generated by GPT-2 improve the Classification of EEG and EMG through Data Augmentation". In: *IEEE Robotics and Automation Letters* 6.2, pp. 3498–3504.

Bol, Linda and Douglas J Hacker (2012). "Calibration research: Where do we go from here?" In: *Frontiers in psychology* 3, p. 229.

Campuzano, Susana et al. (2020). "Beyond sensitive and selective electrochemical biosensors: Towards continuous, real-time, antibiofouling and calibration-free devices". In: *Sensors* 20.12, p. 3376.

Cano, J. et al. (2022). "The Relevance of Calibration in Machine Learning-Based Hypertension Risk Assessment Combining Photoplethysmography and Electrocardiography". In: *Biosensors* 12. DOI: 10.3390/bios12050289.

Cao, L. (2022). "Beyond i.i.d.: Non-IID Thinking, Informatics, and Learning". In: *IEEE Intelligent Systems* 37. DOI: 10.1109/MIS.2022.3194618.

Caywood, Matthew S et al. (2017). "Gaussian process regression for predictive but interpretable machine learning models: an example of predicting mental workload across tasks". In: *Frontiers in human neuroscience* 10, p. 647.

Chadwick, Ruth and Alan O'Connor (2013). "Epigenetics and personalized medicine: prospects and ethical issues". In: *Personalized Medicine* 10.5, pp. 463–471.

Charoenphakdee, Nontawat, Jongyeong Lee, and Masashi Sugiyama (2021). "A Symmetric Loss Perspective of Reliable Machine Learning". In: *arXiv preprint arXiv:2101.01366*.

Ching, Travers et al. (2018). "Opportunities and obstacles for deep learning in biology and medicine". In: *Journal of The Royal Society Interface* 15.141, p. 20170387.

Chollet, Francois et al. (2015). *Keras*. URL: https://github.com/fchollet/keras.

Conti, Rena et al. (2010). "Personalized medicine and genomics: challenges and opportunities in assessing effectiveness, cost-effectiveness, and future research priorities". In: *Medical Decision Making* 30.3, pp. 328–340.

Davis, Sharon E et al. (2017). "Calibration drift in regression and machine learning models for acute kidney injury". In: *Journal of the American Medical Informatics Association* 24.6, pp. 1052–1061.

Dawkins, Christina, Thirukodikaval Nilakanta Srinivasan, and John Whalley (2001). "Calibration". In: *Handbook of econometrics*. Vol. 5. Elsevier, pp. 3653–3703.

Deo, R.C. (2015). "Machine Learning in Medicine". In: *Circulation* 115. DOI: 10.1161/circulationaha.115.001593.

Drake, Robert E, Delia Cimpean, and William C Torrey (2022). "Shared decision making in mental health: prospects for personalized medicine". In: *Dialogues in clinical neuroscience*.

Ginsburg, Geoffrey S and Nicole M Kuderer (2012). "Comparative effectiveness research, genomics-enabled personalized medicine, and rapid learning health care: a common bond". In: *Journal of clinical oncology* 30.34, p. 4233.

Grant, Barbara et al. (2022). "Why does the metabolic cost of walking increase on compliant substrates?" In: *Journal of the Royal Society Interface* 19.196, p. 20220483.

Hamburg, M. A. and F. S. Collins (2010). "The path to personalized medicine". In: *New England Journal of Medicine* 363.

Han, Chuanqi et al. (2020). "Calibration-free Blood Pressure Assessment Using An Integrated Deep Learning Method". In: *2020 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pp. 1001–1005. DOI: 10.1109/BIBM49941.2020.9313586.

Huot, Pierre-Luc et al. (2019). "A hybrid optimization approach for efficient calibration of computationally intensive hydrological models". In: *Hydrological Sciences Journal* 64.10, pp. 1204–1222.

Idili, Andrea et al. (2019). "Calibration-free measurement of phenylalanine levels in the blood using an electrochemical aptamer-based sensor suitable for point-of-care applications". In: *ACS sensors* 4.12, pp. 3227–3233.

Jakka, Sairamesh and Michael Rossbach (2013). "An economic perspective on personalized medicine". In: *The HUGO Journal* 7.1, pp. 1–6.

James, J. T. (2013). "A new, evidence-based estimate of patient harms associated with hospital care." In: *Journal of patient safety* 9.

Kachuee, Mohamad et al. (2015). "Cuff-less high-accuracy calibration-free blood pressure estimation using pulse transit time". In: *2015 IEEE international symposium on circuits and systems (ISCAS)*. IEEE, pp. 1006–1009.

Khazem, S. et al. (2021). "Minimizing Subject-dependent Calibration for BCI with Riemannian Transfer Learning". In: *International IEEE/EMBS Conference on Neural Engineering (NER)* 10. DOI: 10.1109/NER49283.2021.9441279.

Kirkpatrick, J. et al. (2017). "Overcoming catastrophic forgetting in neural networks". In: *PNAS* 114. DOI: 10.1073/pnas.1611835114.

Kwon, O-Yeon et al. (2019). "Subject-independent brain–computer interfaces based on deep convolutional neural networks". In: *IEEE transactions on neural networks and learning systems* 31.10, pp. 3839–3852.

Lehmler, S. et al. (2021). "Deep Transfer-Learning for patient specific model re-calibration: Application to sEMG-Classification". In: DOI: 10.48550/arXiv.2112.15019.

Little, M. et al. (2017). "Using and understanding cross-validation strategies. Perspectives on Saeb et al." In: *GigaScience* 6. DOI: 10.1093/gigascience/gix020.

Luo, Y. et al. (2020). "A database of human gait performance on irregular and uneven surfaces collected by wearable sensors". In: *Scientific Data* 7. URL: https://www.nature.com/articles/s41597-020-0563-y.

Luo, Yongle et al. (2021). "Calibration-free monocular vision-based robot manipulations with occlusion awareness". In: *IEEE Access* 9, pp. 85265–85276.

McGrath, D. et al. (2012). "Gyroscope-based assessment of temporal gait parameters during treadmill walking and running". In: *Sports Engineering* 15. URL: https://link.springer.com/article/10.1007/s12283-012-0093-8.

Meckley, Lisa M and Peter J Neumann (2010). "Personalized medicine: factors influencing reimbursement". In: *Health policy* 94.2, pp. 91–100.

Miotto, R. et al. (2018). "Deep learning for healthcare: review, opportunities and challenges". In: *Briefings in Bioinformatics* 19. DOI: 10.1093/bib/bbx044.

Mirshekari, Mostafa, Pei Zhang, and Hae Young Noh (2017). "Calibration-free footstep frequency estimation using structural vibration". In: *Dynamics of Civil Structures, Volume 2*. Springer, pp. 287–289.

Möller, H. J. (2022). "Effectiveness studies: advantages and disadvantages." In: *Dialogues in clinical neuroscience.*

Moy, Marilyn L et al. (2013). "Daily step count predicts acute exacerbations in a US cohort with COPD". In: *PLoS One* 8.4, e60400.

Obermeyer, Z. and E.J. Emanuel (2016). "Predicting the future–big data, machine learning, and clinical medicine". In: *PubMed* 375. DOI: 10.1056/NEJMp1606181.

Offit, K. (2011). "Personalized medicine: new genomics, old lessons". In: *Human genetics* 130.

Patel, K. et al. (2008). "Examining Difficulties Software Developers Encounter in the Adoption of Statistical Machine Learning". In: *AAAI*.

Pedregosa, F. and al. (2011). "Scikit-learn: Machine Learning in Python". In: *Journal of Machine Learning Research* 12. URL: https://www.jmlr.org/papers/volume12/pedregosa11a/pedregosa11a.pdf?ref=https://githubhelp.com.

Phillips, Kathryn A et al. (2014). "The economic value of personalized medicine tests: what we know and what we need to know". In: *Genetics in Medicine* 16.3, pp. 251–257.

Saeb, S. et al. (2017). "The need to approximate the use-case in clinical machine learning". In: *GigaScience* 6. URL: https://academic.oup.com/gigascience/article/6/5/gix019/3071704.

Shah, V. et al. (2022). "Generalizability of deep learning models for predicting outdoor irregular walking surfaces". In: *Journal of Biomechanics* 139. DOI: 10.1016/j.jbiomech.2022.111159.

Straus, S.E. and F.A. McAlister (2000). "Evidence-based medicine: a commentary on common criticisms." In: *Cmag* 163.

Sutherland, David H et al. (1980). "The development of mature gait." In: *Jbjs* 62.3, pp. 336–353.

Tobore, I. et al. (2019). "Deep Learning Intervention for Health Care Challenges: Some Biomedical Domain Considerations". In: *JMIR Mhealth Uhealth* 7. DOI: 10.2196/11966.

Tolson, Bryan A and Christine A Shoemaker (2007). "Dynamically dimensioned search algorithm for computationally efficient watershed model calibration". In: *Water Resources Research* 43.1.

Tomi-Tricot, Raphaël et al. (2019). "SmartPulse, a machine learning approach for calibration-free dynamic RF shimming: preliminary study in a clinical environment". In: *Magnetic Resonance in Medicine* 82.6, pp. 2016–2031.

Vidaurre, C. et al. (2011). "Machine-Learning-Based Coadaptive Calibration for Brain-Computer Interfaces". In: *Neural Computation* 23. DOI: 10.1162/NECO_a_00089.

Volm, Manfred and Thomas Efferth (2015). "Prediction of cancer drug resistance and implications for personalized medicine". In: *Frontiers in oncology* 5, p. 282.

Wang, S. et al. (2022). "Guest Editorial Emerging Challenges for Deep Learning". In: *IEEE Journal of Biomedical and Health Informatics* 26. DOI: 10.1109/JBHI.2022.3211369.

Watson, D.S. and al. (2019). "Clinical applications of machine learning algorithms: beyond the black box". In: *BMJ* 364. DOI: 10.1136/bmj.l886.

Weissler, E.H. et al. (2021). "The role of machine learning in clinical research: transforming the future of evidence generation." In: *Trials* 22. DOI: 10.1186/s13063-021-05489-x.

White, R. and S. Taylor (2002). "Nursing practice should be informed by the best available evidence, but should all first-level nurses be competent at research appraisal and utilization?" In: *Nurse Education Today* 22.

Yang, Jingmin et al. (2022). "Calibration-Free 3D Indoor Positioning Algorithms Based on DNN and DIFF". In: *Sensors* 22.15, p. 5891.

Zemouri, R., N. Zerhouni, and D. Racoceanu (2019). "Deep Learning in the Biomedical Applications: Recent and Future Status". In: *Appl. Sci.* 9. DOI: 10.3390/app9081526.

Zhang, Quanshi, Ying Nian Wu, and Song-Chun Zhu (2018). "Interpretable convolutional neural networks". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 8827–8836.