Journal of Biomechanics 165 (2024) 111998

Contents lists available at ScienceDirect

Journal of Biomechanics

journal homepage: www.elsevier.com/locate/jbiomech

Strategies to optimise machine learning classification performance when using biomechanical features

Bernard X.W. Liew [a,*], Florian Pfisterer [b,c], David Rügamer [b,c], Xiaojun Zhai [d]

[a] School of Sport, Rehabilitation and Exercise Sciences, University of Essex, Colchester, Essex, United Kingdom
[b] Department of Statistics, LMU Munich, Munich Germany
[c] Munich Center for Machine Learning, Munich, Germany
[d] School of Computer Science and Electrical Engineering, University of Essex, Colchester, Essex, United Kingdom

ARTICLE INFO

Keywords:
Machine learning
Deep learning
Gait
Biomechanics
Orthopedic
Musculoskeletal pain

ABSTRACT

Building prediction models using biomechanical features is challenging because such models may require large sample sizes. However, collecting biomechanical data on large sample sizes is logistically very challenging. This study aims to investigate if modern machine learning algorithms can help overcome the issue of limited sample sizes on developing prediction models. This was a secondary data analysis two biomechanical datasets – a walking dataset on 2295 participants, and a countermovement jump dataset on 31 participants. The input features were the three-dimensional ground reaction forces (GRFs) of the lower limbs. The outcome was the orthopaedic disease category (healthy, calcaneus, ankle, knee, hip) in the walking dataset, and healthy vs people with patellofemoral pain syndrome in the jump dataset. Different algorithms were compared: multinomial/LASSO regression, XGBoost, various deep learning time-series algorithms with augmented data, and with transfer learning. For the outcome of weighted multiclass area under the receiver operating curve (AUC) in the walking dataset, the three models with the best performance were InceptionTime with x12 augmented data (0.810), XGBoost (0.804), and multinomial logistic regression (0.800). For the jump dataset, the top three models with the highest AUC were the LASSO (1.00), InceptionTime with x8 augmentation (0.750), and transfer learning (0.653). Machine-learning based strategies for managing the challenging issue of limited sample size for biomechanical ML-based problems, could benefit the development of alternative prediction models in healthcare, especially when time-series data are involved.

1. Introduction

Gait impairments are common in many orthopedic (Biggs et al., 2022), musculoskeletal (Diamond et al., 2017), neurological (de Freitas Guardini et al., 2021), and cardiovascular disorders (Green et al., 2016). The quantification of gait impairments for use in predictive models can serve in facilitating clinical decision-making (Chia et al., 2020), and stratify patients to homogenous functional severity levels (Tsitlakidis et al., 2019) for allocation resourcing, and prognostication (Capin et al., 2017; de Freitas Guardini et al., 2021). Predictive models are typically required to understand the relationship between a set of risk/prognostic factors and clinically relevant outcomes (Shibuya et al., 2020). A challenge in the development of predictive models is the issue of sample size. For example, using 10, 20, and 50 events per predictor parameter rule (Cruz et al., 2020; Riley et al., 2019), for just 20 included predictors, the number of required participants will exceed some of the largest prospective clinical cohort studies to date (n = 2758 participants (Traeger et al., 2016)).

While new techniques are emerging quickly in machine learning (ML) and deep learning, many studies show that tree-based gradient boosting techniques such as XGBoost (Chen and Guestrin, 2016) still outperform most techniques, especially, when the sample size is small (Benkendorf and Hawkins, 2020). One reason for this is that (deep) neural networks (DNNs) require much more data than ML approaches (Al-Qerem et al., 2021), compared to gradient-boosting methods. If insufficient data are present, DNNs can easily lead to statistical overfitting (Marcus, 2018). For example, some of the largest biomechanics studies to date have recruited over 2000 participants (Horsak et al., 2020), which still pales in comparison to deep learning models trained on millions of samples of images (Simonyan and Zisserman, 2014).

* Corresponding author.
E-mail address: bl19622@essex.ac.uk (B.X.W. Liew).

https://doi.org/10.1016/j.jbiomech.2024.111998
Accepted 13 February 2024
Available online 15 February 2024

Two techniques that have been proven successful in improving the performance of ML in small data regimes are transfer learning (Pan and Yang, 2010) and data augmentation (Iwana and Uchida, 2021). Data augmentation artificially generates new data observations based on existing data. Data augmentation can help to improve ML performance by training the model on a more diverse data set that reduces overfitting (Moreno-Barea et al., 2020). This has proven to be particularly useful in computer vision (Shorten and Khoshgoftaar, 2019). Augmenting time-series data is more challenging (Iwana and Uchida, 2021), as a manipulation of a data instant is more likely to change the whole character of the time-series than an image. Transfer learning, on the other hand, uses "knowledge" from very large pre-trained models and combines it with new data for model fine-tuning (Liew et al., 2021). These pre-trained models can have >1 million observations and typically take the form of images and signals (Simonyan and Zisserman, 2015).

Whether deep learning can match the performance of state-of-the-art boosting approaches in situations of relatively small sample size are unknown. Currently, many studies only show that neural networks themselves can yield improved performance using transfer learning (Krizhevsky et al., 2012) or data augmentation (Iwana and Uchida, 2021), compared to a network developed from scratch. Few have directly compared transfer learning against augmentation (Al-Qerem et al., 2021; Zhong et al., 2021), and none in the area of biomechanics. The primary aim of this study is to provide a fair and realistic comparison of ML and deep learning, by including a statistical regression as a baseline and allowing for the same amount of data augmentation in the boosting approach as used for the DNNs. We hypothesised that both data augmentation and transfer learning would result in greater prediction performance than the state-of-art shallow ML algorithm on the original data.

2. Methods

This was a secondary analysis of two datasets – a publicly available walking dataset with a fairly large sample size of 2295 participants (Horsak et al., 2020); and a small dataset of 31 participants performing maximal countermovement jumps (CMJ) (Liew et al., 2020a). The details of the data collection and processing procedures will be briefly summarized here.

2.1. GaitRec dataset

2.1.1. Participants

Five groups of participants were recruited - healthy, and patients with hip, knee, ankle, and calcaneus orthopaedic disorders (Horsak et al., 2020).

2.1.2. Protocol

Ground reaction forces (GRFs) were recorded by having participants walk unassisted at a self-paced speed along a 10 m walkway across two in-ground force plates (2000 Hz, Kistler, Type 9281B12, Winterthur, CH). All participants had the option of either walking barefoot or in shoes. The GRF signals were filtered using a 2nd order low-pass Butterworth filter (20 Hz). Gait events of initial contact and toe-off were calculated using a 25 N threshold of the vertical GRF. The GRF signals were time normalised to 101 data points in the stance phase, and amplitude normalised to the body weight (N).

2.2. Patellofemoral pain syndrome (PFPS) dataset

2.2.1. Participants

Fourteen participants with PFPS and 17 health controls volunteered for the study (Liew et al., 2020a). Ethical clearance was obtained from the Ethics Committee of the University of Birmingham, United Kingdom (MCR041218-1).

2.2.2. Protocol

Participants performed maximal CMJ on two in-ground force plates (500 Hz, BTS P6000, BTS Bioengineering, Italy). The depth reached during the countermovement phase was self-determined and practiced by each participant. GRF data were low-pass filtered at 75 Hz (4th order, zero-lag, Butterworth), and time-normalised to 101 data points between the start of the eccentric phase and toe-off, and scaled to each participant's bodyweight (N). Only GRF variables from one side (right or left) were selected. For healthy controls and individuals with bilateral PFPS, GRF variables from the right side were selected. For individuals with unilateral PFPS, GRF variables from the side of pain were selected.

2.3. Software

All analyses were conducted on a Linux Server, Ubuntu 22.04, with 32 Cores (Intel(R) Xeon(R) CPU E5-2650 v2 @ 2.60 GHz) and 64 GB RAM. Experiments used the R software (version 4.2.0) and Python (version 3.6.9), with associated codes and results found online (https://github.com/davidruegamer/TransferLearning_MTSC). The following packages were used: *mlr3* for ML and model tuning (Lang et al., 2019), XGBoost for gradient boosting (Chen and Guestrin, 2016), and *mlr3keras* for deep learning in the mlr3 (https://github.com/mlr-org/mlr3keras) framework, *reticulate* which provides an R interface to Python [25], *tensorflow* [28] for training deep neural networks (DNNs), and the *glmnet* for multinomial logistic regression (Simon et al., 2011). To incorporate augmentation into the models, we extended mlr3keras and XGBoost by additional hyperparameters, allowing the use of different data augmentation strategies (see the following Data augmentation section) and their tuning.

2.4. Machine learning algorithms

For the GaitRec dataset, the outcome was categorical with five levels (healthy, calcaneus, ankle, knee, and hip), as well as six GRF (three-axis bilaterally) time-series variables as predictors. For the PFPS dataset, the outcome was binary (healthy vs PFPS), whilst there were three time-series variables as predictors. For both datasets, the predictors were scaled to a mean of zero and a standard deviation of 1 as a pre-processing step.

2.4.1. Data augmentation

Since many different data augmentation techniques for time-series exist (Iwana and Uchida, 2021) and it is not a priori clear, which techniques are optimal to improve classification performance, we here focus on a predefined set of data augmentation strategies (Table 1), namely (1) jittering, (2) magnitude warping, (3) random guided warping, (4) spawner, and (5) window slicing. A description of these strategies can be found in the Supplementary Material. Herein, we used all five methods to enhance the original sample size for model training. To not over-or underfit due to the limited number of observations, four different amounts of augmentation (none; 2, 4, 8, and 12 times the original data size) were used with the predefined augmentation strategies as defined previously.

Table 1
Values used for various data augmentation strategies.

Strategy	Argument	Values
Jittering	Sigma	0.03
Magnitude warping	Sigma	0.2
	Knots	4
Random guided warping	Slope constraint	Symmetric
	Use window	True
	DTW type	Normal
Spawner	Sigma	0.05
Window slicing	Reduce ratio	0.9

2.4.2. Baseline – multinomial/LASSO logistic regression w/o data augmentation

For the GaitRec dataset, as a baseline algorithm, we used a multinomial logistic regression. Every time point of every time-series is considered as a predictor – resulting in 606 predictors (6 GRF variables each with 101 data points). For the PFPS dataset, a traditional logistic regression cannot be used as the number of predictors (3 GRF variables, each 101 data points) exceeds the sample size. Hence, a logistic regression with the the least absolute shrinkage and selection operator (LASSO) penalty was used. The optimal amount of shrinkage was determined by a nested inner 3-fold cross-validation.

2.4.3. XGBoost

We chose the XGBoost as reflective of a state-of-the-art ML algorithm (Chen and Guestrin, 2016). XGBoost has been shown to perform similarly well to a deep neural network for time-series prediction of biomechanical signals (Wang et al., 2020), as well as general time-series classification tasks (Pfisterer et al., 2019). We tuned XGBoost using Hyperband (Li et al., 2017b) with the number of boosting iterations as a budget parameter. Hyperband is a technique from automated ML that automatically tries to find the best configuration for every tuning parameter of the model (Li et al., 2017a). We used Hyperband to automatically tune the algorithm including the amount of augmentation based on a nested inner 3-fold cross-validation (Wainer and Cawley, 2021). Details of the tuned hyperparameters used can be found in the Supplementary Material.

2.4.4. Deep learning approaches

Various time-series classification (TSC) neural network architectures exist (e.g. (Goschenhofer et al., 2021)). One well-known architecture is InceptionTime (Ismail Fawaz et al., 2020). Next to InceptionTime, we also investigate the performance of fully convolutional neural networks (FCNs) (Wang et al., 2017). FCNs are composed of three convolutional blocks that use a convolution operation. Both networks yield state-of-the-art TSC performances (Ismail Fawaz et al., 2019), hence these were selected for this study. As automatic tuning of the two DNNs is computationally expensive, we train the two architectures with a predefined set of default hyperparameters (see Supplementary Material).

2.4.5. Transfer learning (TL)

We here investigated two approaches. The first approach is based on the findings from prior studies (Johnson et al., 2019; Liew et al., 2021), concatenating the time-series and considering the input as an image. This allows the use of large pre-trained convolutional neural networks (here the VGG16 (Simonyan and Zisserman, 2014)) trained on a large corpus of images (here ImageNet). Even though a previous study reported that a custom deep neural network performed better than TL of pre-trained image models (Liew et al., 2021), the previous study focused on predicting a time-series outcome, whilst the pre-trained image model was trained on a simpler classification problem (Liew et al., 2021). Given that this study focused on classification problems, we decided to retain this method as a basis for comparison.

The second TL approach makes use of the results provided (Ismail Fawaz et al., 2018), and then uses transfer-learning on a whole collection of pre-trained models, each trained on a different time-series data set. We performed TL on pre-trained time-series models as these problems were all dealing with a classification problem (Ismail Fawaz et al., 2018), similar to this study. Given that many pre-trained time-series models were available, we took the average performance metrics across all pre-trained models. These datasets may not be related to the actual task of GaitRec/PFPS classification, but the trained models can potentially provide good and general feature extractors from time-series data.

All pre-trained (image and time-series) networks are adapted by changing the input, last hidden, and output layer, and then fine-tuned on the actual GaitRec/PFPS data set. The default hyperparameters used can be found in the Supplementary Material.

2.5. Predictive accuracy and model evaluation

The performance of all methods was evaluated using 10-fold cross-validation of the full data set. The split was performed at the subject-level, meaning that every participant is exclusively in the training, validation, or test set. For each fold, the early stopping of neural network-based approaches was done using a validation data set consisting of 20 % of the current fold's training data. The primary measure of model performance was the log loss of the test set. The log-loss measures how well the predicted distribution of the classes fits the true distribution of the classes.

2.5.1. GaitRec dataset

As secondary measures, we report the calculated multi-class Brier score, the weighted multi-class area under the receiver operating curve (AUC), and the balanced accuracy. The AUC measures as well as the Brier score work on class probabilities. The Brier score measures the quadratic difference of predicted probability for one class and is an indicator of whether this class was observed, and then averaged over all classes and observations. The ideal value is thus 0, while a perfectly confident but wrong prediction for every class and sample yields the value 2. The multi-class AUC averages over all individual AUC values when comparing only one of the five classes against another one, and ranges from 0 to 1, with a value of 1 being when the model can perfectly distinguish between all classes. Its weighted version accounts for different class frequencies as not all classes have been observed equally often. The accuracy reflects the ratio between the number of correct predictions made by the model to the total number of predictions made – this ranges from 0 (no correct prediction) to 1 (perfect prediction). The balanced accuracy, in turn, accounts for the different class frequencies and weights the individual class accuracies according to their observed frequency in the actual data.

2.5.2. PFPS dataset

Here, we also reported similar performance metrics for binary classification tasks, including the Brier score, AUC, and accuracy, with identical interpretation of the values as described in the multiclass classification case.

3. Results

3.1. GaitRec dataset

In general, a large correlation was observed between the metrics that measure the goodness of predicted probabilities, such as between the log loss and multiclass brier score (with a correlation 0.84); as well as

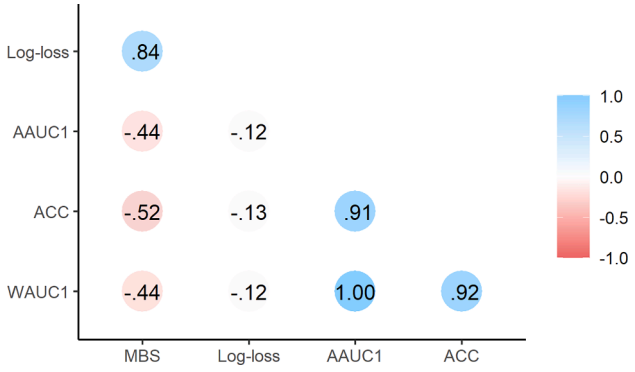

Fig. 1. Correlation matrix of the results of different machine learning strategies across different performance metric outcomes for the GaitRec dataset. **Abbreviation:** multi-class Brier score (MBS), the weighted multi-class area under the receiver operating curve (WAUC1), the averaged multi-class area under the receiver operating curve (AAUC1), and the balanced accuracy (ACC).

between the metrics that measure the goodness of class ranking (i.e., all accuracy and AUC metrics; correlation between 0.91 and 1.00) (Fig. 1). However, a negative correlation was observed between the probability assessment measures and the class-based assessment measures (between −0.52 and −0.12) (Fig. 1).

For the primary outcome of log-loss, the top three performing models were XGBoost (1.172), transfer learning image-based (TL-image) (1.286), and transfer learning time-series (TL-time) without augmentation (1.366) (Fig. 2). For the Brier score, the three models with the best performance were XGBoost (0.625), TL-image (0.681), and with logistic regression (0.684) (Fig. 2). For the weighted multiclass AUC, the three models with the best performance were InceptionTime with x12 augmented data (0.810), XGBoost (0.804), and multinomial logistic regression (0.800). For balanced accuracy, the three models with the best performance were XGBoost (0.520), InceptionTime with x12 augmented data (0.512), and InceptionTime with x8 augmented data (0.509) (Fig. 2).

3.2. PFPS dataset

For the primary outcome of log-loss, the top three performing models were the LASSO (0.544), and FCN models without augmentation (0.729), and x2 augmentation (0.732) (Fig. 3). For the Brier score, the LASSO (0.180), and FCN models without augmentation (0.268), and x2 augmentation (0.269) were the top three performing models (Fig. 3). For accuracy, the top three performing models were the LASSO, FCN model with x12 augmentation, and InceptionTime, all with a value of 0.667 (Fig. 3). For the AUC, the top three models were the LASSO (1.00), InceptionTime with x8 augmentation (0.750), and TL-time (0.653) (Fig. 3). These results, have a large variability due to the small size of the data set and can only give a rough indication.

4. Discussion

The GaitRec study has demonstrated that clinical biomechanics data can be collected at scale in the clinical environment. The emergence of smart technologies means that large-scale clinical biomechanics data collection will soon become the norm and that ML may be increasingly relied upon to drive healthcare applications. Several important findings emerged from the present study. First, XGBoost had superior prediction performance in a larger dataset but not in a smaller dataset. Second, the classification performance, as defined by class-based metrics like the AUC, of many algorithms, such as XGBoost, FCN, InceptionTime declined markedly when applied to a much smaller dataset. Third, DNNs can be effectively trained using default hyperparameters, on a moderately large biomechanics time-series data set with a performance matching the performance of XGBoost. Fourth, the study shows that augmentation of biomechanical time-series data works in practice, albeit on larger datasets, and can notably boost the classification performance of ML prediction models. Lastly, even though the present study used time-series predictors, TL-image still provides a better architecture compared to TL-series.

A previous ML study using one-dimensional convolutional DNN on the GaitRec data reported a log-loss of 0.25 and an accuracy of 0.92 (Pandey et al., 2022). The much better prediction performance in the previous study was likely because they binarised the outcome (healthy vs disorder) (Pandey et al., 2022), rather than dealing with a more challenging prediction problem of classification of five classes. In small binary classification problems in biomechanics, previous studies have reported AUC values of 80 % (n = 47) (Liew et al., 2020b), 93 % (n = 31)

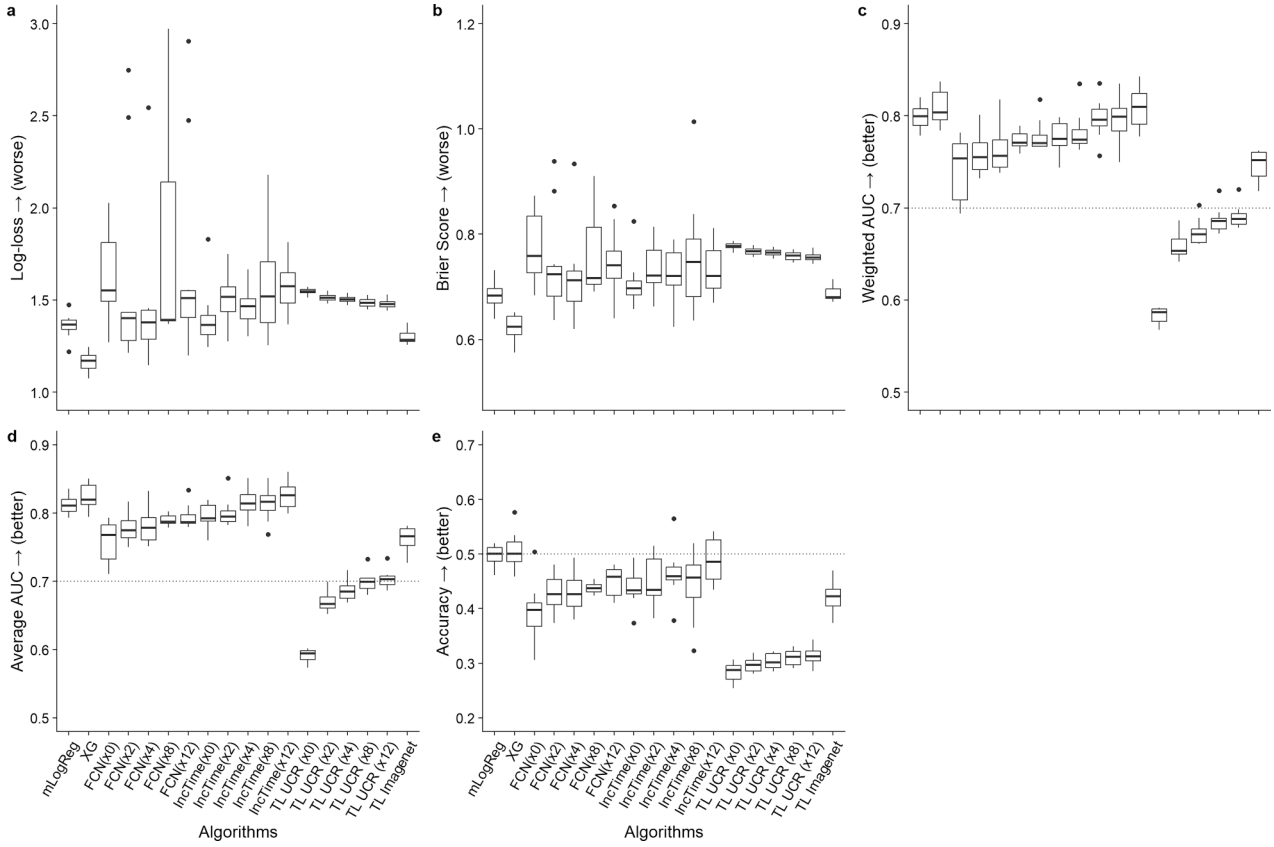

Fig. 2. Predictive performance of different statistical and machine learning algorithms. (a) Log-loss values, (b) the multi-class Brier Score, (c) the weighted multi-class area under the receiver operating curve, (d) the averaged multi-class area under the receiver operating curve, and (e) the balanced accuracy. A**bbreviation:** mLogReg – multinomial logistic regression; XG – extreme gradient boosting; FCN – fully connected network; IncTime – InceptionTime architecture; TL – transfer learning; UCR – University of California, Riverside datasets.

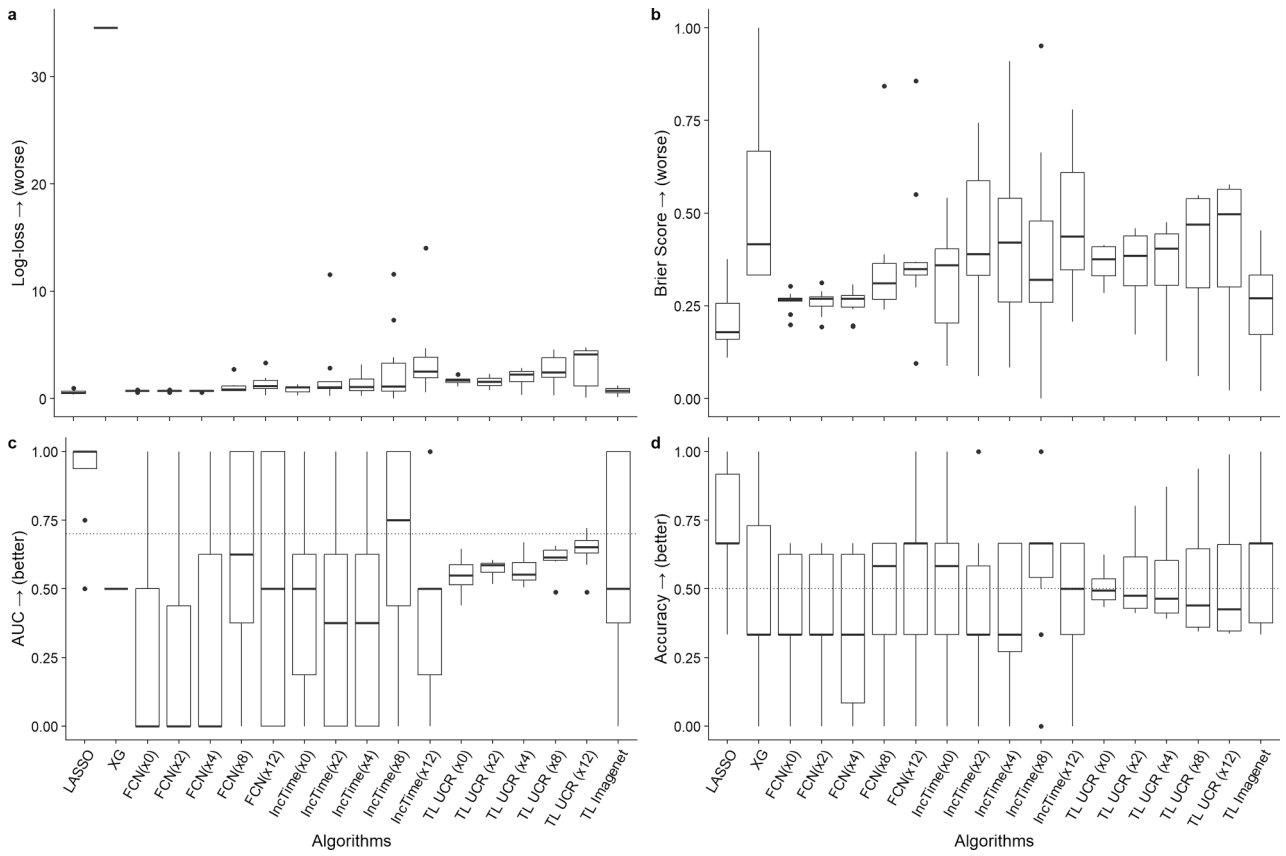

Fig. 3. Predictive performance of different statistical and machine learning algorithms. (a) Log-loss values, (b) the Brier Score, (c) the area under the receiver operating curve, and (d) accuracy. A**bbreviation:** LASSO – least absolute shrinkage and selection operator; XG – extreme gradient boosting; FCN – fully connected network; IncTime – InceptionTime architecture; TL – transfer learning; UCR – University of California, Riverside datasets.

(Liew et al., 2020a), and 90 % (n = 49) (Liew et al., 2019). This was comparable to the LASSO model in the present study which reported a perfect AUC.

The negative correlation of probability (e.g. Brier score) - and class-based metrics (e.g. AUC) in our study may not be surprising, since only classifiers that are naturally calibrated (such as the multinomial logistic regression) will provide good scores in terms of both measures (Van Calster et al., 2019). Regardless of the size of the dataset investigated presently, TL-image produced models with probabilities that were calibrated better than logistic regression, whereas all TL-time in general performed worse both in ranking and in calibration. Augmentation appears to work better than TL-time and TL-image both in terms of calibration and ranking in a larger dataset. This contrasted with other studies that reported that TL was better than augmentation, albeit in different scientific domains (Al-Qerem et al., 2021; Zhong et al., 2021).

Interestingly, TL from pre-trained imaged models was superior to TL from pre-trained time-series models, even though this study used time-series predictors. Speculatively, the superior performance of TL-image compared to TL-time could be because pre-trained image models have been trained on classification tasks with many more different classes than pretrained time-series models. The inferior performance of TL-image relative to augmentation in the present study could be the present study focused on a classification task with few restricted classes, and that the former is more suited either for a classification task with many classes or a regression task.

TL overall appears to provide more precise predictions across the testing folds, compared to all other methods in a larger dataset. TL may provide some form of inductive bias by already having converged to a solution for a larger, pre-trained, block of the network (Xuhong et al., 2018). This explains why TL requires fewer parameters to be trained compared to training an entire DNN from scratch. The inductive bias could guide the model towards a specific solution, which, however, can also be a local optimum, potentially explaining the inferior performance of TL approaches in some cases. The augmentation models in contrast are all trained from scratch and may not have yet converged to one solution (a DNN has many different (local) optima). This lack of convergence to a single solution may be worsened by a small dataset which explains the low precision of predictions across folds in the PFPS dataset.

While the performance of DNNs is still inferior to the tuned and augmented XGBoost model, the study shows that DNNs can be effectively trained on moderately large time-series data sets with a performance matching a tuned XGBoost model in terms of class rankings only. The study further shows that data augmentation of biomechanical gait data works in practice and can notably boost the performance of models in larger data regimes. We note that XGBoost with augmentation and automatic tuning via Hyperband took several days of training. If computational time and resources are limited, DNNs with default hyperparameters could be used in larger data regimes, whereas if time and resource are unlimited, the best option is a tuned gradient boosting model.

This study is not without limitations. The cross-sectional nature of the data precludes the ability to extrapolate our findings to longitudinal prediction models. Second, as neither the type of augmentation nor the architecture of the DNNs was tuned, our results further suggest that DNNs with augmentation can potentially outperform ML methods on disease classification using time-series predictors. Third, other methods of data augmentation exist which were not done in the present study. For example, prior studies have used more complex deep learning methods like Generative Adversarial Networks (Bicer et al., 2022), and also using physics-based musculoskeletal simulations (Dorschky et al., 2020).

5. Conclusions

The GaitRec study has already demonstrated the feasibility of collecting both biomechanics and clinical outcomes data at scale. With the emergence of technologies like wearable sensors and markerless motion capture, large-scale biomechanics data collection would soon become the norm, rather than the exception, in clinical environments. This study has demonstrated the feasibility of two strategies that could benefit ML prediction performance when using biomechanical features. TL using pre-trained image models appears to perform well in large biomechanics data regimes, like the GaitRec dataset. Data augmentation does not perform well in very small data regimes. Our approaches could benefit the development of alternative prediction models in healthcare, especially when non-conventional data types are incorporated, such as time-series and spatial data.

CRediT authorship contribution statement

Bernard X.W. Liew: . **Florian Pfisterer:** Visualization, Validation, Software, Methodology, Investigation, Formal analysis. **David Rügamer:** Writing – review & editing, Writing – original draft, Visualization, Validation, Software, Formal analysis, Conceptualization. **Xiaojun Zhai:** Writing – review & editing, Writing – original draft, Validation, Supervision, Methodology, Formal analysis.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Appendix A. Supplementary material

Supplementary data to this article can be found online at https://doi.org/10.1016/j.jbiomech.2024.111998.

References

Al-Qerem, A., Salem, A.A., Jebreen, I., Nabot, A., Samhan, A., 2021. Comparison between transfer learning and data augmentation on medical images classification. In: Proceedings of the 2021 22nd International Arab Conference on Information Technology (ACIT).

Benkendorf, D.J., Hawkins, C.P., 2020. Effects of sample size and network depth on a deep learning approach to species distribution modeling. Eco. Inform. 60, 101137.

Bicer, M., Phillips, A.T.M., Melis, A., McGregor, A.H., Modenese, L., 2022. Generative deep learning applied to biomechanics: a new augmentation technique for motion capture datasets. J. Biomech. 144, 111301.

Biggs, P., Holsgaard-Larsen, A., Holt, C.A., Naili, J.E., 2022. Gait function improvements, using Cardiff Classifier, are related to patient-reported function and pain following hip arthroplasty. J. Orthop. Res. 40, 1182–1193.

Capin, J.J., Khandha, A., Zarzycki, R., Manal, K., Buchanan, T.S., Snyder-Mackler, L., 2017. Gait mechanics and second ACL rupture: Implications for delaying return-to-sport. J. Orthop. Res. 35, 1894–1901.

Chen, T., Guestrin, C., 2016. XGBoost: A Scalable Tree Boosting System, Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Association for Computing Machinery, San Francisco, California, USA, pp. 785–94.

Chia, K., Fischer, I., Thomason, P., Graham, H.K., Sangeux, M., 2020. A decision support system to facilitate identification of musculoskeletal impairments and propose recommendations using gait analysis in children with cerebral palsy. Front Bioeng. Biotechnol. 8.

Cruz, E.B., Canhão, H., Fernandes, R., Caeiro, C., Branco, J.C., Rodrigues, A.M., Nunes, C., 2020. Prognostic indicators for poor outcomes in low back pain patients consulted in primary care. PLoS One 15, e0229265.

de Freitas Guardini, K.M., Kawamura, C.M., Lopes, J.A.F., Fujino, M.H., Blumetti, F.C., de Morais Filho, M.C., 2021. Factors related to better outcomes after single-event multilevel surgery (SEMLS) in patients with cerebral palsy. Gait Post. 86, 260–265.

Diamond, L.E., Van den Hoorn, W., Bennell, K.L., Wrigley, T.V., Hinman, R.S., O'Donnell, J., Hodges, P.W., 2017. Coordination of deep hip muscle activity is altered in symptomatic femoroacetabular impingement. J. Orthop. Res. 35, 1494–1504.

Dorschky, E., Nitschke, M., Martindale, C.F., van den Bogert, A.J., Koelewijn, A.D., Eskofier, B.M., 2020. CNN-based estimation of sagittal plane walking and running biomechanics from measured and simulated inertial sensor data. Front. Bioeng. Biotechnol. 8, 604.

Goschenhofer, J., Hvingelby, R., Ruegamer, D., Thomas, J., Wagner, M., Bischl, B., 2021. Deep Semi-supervised Learning for Time Series Classification. In Proceedings of the 2021 20th IEEE International Conference on Machine Learning and Applications (ICMLA).

Green, D.J., Panizzolo, F.A., Lloyd, D.G., Rubenson, J., Maiorana, A.J., 2016. Soleus muscle as a surrogate for health status in human heart failure. Exerc. Sport. Sci. Rev. 44, 45–50.

Horsak, B., Slijepcevic, D., Raberger, A.-M., Schwab, C., Worisch, M., Zeppelzauer, M., 2020. GaitRec, a large-scale ground reaction force dataset of healthy and impaired gait. Sci. Data 7, 143.

Ismail Fawaz, H., Forestier, G., Weber, J., Idoumghar, L., Muller, P.-A., 2018. Transfer learning for time series classification. arXiv arXiv:1811.01533.

Ismail Fawaz, H., Forestier, G., Weber, J., Idoumghar, L., Muller, P.-A., 2019. Deep learning for time series classification: a review. Data Min. Knowl. Discov. 33, 917–963.

Ismail Fawaz, H., Lucas, B., Forestier, G., Pelletier, C., Schmidt, D., Weber, J., Petitjean, F., 2020. InceptionTime: finding AlexNet for time series classification. Data Min. Knowl. Discov. 34, 1936–1962.

Iwana, B.K., Uchida, S., 2021. An empirical survey of data augmentation for time series classification with neural networks. PLoS One 16, e0254841.

Johnson, W.R., Alderson, J., Lloyd, D., Mian, A., 2019. Predicting athlete ground reaction forces and moments from spatio-temporal driven CNN models. IEEE Trans. Biomed. Eng. 66, 689–694.

Krizhevsky, A., Sutskever, I., Hinton, G.E., 2012. Imagenet classification with deep convolutional neural networks. Adv. Neural Inf. Process. Syst. 25, 1097–1105.

Lang, M., Binder, M., Richter, J., Schratz, P., Pfisterer, F., Coors, S., Bischl, B., 2019. mlr3: a modern object-oriented machine learning framework in R. J. Open Source Software 4, 1903.

Li, L., Jamieson, K., DeSalvo, G., Rostamizadeh, A., Talwalkar, A., 2017. Hyperband: a novel bandit-based approach to hyperparameter optimization. J. Mach. Learn. Res. 18, 6765–6816.

Liew, B., Rugamer, D., De Nunzio, A., Falla, D., 2019. Interpretable machine learning models for classifying low back pain status using functional physiological variables, 2 ed, Mendeley Data. doi: 10.17632/stbx779nt6.2.

Liew, B.X.W., Rugamer, D., Abichandani, D., De Nunzio, A.M., 2020a. Classifying individuals with and without patellofemoral pain syndrome using ground force profiles – development of a method using functional data boosting. Gait Post. 80, 90–95.

Liew, B.X.W., Rugamer, D., Stocker, A., De Nunzio, A.M., 2020b. Classifying neck pain status using scalar and functional biomechanical variables - development of a method using functional data boosting. Gait Post. 76, 146–150.

Liew, B.X.W., Rügamer, D., Zhai, X., Wang, Y., Morris, S., Netto, K., 2021. Comparing shallow, deep, and transfer learning in predicting joint moments in running. J. Biomech. 129, 110820.

Marcus, G.F., 2018. Deep Learning: A Critical Appraisal. ArXiv abs/1801.00631.

Moreno-Barea, F.J., Jerez, J.M., Franco, L., 2020. Improving classification accuracy using data augmentation on small data sets. Exp. Syst. Appl. 161, 113696.

Pan, S.J., Yang, Q., 2010. A survey on transfer learning. IEEE Trans. Knowl. Data Eng. 22, 1345–1359.

Pandey, C., Roy, D.S., Poonia, R.C., Altameem, A., Nayak, S.R., Verma, A., Saudagar, A.K. J., 2022. GaitRec-net: A deep neural network for gait disorder detection using ground reaction force. PPAR Res. 2022, 9355015.

Pfisterer, F., Beggel, L., Sun, X., Scheipl, F., Bischl, B., 2019. Benchmarking time series classification–functional data vs machine learning approaches. arXiv preprint arXiv: 1911.07511.

Riley, R.D., Snell, K.I.E., Ensor, J., Burke, D.L., Harrell Jr, F.E., Moons, K.G.M., Collins, G. S., 2019. Minimum sample size for developing a multivariable prediction model: PART II - Binary and time-to-event outcomes. Stat. Med. 38, 1276–1296.

Shibuya, M., Nanri, Y., Kamiya, K., Fukushima, K., Uchiyama, K., Takahira, N., Matsunaga, A., 2020. The maximal gait speed is a simple and useful prognostic indicator for functional recovery after total hip arthroplasty. BMC Musculoskel. Disord. 21, 84.

Shorten, C., Khoshgoftaar, T.M., 2019. A survey on image data augmentation for deep learning. J. Big Data 6, 60.

Simon, N., Friedman, J., Hastie, T., Tibshirani, R., 2011. Regularization paths for Cox's proportional hazards model via coordinate descent. J. Stat. Softw. 39, 1–13.

Simonyan, K., Zisserman, A., 2014. Very Deep Convolutional Networks for Large-Scale Image Recogn. arXiv 1409, 1556.

Simonyan, K., Zisserman, A., 2015. Very Deep Convolutional Networks for Large-Scale Image Recognition.

Traeger, A.C., Henschke, N., Hübscher, M., Williams, C.M., Kamper, S.J., Maher, C.G., McAuley, J.H., 2016. Estimating the risk of chronic pain: development and validation of a prognostic model (PICKUP) for patients with acute low back pain. PLoS Med. 13, e1002019.

Tsitlakidis, S., Horsch, A., Schaefer, F., Westhauser, F., Goetze, M., Hagmann, S., Klotz, M.C.M., 2019. Gait classification in unilateral cerebral palsy. J. Clin. Med. 8.

Van Calster, B., McLernon, D.J., van Smeden, M., Wynants, L., Steyerberg, E.W., Bossuyt, P., . . . prediction models' of the, S.i., 2019. Calibration: the Achilles heel of predictive analytics. BMC Med 17, 230.

Wainer, J., Cawley, G., 2021. Nested cross-validation when selecting classifiers is overzealous for most practical applications. Exp. Syst. Appl. 182, 115222.

Wang, C., Chan, P.P.K., Lam, B.M.F., Wang, S., Zhang, J.H., Chan, Z.Y.S., Cheung, R.T.H., 2020. Real-time estimation of knee adduction moment for gait retraining in patients

with knee osteoarthritis. IEEE Trans. Neural Syst. Rehabil. Eng.: A Publ. IEEE Eng. Med. Biol. Soc. 28, 888–894.
Wang, Z., Yan, W., Oates, T., 2017. Time series classification from scratch with deep neural networks: a strong baseline. International Joint Conference on Neural Networks (IJCNN) 1578–1585.
Xuhong, L., Grandvalet, Y., Davoine, F., 2018. Explicit inductive bias for transfer learning with convolutional networks. In: Proceedings of the International Conference on Machine Learning.
Zhong, S., Hu, J., Yu, X., Zhang, H., 2021. Molecular image-convolutional neural network (CNN) assisted QSAR models for predicting contaminant reactivity toward OH radicals: transfer learning, data augmentation and model interpretation. Chem. Eng. J. 408, 127998.