Title

Comparison of evaluation metrics of deep learning for imbalanced imaging data in osteoarthritis studies.

Journal

Osteoarthritis and Cartilage, 31(9)

Authors

Liu, Shen
Roemer, Frank
Ge, Yong
et al.

Publication Date

2023-09-01

DOI

10.1016/j.joca.2023.05.006

Peer reviewed

HHS Public Access

Author manuscript

Osteoarthritis Cartilage. Author manuscript; available in PMC 2024 September 01.

Published in final edited form as:

Osteoarthritis Cartilage. 2023 September ; 31(9): 1242–1248. doi:10.1016/j.joca.2023.05.006.

Comparison of Evaluation Metrics of Deep Learning for Imbalanced Imaging Data in Osteoarthritis Studies

Shen Liu, MS,

Frank Roemer, MD,

Yong Ge, PhD,

Edward J. Bedrick, PhD,

Zong-Ming Li, PhD,

Ali Guermazi, MD, PhD,

Leena Sharma, MD,

Charles Eaton, MD,

Marc C. Hochberg, MD,

David J. Hunter, PhD,

Michael C. Nevitt, PhD,

Wolfgang Wirth, PhD,

C. Kent Kwoh, MD*,

Xiaoxiao Sun, PhD*

Department of Epidemiology and Biostatistics, University of Arizona, 1295 N. Martin Ave., Tucson, AZ 85724, USA (S. L., E. J. B., X. S.); Department of Radiology, University of Erlangen

Address correspondent to C. K. K. (CKwoh@arthritis.arizona.edu), X. S. (xiaosun@arizona.edu).

*C. K. K. and X. S. are co-senior authors

5.Author Contributions

Guarantors of the integrity of the entire study, **S. L., F. R.**; study concepts/study design or data acquisition or data analysis/ interpretation, all authors; manuscript drafting or manuscript revision for important intellectual content, all authors; approval of the final version of the submitted manuscript, all authors; agrees to ensure any questions related to the work are appropriately resolved, all authors; literature research, ; clinical studies, ; experimental studies, ; statistical analysis, **S. L., X. S.**; and manuscript editing, all authors.

7.Disclosures of Conflicts of Interest

S. L. No relevant relationships. **F. R.** Shareholders of BICL, LLC. Consultant to Calibr-California Institute of Biomedical Research and Grunenthal. **Y. G.** No relevant relationships. **E. J. B.** No relevant relationships. **Z.-M. L.** No relevant relationships. **A. G.** Consultant to Pfizer, Novartis, Regeneron, TissueGene, Merck Serono, and AstraZeneca. Shareholders of BICL, LLC. **L. S.** No relevant relationships**. C. E.** No relevant relationships. **M.C.H.** Perform consulting activities, including attendance at virtual Advisory Board meetings, for the following entities: Acadia Pharmaceuticals, Bioclinica, Biosplice Therapeutics, BriOri Biotech, Centrexion Therapeutics, Eli Lilly, Dompe, Flexion Therapeutics Inc., Genascience, Gilead, GlaxoSmithKline, Kolon TissueGene, Novartis Pharma AG, Pfizer Inc., Regenosine, SFJ Pharmaceuticals Inc., Theralogix LLC, TrialSpark, WCG Analgesics Solutions, and Xalud Therapeutics. Member of Data Safety Monitoring Committees for clinical trials coordinated by ACI Clinical, Covance Inc., Cytel, ICON plc, IQVIA, and MiMedRx. Receive royalties from Elsevier (Editor, Rheumatology 7e and Editor-in-Chief, Seminars in Arthritis and Rheumatism) and Wolters Kluwer (UpToDate™). Stock options in BriOri Biotech, Regenosine, and Theralogix LLC. **D. J. H.** Provides consulting advice on scientific advisory boards for Pfizer, Lilly, TLCBio, Novartis, Tissuegene, and Biobone. **M.C.N.** No relevant relationships. **W. W.** Employee and shareholder of Chondrometrics GmbH. **C. K. K.** Consultant to Regeneron, LG Chem, Novartis, Xalud Therapeutics, and Express Scripts. He is the principal investigator for pharma-sponsored clinical trials to Abbvie, Cumberland, and GSK and DSMB to Kolon TissueGene and Avalor Therapeutics. **X. S.** No relevant relationships.

– Nuremberg, Erlangen, Germany (F. R.); Department of Management Information Systems, University of Arizona, AZ, USA (Y. G.); Department of Radiology, Boston University School of Medicine, MA, USA (F. R., A. G.); Department of Epidemiology and Biostatistics, University of California San Francisco, CA, USA (M. C. N.); Kent Memorial Hospital, and Department of Family Medicine, Warren Alpert Medical School, and Department of Epidemiology, School of Public Health, Brown University, RI, USA (C. E.); School of Medicine, University of Maryland, MD, and Medical Care Clinical Center, VA Maryland Health Care System (M. C. H.); Feinberh School of Medicine, Northwestern University, IL, USA (L. S.); Sydney Musculoskeletal Health, Kolling Institute, Faculty of Medicine and Health, The University of Sydney, Sydney, 2065 NSW, Australia, and Rheumatology Department, Royal North Shore Hospital, St Leonards, NSW 2065 Australia. (D. J. H.); Department of Imaging & Functional Musculoskeletal Research, Institute of Anatomy & Cell Biology, Paracelsus Medical University Salzburg & Nuremberg, Salzburg, Austria, and Ludwig Boltzmann Inst. for Arthritis and Rehabilitation, Paracelsus Medical University Salzburg & Nuremberg, Salzburg, Austria, and Chondrometrics GmbH, Ainring, Germany (W. W.); and the University of Arizona Arthritis Center, University of Arizona College of Medicine - Tucson, AZ, USA (Z.-M. L., C. K. K.)

Abstract

Purpose: To compare the evaluation metrics for deep learning methods that were developed using imbalanced imaging data in osteoarthritis (OA) studies.

Materials and Methods: This retrospective study utilized 2996 sagittal intermediate-weighted (IW) fat-suppressed (FS) knee MRIs with MRI Osteoarthritis Knee Score (MOAKS) readings from 2467 participants in the Osteoarthritis Initiative (OAI) study. We obtained probabilities of the presence of bone marrow lesions (BMLs) from MRIs in testing dataset at the sub-region (15 sub-regions), compartment, and whole-knee levels based on the trained deep learning models. We compared different evaluation metrics (e.g., receiver operating characteristic (ROC) and precision-recall (PR) curves) in the testing dataset with various class ratios (presence of BMLs vs. absence of BMLs) at these three data levels to assess the model's performance.

Results: In a subregion with an extremely high imbalance ratio, the model achieved an ROC-AUC of 0.84, a PR-AUC of 0.10, a sensitivity of 0, and a specificity of 1.

Conclusion: The commonly used ROC curve is not sufficiently informative, especially in the case of imbalanced data. We provide the following practical suggestions based on our data analysis: 1) ROC-AUC is recommended for balanced data, 2) PR-AUC should be used for moderately imbalanced data (i.e., when the proportion of the minor class is above 5% and less than 50%), and 3) for severely imbalanced data (i.e., when the proportion of the minor class is below 5%), it is not practical to apply a deep learning model, even with the application of techniques addressing imbalanced data issues.

Keywords

Osteoarthritis; Bone marrow lesion; Imbalanced data; Deep learning; Receiver operating characteristic; Precision recall curve

1. Introduction

Osteoarthritis (OA) is a leading cause of mobility limitations and disability in the aging population, affecting more than 50 million people in the United States(1). Prior studies have shown that MRI-defined features (e.g., bone marrow lesions (BMLs)) often appear before irreversible cartilage degeneration occurs(2). Thus, rapid detection and quantification of these features may allow the identification of individuals at high risk of knee OA and pain(3). One popular approach for accurate characterization is through a semi-quantitative (SQ) scoring system (e.g., MRI Osteoarthritis Knee Score (MOAKS)) on MRI(4). However, such SQ scoring systems require experienced and trained musculoskeletal radiologists to perform these MRI readings, which is time-consuming, expensive, and resource-intensive.

Recently, deep learning (DL) methods (e.g., convolutional neural network (CNN) models) have been implemented to automatically predict MRI-defined abnormalities from MRIs(5). Compared with the statistical methods, DL-based methods have higher prediction accuracy and need fewer model assumptions. The area under the receiver operating characteristic (ROC) curve is often utilized to assess the performance of a DL-based binary classifier for classifying positive (presence of MRI-defined features) and negative classes. The area under the ROC curve (ROC-AUC) ranges from 0 to 1, with 1 indicating the perfect prediction. In general, a binary classifier with a ROC-AUC value of 0.8 to 0.9 is considered excellent and has an outstanding performance with a value of more than 0.9. However, ROC-AUC alone is not sufficiently informative to evaluate the performance of approaches when the underlying data have class imbalance problems, for example, when the positive class occurs with a markedly reduced frequency, as ROC-AUC is insensitive to the changes in class imbalance ratios between positive and negative classes in the data, resulting in uninformative evaluation of model performance(6). Since most of the datasets from large-scale OA studies are imbalanced(7, 8), it is imperative to examine the evaluation metrics and benchmark DL methods when balanced and/or imbalanced OA images are utilized in DL methods.

The objective of this study was to compare several metrics for evaluating the performance of DL-based models using the data with various class imbalance ratios. To perform this comparison, we implemented and modified an existing DL framework to predict the presence of BMLs using MRIs from the Osteoarthritis Initiative (OAI) study.

2. Materials and Method

2.1. Datasets

This retrospective study used the MRIs with available MOAKS BML size grades at baseline collected through the OAI, a longitudinal study, which has the largest amount of MRIs on the natural history of those with or at risk of developing knee OA(9). A total of 4,796 participants were enrolled from 2004 to 2006. Among 4,796 participants, MRIs of 2,473 individuals over eight years were assessed by radiologists using the SQ scoring system (i.e., MOAKS). It should be noted that the readings we used were obtained from both case-control studies (e.g., FNIH) and case-cohort studies. However, since the same data were used for all metrics, the impact of the different study designs on performance evaluation is likely to be minimal. After removing 6 participants without MOAKS BML size grades at

baseline, the sagittal intermediate-weighted (IW) and fat-suppressed (FS) MRI data of 2,467 participants (mean age, 61 years [range, 45 to 79 years]; 1468 women) were utilized in the data analysis. subchondral bone marrow signal alterations are characterized by ill-defined subchondral areas of high signal intensity on these MRIs(2). We used the BML size grades, which represent the percentage of volume relative to the size of the sub-region, including associated cysts(4). The dataset was split into a training dataset (1838 exams, 1500 subjects), a validation dataset (582 exams, 480 subjects), and a testing dataset (576 exams, 487 subjects) by random sampling.

2.2. Methods

DL framework.—We applied a DL framework based on MRNet(8) to the preprocessed MRIs. The details of data preprocessing and DL configurations can be found in Supplemental Methods. We dichotomized the MOAKS BML size grades into presence or absence categories. The split was done by categorizing grades > 0 as presence (i.e., positive class) and grades = 0 as absence (i.e., negative class). The images with the dichotomized MOAKS BML size grades were used in the model training process.

Sub-region level, compartment-level, and whole-knee level prediction.—We obtained predicted BML statuses (presence vs. absence) for each of the 15 sub-regions from the trained DL models. With the above prediction of 15 sub-regions as independent variables and overall MOAKS BML size grades as the dependent variable, which was dichotomized as 0 (all 15 sub-regions have MOAKS BML size grades of 0) and 1 (at least one of 15 sub-regions has MOAKS BML size grades of 1), a logistic regression model was implemented to predict the BML status at the whole-knee level. The compartment-level prediction is similar to the whole-knee level prediction procedure. This prediction was performed by the scikit-learning (version 0.24.1) package in Python (version 3.6.5). Details of sub-region and compartment definition are shown in Supplementary Material.

Performance evaluation.—We used several evaluation metrics, including ROC-AUC, Precision-Recall (PR)-AUC, Precision-Recall Gain (PRG)-AUC, F1 score, and Matthews correlation coefficient (MCC), precision, sensitivity, and specificity(10–13). We evaluated the performance of DL models based on these metrics in the testing dataset. The definitions of these metrics are shown in Supplemental Methods. We also used the Pearson correlation coefficient to measure the correlation estimates between evaluation metrics at the sub-region level.

3. Results

The baseline demographic statistics of selected participants are shown in Table 1. A majority of subjects (i.e., 56.67%−61.67% in different datasets) were female. The mean age and BMI were approximately 61 years and 28 kg/m^2, respectively. In Table 2, we report the results of different performance metrics at the whole-knee, compartment, and sub-region levels. At the whole-knee level, the ROC-AUC value was 0.86, indicating excellent performance. The model also had a PR-AUC value of 0.96, an F1 score of 0.88, a sensitivity value of 0.88, and a precision value of 0.86. However, a different conclusion about the model performance

at the whole-knee level could be drawn according to the values of PRG-AUC (0.42) and MCC (0.36). As shown in Figure 1A, the imbalance class ratio (positives/negatives) for the whole-knee level data was 0.79/0.21. If BMLs were present in one of the 15 sub-regions, the class label for this knee was positive. Thus, the negative class was the minority class in the whole-knee level imbalanced data. The specificity value was 0.43, indicating a limited ability to predict negatives for the whole-knee model, which was also confirmed by the MCC value. Since the PRG curve is adjusted by the proportion of positives, its AUC value might be affected when the data are imbalanced (i.e., the proportion of positives is close to zero or one). As shown in Figure 2A, the PRG curve had the smallest AUC value.

At the compartment level, there was a mild class imbalance issue (i.e., about 60% negative) in the lateral compartment (see Figure 1B). Although the imbalance class ratio was around 1:2 for the lateral compartment, there were still a sufficient number of positives to train and tune the DL model. The class ratios were approximately 1:1 for the medial and patella compartments. Therefore, all metrics presented consistent information about the model performance for each compartment. For instance, the DL model obtained better performance in the lateral and patella compartments than in the medial compartment.

At the sub-region level, data were severely imbalanced, with less than 10% positives in some sub-regions (see Figure 1C). The PR-AUC and MCC had the highest correlation value (0.92), whereas the correlation value between ROC-AUC and MCC was 0.74. The metrics F1 and MCC had consistent conclusions in terms of model performance, with a correlation value of 0.98. The ROC-AUC values of the FemAntMed, FemPostMed, TibPostLat, and PatellaMed sub-regions were below 0.8. The results of these metrics (i.e., MCC, F1, PR-AUC) were consistent with those of ROC-AUC. However, PRG-AUC for the TibPostLat sub-region (i.e., 0.9) was too high due to its dependence on the proportion of positives. Based on the value of MCC, the prediction performance for this sub-region might not be considered outstanding. The ROC-AUC values of the remaining sub-regions were above 0.8. For example, for the PatellaMed sub-region, the classification results were not as good as those in other sub-regions. The ROC, PR, and PRG curves provided complementary information, see Figure 2B. The ROC-AUC value was 0.84 in the TibAntLat sub-region, showing excellent prediction performance. However, the value of PR-AUC was only 0.10 in the TibAntLat sub-region (see Figure 2C). For this sub-region, PR-AUC was more informative since the precision and sensitivity were zeros, indicating that all data were assigned to the negative class. The MCC and F1 metrics were also good indicators for the performance evaluation for classifying the positive class, with the values of zeros. This might be due to the high imbalance ratio (46 positives:1223 negatives) in the data for the TibAntLat sub-region. Four methods addressing the imbalance issues have been implemented. The results in Supplementary Results show that the improvement was trivial after these methods were applied. We also compared the evaluation results for females and males at the whole-knee level in Supplementary Material(14). The gender differences were also not significant in this study.

4. Discussion

Recently, numerous DL methods have been applied to automatically detect different MRI-defined abnormalities in OA. The development and validation of accurate DL frameworks to detect MRI-defined abnormalities are critical for the identification of early pre-radiographic OA and optimal participant screening for clinical trials of disease-modifying OA drugs (DMOAD). The ultimate application of these DL models is dependent on informative reporting of the performance metrics used to benchmark different DL models, particularly when the underlying MRI data used to derive these models are imbalanced. To compare the performance metrics of DL methods for imbalanced and/or balanced MRI data in OA studies, we have modified an existing DL framework (MRNet) to automatically detect BMLs from the MRIs.

We have demonstrated that there were class imbalance issues in the OAI. In the prior OA studies, the class ratios in the data vary from 1:3 to 1:30(7, 8). In the OAI study, the severity of class imbalance was associated with the data levels (i.e., whole-knee, compartment, and sub-region levels), see details in Supplementary Results. Our findings suggest that class imbalance ratios can affect the evaluation results of some metrics since the DL learners may over-classify the majority group (i.e., assign all minorities to majorities) to obtain high overall accuracy. When data were balanced at the compartment level, all the evaluation metrics had consistent results. With the imbalanced data at the whole-knee and sub-region levels, the commonly used metric ROC-AUC alone might provide a distorted view of the performance of DL methods, demonstrating excellent performance (> 0.8) even for the classifiers that assign all positives to negatives. The ROC curve is necessary but not sufficient to evaluate the performance of DL methods, particularly when data are imbalanced(15). In our case study, the PR curve can be used to address the issues due to the class imbalance. The PRG curve is overly optimistic when data are severely imbalanced, however.

The ROC and PR curves are rank metrics under different threshold settings, while F1 and MCC are threshold metrics for a fixed threshold setting. The ordering of prediction instead of the actual predicted values is used in rank metrics, while the threshold metric depends on a threshold level (e.g. data are predicted as positives above the threshold level). Since the rank metrics provide a summary of model performance for all possible threshold settings, the ROC-AUC and PR-AUC can be used as the major evaluation metrics. The F1 and MCC can be used as auxiliary metrics to evaluate model performance. In addition, the class imbalance ratio should also be reported. In summary, we provide the following practical suggestions based on our data analysis: 1) ROC-AUC is recommended for balanced data, 2) PR-AUC should be used for moderately imbalanced data (i.e. when the proportion of the minor class is above 5%, and less than 50%), and 3) for severely imbalanced data (i.e. when the proportion of the minor class is below 5%), it is not feasible to apply a DL model, even with the application of techniques that attempt to address the imbalanced data issues.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Funding Information

The analyses performed in this study were funded by the NIH grant R01AR078187. This work was performed using publicly available data from the Osteoarthritis Initiative (OAI). The OAI is a public-private partnership comprised of five contracts (N01-AR-2-2258; N01-AR-2-2259; N01-AR-2-2260; N01-AR-2-2261; N01-AR-2-2262) funded by the National Institutes of Health. Funding partners include Merck Research Laboratories; Novartis Pharmaceuticals Corporation, GlaxoSmithKline; and Pfizer, Inc. Scientific and financial support for the Foundation for the National Institutes of Health (FNIH) Osteoarthritis (OA) Biomarkers Consortium has been made possible through grants as well as direct and indirect in-kind contributions from AbbVie, Amgen Inc., the Arthritis Foundation, Bioiberica SA, DePuy Mitek, Inc., Flexion Therapeutics, Inc., GlaxoSmithKline, Merck Serono, Rottapharm | Madaus, Sanofi, Stryker, by a vendor contract from the OAI coordinating center at the University of California, San Francisco (N01-AR-22258), and the Pivotal Osteoarthritis Initiative Magnetic Resonance Imaging Analyses (POMA) Study (NIH/National Heart, Lung, and Blood Institute grant HHSN-2682010000). Private sector funding for the Biomarkers Consortium and the OAI is managed by the FNIH. This work also used data supported by RO1AR065473 and R01AR066601.

8. References

1. Deshpande BR, Katz JN, Solomon DH, Yelin EH, Hunter DJ, Messier SP, et al. Number of persons with symptomatic knee osteoarthritis in the US: impact of race and ethnicity, age, sex, and obesity. Arthritis Care Research (Hoboken). 2016;68(12):1743–1750. doi:10.1002/acr.22897
2. Roemer FW, Frobell R, Hunter DJ, Crema MD, Fischer W, Bohndorf K, et al. MRI-detected subchondral bone marrow signal alterations of the knee joint: terminology, imaging appearance, relevance, and radiological differential diagnosis. Osteoarthritis Cartilage. 2009;17(9):1115–1131. doi:10.1016/j.joca.2009.03.012 [PubMed: 19358902]
3. Tanamas SK, Wluka AE, Pelletier JP, Pelletier JM, Abram F, Berry PA, et al. Bone marrow lesions in people with knee osteoarthritis predict progression of disease and joint replacement: a longitudinal study. Rheumatology (Oxford). 2010;49(12):2413–2419. doi:10.1093/rheumatology/keq286 [PubMed: 20823092]
4. Hunter DJ, Guermazi A, Lo GH, Grainger AJ, Conaghan PG, Boudreau RM, et al. Evolution of semi-quantitative whole joint assessment of knee OA: MOAKS (MRI osteoarthritis knee score). Osteoarthritis Cartilage. 2011;19:990–1002. doi: 10.1016/j.joca.2011.05.004. [PubMed: 21645627]
5. Liu S, Sun X, Roemer F, Ashbeck EL, Bedrick EJ, Li ZM, et al. Automatic detection of bone marrow lesions from knee MRI data from the OAI study [abstract]. ACR Convergence 2021. Arthritis Rheumatology. 2021;73 (suppl 9).
6. Brabec J, Komárek T, Franc V, Machlica L. On model evaluation under non-constant class imbalance. Computational Science – International Conference on Computational Science (ICCS) 2020. 2020;12140:74–87. doi:10.1007/978-3-030-50423-6_6
7. Namiri NK, Lee J, Astuto B, Liu F, Shah R, Majumdar S, et al. Deep learning for large scale MRI-based morphological phenotyping of osteoarthritis. Scientific Reports. 2021;11. 10.1038/s41598-021-90292-6
8. Bien N, Rajpurkar P, Ball RL, Irvin J, Park A, Jones E, et al. Deep-learning-assisted diagnosis for knee magnetic resonance imaging: development and retrospective validation of MRNet. PLoS Medicine. 2018;15(11):e1002699. doi: 10.1371/journal.pmed.1002699.
9. Peterfy CG, Schneider E, Nevitt M. The osteoarthritis initiative: report on the design rationale for the magnetic resonance imaging protocol for the knee. Osteoarthritis Cartilage. 2008;16(12):1433–1441. doi:10.1016/j.joca.2008.06.016 [PubMed: 18786841]
10. Raghavan V, Bollmann P, Jung GS. A critical investigation of recall and precision as measures of retrieval system performance. Association for Computing Machinery (ACM) Transactions on Information Systems. 1989;7;3:205–229. 10.1145/65943.65945.
11. Flach PA, Kull M. Precision-Recall-Gain curves: PR analysis done right. In Proceedings of the 28th International Conference on Neural Information Processing Systems. 2015;1:838–846.
12. Tharwat A. Classification assessment methods. Applied Computing and Informatics. 2018;17(1): 168–192. 10.1016/j.aci.2018.08.003

13. Chicco D, Tötsch N, Jurman G. The matthews correlation coefficient (Mcc) is more reliable than balanced accuracy, bookmaker informedness, and markedness in two-class confusion matrix evaluation. BioData Mining. 2021;14. 10.1186/s13040-021-00244-z

14. Christodoulou E, Moustakidis S, Papandrianos N, Tsaopoulos D, Papageorgiou E. Exploring deep learning capabilities in knee osteoarthritis case study for classification. 10th International Conference on Information, Intelligence, Systems and Applications (IISA). 2019; 1–6. doi: 10.1109/IISA.2019.8900714.

15. Davis J, Goadrich M. The relationship between precision-recall and ROC curves. In Proceedings of the 23rd international conference on Machine learning (ICML ‘06). 2006;233–240. 10.1145/1143844.1143874

Figure 1.

Class ratios (negative (red) : positive (cyan)) at A) whole-knee level; B) compartment level; C) sub-region level

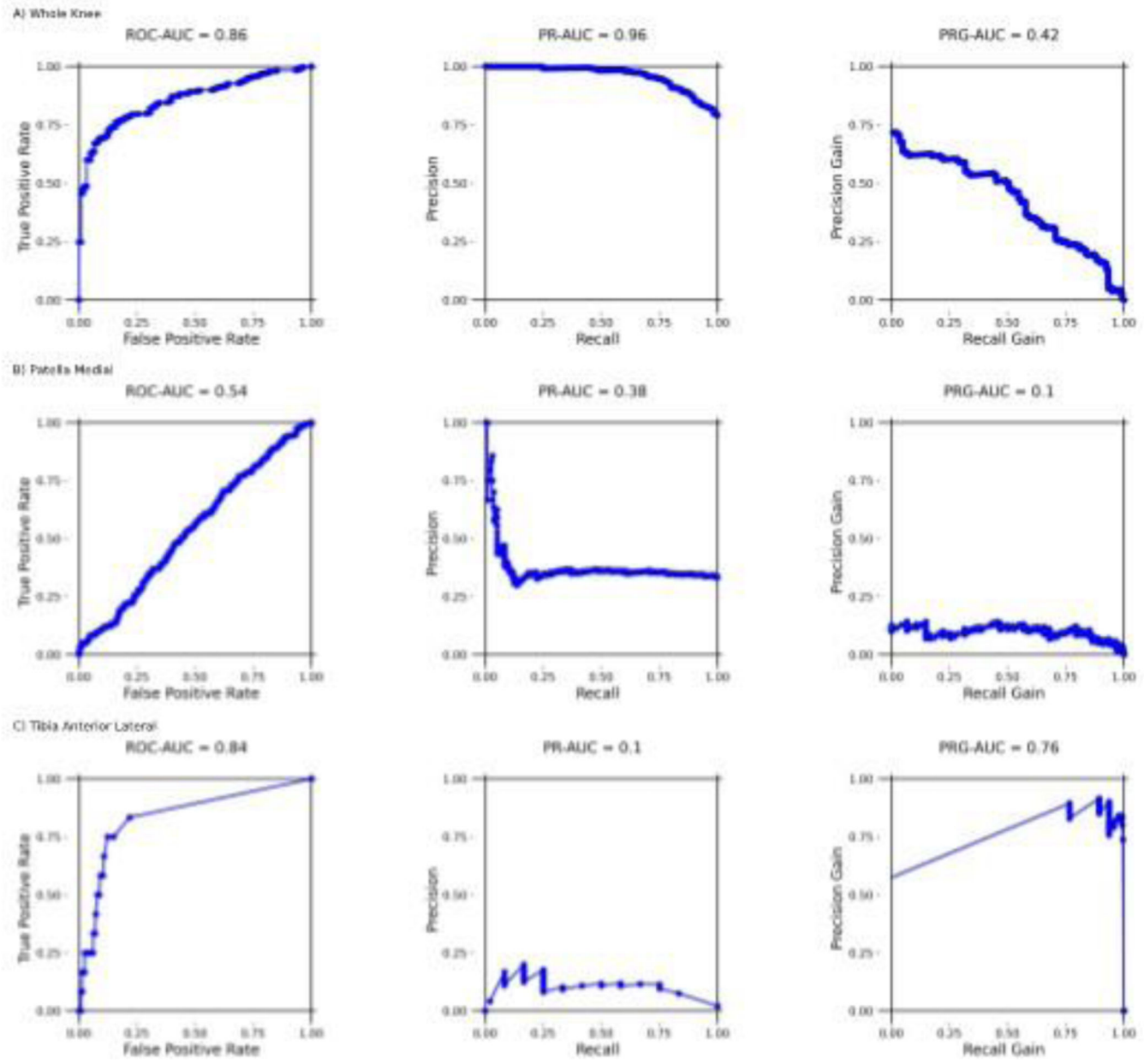

Figure 2.

Performance evaluation of deep learning by ROC (left panels), PR (middle panels), and PRG (right panels) curves at A) whole-knee level; B) PatellaMed; C) TibAntLat. The AUC values of the curves are shown on top of each figure.

Table 1.

Summary statistics of training, validation, and testing datasets

	Training	Validation	Testing
Exams, n	1838	582	576
Participants, n	1500	480	487
Female, n (%)	896 (59.73)	296 (61.67)	276 (56.67)
Age, mean (SD)	61.12 (8.80)	61.02 (8.88)	61.41 (8.91)
BMI, mean (SD)	28.87 (4.72) *	28.95 (4.77)	28.94 (4.81)

* One observation is missing

Table 2.

Performance evaluation at the whole-knee, compartment, and sub-region levels in the testing dataset

	ROC-AUC	PR-AUC	PRG-AUC	F1	MCC	Precision	Sensitivity (Recall)	Specificity
Whole-knee	0.86 (0.83, 0.89)	0.96 (0.95, 0.97)	0.42 (0.31, 0.55)	0.88 (0.86, 0.90)	0.36 (0.25, 0.45)	0.86 (0.83, 0.88) 409/478	0.90 (0.87, 0.93) 409/455	0.43 (0.34, 0.51) 52/121
Medial	0.79 (0.75, 0.82)	0.82 (0.78, 0.85)	0.62 (0.51, 0.71)	0.69 (0.64, 0.73)	0.48 (0.40, 0.54)	0.79 (0.73, 0.84) 167/212	0.61 (0.55, 0.66) 167/274	0.85 (0.81, 0.89) 257/302
Lateral	0.87 (0.84, 0.90	0.77 (0.70, 0.83)	0.86 (0.81, 0.91)	0.71 (0.66, 0.76)	0.59 (0.52, 0.67)	0.71 (0.64, 0.77) 121/171	0.72 (0.65, 0.78) 121/168	0.88 (0.85, 0.91) 358/408
Patella	0.78 (0.74, 0.81)	0.80 (0.75, 0.84)	0.59 (0.49, 0.69)	0.69 (0.64, 0.74)	0.46 (0.39, 0.53)	0.78 (0.73, 0.83) 173/223	0.62 (0.56, 0.68) 174/282	0.83 (0.79, 0.88) 245/294
FemAntLat	0.95 (0.92, 0.97)	0.85 (0.79, 0.90)	0.98 (0.96, 0.99)	0.61 (0.57, 0.65)	0.55 (0.50, 0.60)	0.45 (0.42, 0.49) 100/222	0.95 (0.90, 0.99) 100/105	0.74 (0.70, 0.78) 349/471
FemAntMed	0.66 (0.60, 0.71)	0.32 (0.27, 0.39)	0.46 (0.29, 0.59)	0.12 (0.06, 0.20)	0.11 (0.01, 0.21)	0.45 (0.24, 0.67) 9/20	0.07 (0.03, 0.12) 9/125	0.98 (0.96, 0.99) 440/451
FemCentLat	0.92 (0.86, 0.98)	0.63 (0.45, 0.78)	0.99 (0.98, 1)	0.26 (0.07, 0.46)	0.38 (0.19, 0.54)	1 (1, 1) 4/4	0.15 (0.04, 0.3) 4/27	1 (1, 1) 549/549
FemCentMed	0.91 (0.87, 0.94)	0.82 (0.77, 0.87)	0.95 (0.93, 0.97)	0.74 (0.67, 0.80)	0.68(0.60, 0.75)	0.82 (0.76, 0.89) 84/102	0.67 (0.58, 0.75) 84/126	0.96 (0.94, 0.98) 432/450
FemPostLat	0.96 (0.90, 1)	0.79 (0.59, 0.93)	1 (0.99, 1)	0.67 (0.44, 0.84)	0.68 (0.50, 0.84)	0.91 (0.71, 1) 10/11	0.53 (0.32, 0.74) 10/19	1 (0.99, 1) 556/557
FemPostMed	0.79 (0.74, 0.85)	0.36 (0.27, 0.47)	0.8 (0.69, 0.88)	0.34 (0.22, 0.44)	0.28 (0.15, 0.40)	0.44 (0.30, 0.59) 19/43	0.27 (0.17, 0.37) 19/70	0.95 (0.93, 0.97) 482/506
TibSubSpCent	0.87 (0.84, 0.90)	0.76 (0.70, 0.82)	0.87 (0.81, 0.91)	0.67 (0.60, 0.72)	0.55 (0.47, 0.62)	0.72 (0.66, 0.79) 105/145	0.62 (0.54, 0.69) 105/170	0.90 (0.87, 0.93) 366/406
TibAntLat	0.84 (0.71, 0.96)	0.10 (0.06, 0.20)	0.76 (0, 0.91)	0 (0, 0)	0 (0, 0)	0 (0, 0) 0/0	0 (0, 0) 0/12	1 (1, 1) 564/564
TibAntMed	0.89 (0.84, 0.93)	0.63 (0.52, 0.74)	0.96 (0.92, 0.98)	0.62 (0.52, 0.71)	0.57 (0.47, 0.68)	0.65 (0.54, 0.76) 39/60	0.59 (0.47, 0.71) 39/66	0.96 (0.94, 0.98) 489/510
TibCentLat	0.86 (0.80, 0.92)	0.56 (0.43, 0.71)	0.96 (0.92, 0.98)	0.54 (0.44, 0.64)	0.49 (0.38, 0.61)	0.5 (0.41, 0.61) 32/64	0.59 (0.46, 0.72) 32/54	0.94 (0.92, 0.96) 490/522
TibCentMed	0.94 (0.91, 0.97)	0.85 (0.78, 0.91)	0.97 (0.95, 0.99)	0.78 (0.72, 0.83)	0.73 (0.66, 0.80)	0.74 (0.67, 0.81) 82/111	0.83 (0.75, 0.89) 82/99	0.94 (0.92, 0.96) 448/477
TibPostLat	0.76 (0.68, 0.84)	0.28 (0.17, 0.45)	0.9 (0.75, 0.96)	0.37 (0.23, 0.51)	0.33 (0.18, 0.49)	0.42 (0.27, 0.60) 13/31	0.33 (0.18, 0.49) 13/39	0.97 (0.95, 0.98) 519/537
TibPostMed	0.84 (0.77, 0.92)	0.33 (0.21, 0.47)	0.88 (0.74, 0.97)	0.14 (0, 0.28)	0.15 (0.02, 0.33)	0.38 (0, 0.75) 3/8	0.08 (0, 0.19) 3/36	0.99 (0.98, 1) 535/540
PatellaLat	0.88 (0.85, 0.91)	0.78 (0.72, 0.83)	0.9 (0.85, 0.94)	0.70 (0.65, 0.75)	0.59 (0.52, 0.66)	0.66 (0.61, 0.72) 115/174	0.75 (0.69, 0.82) 115/153	0.86 (0.83, 0.89) 364/423

	ROC-AUC	PR-AUC	PRG-AUC	F1	MCC	Precision	Sensitivity (Recall)	Specificity
PatellaMed	0.54 (0.49, 0.59)	0.38 (0.33, 0.43)	0.1 (0, 0.23)	0.03 (0, 0.07)	0.07 (0.03, 0.15)	0.75 (0, 1) 3/4	0.02 (0, 0.04) 3/192	1 (0.99, 1) 383/384