This MICCAI paper is the Open Access version, provided by the MICCAI Society. It is identical to the accepted version, except for the format and this watermark; the final published version is available on SpringerLink.

# Supplementary Material: Prototypical Severity Representation Learning for Automated Calcium Pyrophosphate Deposition Disease Assessment

## 1 Additional Dataset and Evaluation Details

### 1.1 Dataset Composition

The retrospective dataset comprised 2,172 knee radiographs annotated according to a four-grade CPPD severity scale (Grades 0–3). A patient-level split was used to obtain training, validation, and test sets. The exclusion criteria were age $< 18$ years, previous knee arthroplasty, fixation devices, and fractures involving the knee or adjacent bones. Table 1 reports the class distribution across the three partitions.

**Table 1.** CPPD severity distribution across the training, validation, and test sets. Percentages in the last column refer to the complete dataset.

| CPPD grade | Train | Validation | Test | Total (%) |
|---|---|---|---|---|
| Grade 0 | 464 | 116 | 236 | 816 (37.6%) |
| Grade 1 | 215 | 52 | 106 | 373 (17.2%) |
| Grade 2 | 264 | 71 | 145 | 480 (22.1%) |
| Grade 3 | 291 | 66 | 146 | 503 (23.1%) |
| **Total** | **1,234** | **305** | **633** | **2,172 (100%)** |

### 1.2 Evaluation Protocol

Each model configuration was trained with 10 independent random seeds, which control weight initialization, data shuffling, and stochastic augmentation; the train/validation/test partition was held fixed across runs. All metrics are computed on the held-out test set and reported as mean $\pm$ standard deviation over the 10 runs. We considered Quadratic Weighted Kappa (QWK), Weighted F1 (WF1), Macro F1 (MF1), Accuracy, and Mean Absolute Error (MAE).

*Ordinal metrics.* QWK was adopted as the primary evaluation metric because it accounts for the ordered nature of CPPD severity: its quadratic disagreement weights grow with the squared distance between the ground-truth and predicted grades, so that confusions between distant grades are penalized far more heavily

than those between adjacent ones. MAE provides a complementary distance-aware measure by quantifying the average absolute grading error, with severity grades treated as integer scores. Both metrics therefore distinguish errors between neighbouring grades from larger severity disagreements.

*Classification metrics.* Given the imbalanced class distribution, WF1 accounts for class prevalence, whereas MF1 assigns equal importance to each severity grade. Accuracy is additionally reported as a conventional exact-match measure, enabling comparison with standard categorical evaluation.

*Statistical analysis.* Uncertainty in model performance using classification accuracy and QWK was quantified via 95% non-parametric bootstrap confidence intervals. Because metric distributions (particularly QWK) can exhibit skewness in finite samples, we avoided asymptotic normality assumptions and instead resampled the hold-out test set with replacement across 10,000 iterations. Two-sided 95% intervals were obtained via the standard percentile method (2.5th and 97.5th percentiles). Non-overlapping confidence intervals were considered indicative of statistically meaningful performance differences between model configurations.

*Representation quality.* The penultimate-layer embeddings of the test set were further evaluated with respect to the ground-truth severity grades using the Silhouette score and the Calinski–Harabasz index (higher is better) and the Davies–Bouldin index (lower is better), all computed under Euclidean distance.