

AI-based diagnosis of hallux valgus: evaluation of measurements derived from foot X-ray images

Mahmut Kurtboğan¹, Gülçin Ray², Abdullah Ray², İbrahim Kürtül², Mihriban Karatop³

¹Department of Orthopedics and Traumatology, Faculty of Medicine, Bolu Abant İzzet Baysal University, Bolu, Türkiye

²Department of Anatomy, Faculty of Medicine, Bolu Abant İzzet Baysal University, Bolu, Türkiye

³Faculty of Medicine, Bolu Abant İzzet Baysal University, Bolu, Türkiye

Cite as: Kurtboğan M, Ray G, Ray A, Kürtül İ, Karatop M. AI-based diagnosis of hallux valgus: evaluation of measurements derived from foot X-ray images. Northwestern Med J. 2026;6(4):330-342.

ABSTRACT

Objective: This study aimed to diagnose hallux valgus (HV) using machine learning algorithms (MLAs) applied to angular measurements obtained from foot X-ray images.

Materials and Methods: The study was a retrospective observational investigation. The sample comprised 50 patients diagnosed with HV and 50 controls. The intermetatarsal angle (IMA), interphalangeal angle (IPA), talometatarsal angle (TMTA), metatarsal tangent angle (MTTA), metatarsus adductus angle (MTAA), and metatarsal parabola angle (MTPA) were measured on the anteroposterior X-ray image, and calcaneal inclination angle (CIA) and the lateral talocalcaneal angle (LTCA) were measured on the lateral X-ray image. The following algorithms were used as machine learning algorithms (MLAs): Logistic Regression (LR), Support Vector Machine (SVM), Random Forest (RF), Gradient Boosting (GB), XGBoost (XGB), LightGBM (LGBM), AdaBoost (ADA), Extra Trees (ET), K-Nearest Neighbor (KNN), and Linear Discriminant Analysis (LDA). Statistical analyses were performed using the open-source RStudio software.

Results: The diagnostic accuracy rates of the algorithms (CV) were 90% for XGB, ET, RF, LGBM, and LDA; 85% for ADA, LR, SVM, and GB; and 80% for KNN. The mean $\pm$ SD values of the variables for the control and patient groups, respectively, were calculated as 21.71 ± 1.83 and 22.52 ± 3.32 for IMA; 17.86 ± 1.79 and 18.35 ± 2.28 for MTAA; 143.43 ± 9.20 and 143.25 ± 9.14 for MTPA; and 50.96 ± 5.27 and 49.42 ± 6.98 for LTCA. The median (min-max) values were: for IPA, 36.48 (25.73–45.82) and 45.41 (29.60–76.90); for TMTA, 7.60 (5.91–11.45) and 12.54 (6.62–19.50); for MTTA, 7.93 (1.07–13.20) and 7.13 (2.60–12.07); and for CIA, 28.52 (22.31–37.65) and 27.19 (10.85–47.46).

Conclusions: Angular measurements derived from X-ray images enabled MLAs to diagnose HV with an accuracy of up to 90%. This finding demonstrates the applicability of artificial intelligence for diagnostic purposes in imaging-based applications.

Keywords: hallux valgus deformity, machine learning algorithms, foot anatomy, x-ray images, intermetatarsal angle

Corresponding author: Mahmut Kurtboğan **E-mail:** kurtbogann@gmail.com

Received: 25.06.2026 **Accepted:** 27.07.2026 **Published:** 25.10.2026

Copyright © 2026 The Author(s). This is an open-access article published by Bolu İzzet Baysal Training and Research Hospital under the terms of the [Creative Commons Attribution License \(CC BY\)](#) which permits unrestricted use, distribution, and reproduction in any medium or format, provided the original work is properly cited.

INTRODUCTION

Hallux valgus (HV), one of the most common foot deformities, develops in general with the progressive subluxation of the metatarsophalangeal joint at the pollex, medial deviation of the metatarsal bone, and medial displacement of the pollex (1). A bony prominence developing on the medial aspect of the first metatarsal is the characteristic appearance of the deformity (1,2). A study has reported that 23% of the adult population and 36% of the geriatric population are diagnosed with HV (2). The incidence of the disease is higher in women and increases with age (3). The biomechanical cause of the pathology can be described as the deviation of the bones and joints in the anteromedial region of the foot from their normal axes, leading to impaired muscular stabilization and ligamentous laxity (3,4). This biomechanical disturbance is accompanied by pain, impairment in physical and daily-life activities, and an increased risk of falls due to disturbed balance (4). Because the severity of deformity progresses over time, early diagnosis and intervention are essential (5). An increase in deformity severity corresponds then to an increase in the HV (5,6). Although several methods such as ultrasound, footprint analysis, and gait and plantar pressure analysis are used to diagnose the deformity, the most accurate and widely used method is radiographic imaging (7). For the diagnosis, orthopedic surgeons rely on the HV and intermetatarsal angle IMA, which can be measured on weight-bearing anteroposterior radiographs of the foot (5–7). On the basis of these angles, the deformity is classified as mild, moderate, or severe (8). The HV is measured between the first metatarsal and the proximal phalanx of the pollex, and the IMA is measured between the parallel axes of the first and second metatarsals (7,8). As a surgical intervention in particular is planned, the severity of the deformity, degenerative changes in the joint, and the size and shape of the metatarsal bone are also evaluated (9). After surgical correction, the recurrence rate of HV is considerably high (10). Risk factors such as a high preoperative HV, the presence and degree of subluxation of the sesamoid bone, hereditary predisposition, body mass index, and the patient's preference for narrow, high-heeled footwear, contribute to recurrence (9–11). The rapid progression of disease severity, high recurrence rate despite

surgical correction, and resulting deterioration in quality of life require an easily applicable tool for early diagnosis and follow-up.

A review of the literature reveals that machine learning algorithms (MLAs)—a class of artificial intelligence (AI) applications providing description, classification, and diagnostic functions—have been used on radiographic images for this purpose (12–16). Recently popular AI applications in healthcare perform analysis and classification on medical images, develop predictive models, and predict prognosis with high accuracy (12,13). Algorithms allow large datasets to be processed efficiently, capture correlations among variables, and consequently provide highly accurate predictions (17,18). Nevertheless, further research is needed to enable AI applications to differentiate HV from concomitant deformities such as osteoarthritis, bunionette deformity, or pes planus. With those in mind, the study aimed to demonstrate the accuracy of HV diagnosis by MLAs using angular variables derived from foot X-ray images.

MATERIALS AND METHODS

This retrospective study was initiated following decision No. 2026/217 of the Bolu Abant İzzet Baysal University Non-Interventional Clinical Research Ethics Committee. The study included 50 patients (39 women, 11 men) who presented to the Bolu Abant İzzet Baysal University Department of Orthopedics and Traumatology between January 2020 and February 2026 and were diagnosed with hallux valgus, and 50 controls (29 women, 21 men) who had no history of disease involving the foot. For both groups, individuals with a history of surgical intervention, fracture, trauma, or congenital anomaly in the foot or ankle region, as well as those with unclear or inadequate radiographic images, were excluded. The images of the sample group were retrieved retrospectively from the hospital's PACS archive, transferred to the Radiant DICOM Viewer software, and used to obtain the predetermined angular measurements. An expert orthopedic surgeon classified participants as patients or controls based on HV measurements performed on the images. The intermetatarsal angle (IMA), interphalangeal angle (IPA), talometatarsal angle (TMTA), metatarsal tangent angle (MTTA), metatarsus adductus angle (MTAA),

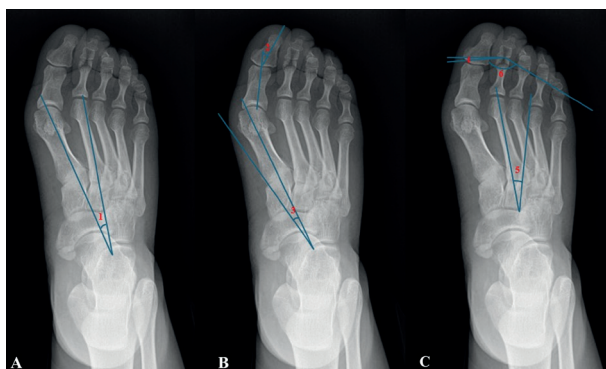


Figure 1. Reference points for angular measurements on the anteroposterior foot radiograph of a 32-year-old male participant. A. intermetatarsal angle; B. interphalangeal angle, talometatarsal angle; C. metatarsal tangent angle, metatarsus adductus angle, metatarsal parabola angle.

and metatarsal parabola angle (MTPA) variables (Figure 1) were measured on the anteroposterior foot radiographs, and the inclination angle (CIA) and the lateral talocalcaneal angle (LTCA) variables were measured on the lateral foot radiographs (Figure 2). All radiographs, both anteroposterior and lateral, were obtained under weight-bearing conditions, with patients standing on the affected foot at the time of image acquisition. To eliminate inter-observer error, each measurement was performed twice by the same investigator, and the mean values were recorded in Excel format.

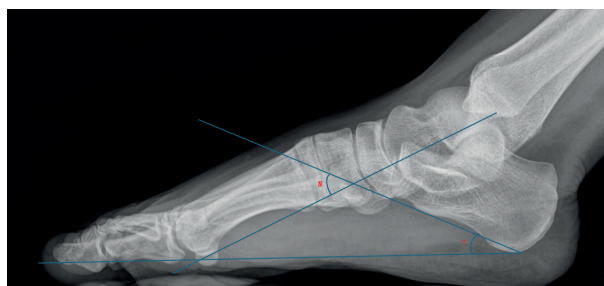


Figure 2. Reference points for angular measurements on the lateral foot radiograph of a 32-year-old male participant. Calcaneal inclination angle; lateral talocalcaneal angle.

Logistic Regression (LR), Support Vector Machine (SVM), Random Forest (RF), Gradient Boosting (GB), XGBoost (XGB), LightGBM (LGBM), AdaBoost (ADA), Extra Trees (ET), K-Nearest Neighbor (KNN), and Linear Discriminant Analysis (LDA) were used as MLAs, and their hyperparameters are shown in Table 1.

Training/test strategy

The dataset was split into training (80%) and test (20%) sets. To preserve class proportions in both subsets, stratified splitting (random state = 42) was applied. For evaluation, 10-fold cross-validation (CV) was used to partition the dataset and measure classification performance. All variables were standardized prior to model training using z-score normalization (StandardScaler: $\mu = 0$, $\sigma = 1$).

Table 1. Best hyperparameters for the MLAs.

Algorithm	Best Hyperparameters (after GridSearchCV)
LR	{'clf__C': 0.01, 'clf__penalty': 'l2', 'clf__solver': 'lbfgs'}
SVM	{'clf__C': 0.1, 'clf__gamma': 'scale', 'clf__kernel': 'rbf'}
RF	{'clf__max_depth': None, 'clf__max_features': 'sqrt', 'clf__min_samples_split': 5, 'clf__n_estimators': 100}
GB	{'clf__learning_rate': 0.05, 'clf__max_depth': 3, 'clf__n_estimators': 100, 'clf__subsample': 1.0}
XGB	{'clf__colsample_bytree': 1.0, 'clf__learning_rate': 0.2, 'clf__max_depth': 2, 'clf__n_estimators': 100, 'clf__subsample': 1.0}
LGBM	{'clf__learning_rate': 0.05, 'clf__max_depth': -1, 'clf__n_estimators': 100, 'clf__num_leaves': 15}
ADA	{'clf__learning_rate': 0.5, 'clf__n_estimators': 50}
ET	{'clf__max_depth': None, 'clf__min_samples_split': 2, 'clf__n_estimators': 100}
KNN	{'clf__metric': 'euclidean', 'clf__n_neighbors': 3, 'clf__weights': 'uniform'}
LDA	{'clf__solver': 'svd'}

ADA: AdaBoost; ET: Extra Trees; GB: Gradient Boosting; KNN: K-Nearest Neighbor; LDA: Linear Discriminant Analysis; LGBM: LightGBM; LR: Logistic Regression; RF: Random Forest; SVM: Support Vector Machine; XGB: XGBoost.

Statistical analyses

Analyses were performed using the open-source RStudio software. A One-way analysis of variance was applied to all variables, and residual normality was assessed using the Anderson–Darling test. Logarithmic and square-root transformations were applied to variables that did not conform to a normal distribution. For pairwise comparisons, parametric variables were analyzed with the Independent Samples t-test, and nonparametric variables were analyzed with the Mann–Whitney U test. Descriptive statistics were calculated for all variables.

RESULTS

The performance of the ten MLAs in distinguishing the HV and control groups is presented in an integrated manner in Table 2 using cross-validation accuracy (CV Accuracy, mean $\pm$ SD), test accuracy, area under the curve (AUC), F1 score, sensitivity (recall), specificity, precision, balanced accuracy, and the Matthews correlation coefficient (MCC). In the CV analysis, mean accuracy values ranged from 0.8500 ± 0.1561 to 0.9125 ± 0.0800 . The highest CV accuracy was obtained for GB at 0.9125 ± 0.0800 , followed by ADA at 0.9000 ± 0.0935 and SVM at 0.9000 ± 0.1225 . ET, RF, and LGBM yielded a mean CV accuracy of 0.8875, with standard deviations of ± 0.1038 , ± 0.1179 , and ± 0.0875 , respectively. Logistic Regression (LR) was 0.8875 ± 0.0875 , XGB was 0.8750 ± 0.1118 , LDA was 0.8750 ± 0.1250 , and KNN was 0.8500 ± 0.1561 . Based on the examined standard deviation values, the lowest variation was observed for GB (0.0800) and the highest for KNN (0.1561), indicating that GB exhibited the most stable performance across CV folds, whereas KNN produced the most variable results.

With respect to test set performance, the models clustered into two distinct groups. XGB, ET, RF, LGBM, and LDA formed the top tier with a test accuracy of 0.9000, while ADA, LR, SVM, and GB achieved 0.8500, and KNN reached 0.8000. In terms of AUC, the highest discriminative capacity was obtained for XGB and ET at 0.9400, followed by RF, LGBM, and LDA at 0.9300; ADA

at 0.9050; LR at 0.9000; SVM at 0.8900; GB at 0.8500; and KNN at 0.8150. In the F1 score analysis, XGB, ET, RF, LGBM, and LDA shared the highest value of 0.9000, whereas ADA, LR, SVM, and GB scored 0.8421, while KNN scored 0.7778.

When sensitivity and specificity were considered together, XGB, ET, RF, LGBM, and LDA displayed a balanced profile, with a value of 0.9000 for both metrics. ADA, LR, SVM, and GB yielded sensitivity of 0.8000 and specificity of 0.9000; KNN yielded sensitivity of 0.7000 and specificity of 0.9000. Precision values were: 0.9000 for XGB, ET, RF, LGBM, and LDA; 0.8889 for the four mid-tier models (ADA, LR, SVM, and GB); and 0.8750 for KNN. Balanced accuracy, reflecting equally weighted performance across classes, was 0.9000 for XGB, ET, RF, LGBM, and LDA; 0.8500 for ADA, LR, SVM, and GB; and 0.8000 for KNN. With respect to the Matthews correlation coefficient (MCC) which evaluates classification quality in a manner sensitive to class imbalance, the highest MCC of 0.8000 was observed for XGB, ET, RF, LGBM, and LDA; ADA, LR, and SVM yielded MCCs of 0.7035; GB yielded 0.6124; and KNN yielded 0.6118. Likewise, the XGBoost algorithm (test accuracy 0.9000; AUC 0.9400; F1 score 0.9000; sensitivity 0.9000; specificity 0.9000; precision 0.9000; balanced accuracy 0.9000; MCC 0.8000) emerged as the highest-performing model in Table 2, providing the most consistent diagnostic performance across all metrics for HV diagnosis. By contrast, the KNN algorithm with a test accuracy of 0.8000, an AUC of 0.8150, an F1 score of 0.7778, and an MCC of 0.6118 delivered the lowest performance across all evaluated parameters and was identified as the model with the most limited diagnostic utility for clinical decision support.

To evaluate models' sample-independent generalization capacity and stability across folds, the ten machine learning algorithms were subjected to 10-fold CV. The accuracy values obtained for each fold (Folds 1–10) are presented in Table 2. For the LR algorithm, fold accuracy values ranged from 0.7500 to 1.0000 (Fold 1: 0.8750; Fold 2: 1.0000; Fold 3: 1.0000; Fold 4: 0.8750; Fold 5: 0.8750; Fold 6: 0.8750;

Table 2. MLA classification results for Hallux Valgus vs Control groups.

Algorithm	CV Accuracy (Mean $\pm$ SD)	Test Accuracy	AUC (ROC)	F1-Score	Sensitivity (Recall)	Specificity	Precision	Balanced Accuracy	MCC
XGB	0.8750 $\pm$ 0.1118	0.9000	0.9400	0.9000	0.9000	0.9000	0.9000	0.9000	0.8000
ET	0.8875 $\pm$ 0.1038	0.9000	0.9400	0.9000	0.9000	0.9000	0.9000	0.9000	0.8000
RF	0.8875 $\pm$ 0.1179	0.9000	0.9300	0.9000	0.9000	0.9000	0.9000	0.9000	0.8000
LGBM	0.8875 $\pm$ 0.0875	0.9000	0.9300	0.9000	0.9000	0.9000	0.9000	0.9000	0.8000
LDA	0.8750 $\pm$ 0.1250	0.9000	0.9300	0.9000	0.9000	0.9000	0.9000	0.9000	0.8000
ADA	0.9000 $\pm$ 0.0935	0.8500	0.9050	0.8421	0.8000	0.9000	0.8889	0.8500	0.7035
LR	0.8875 $\pm$ 0.0875	0.8500	0.9000	0.8421	0.8000	0.9000	0.8889	0.8500	0.7035
SVM	0.9000 $\pm$ 0.1225	0.8500	0.8900	0.8421	0.8000	0.9000	0.8889	0.8500	0.7035
GB	0.9125 $\pm$ 0.0800	0.8500	0.8500	0.8421	0.8000	0.9000	0.8889	0.8500	0.7035
KNN	0.8500 $\pm$ 0.1561	0.8000	0.8150	0.7778	0.7000	0.9000	0.8750	0.8000	0.6124

ADA: AdaBoost; ET: Extra Trees; GB: Gradient Boosting; KNN: K-Nearest Neighbor; LDA: Linear Discriminant Analysis; LGBM: LightGBM; LR: Logistic Regression; RF: Random Forest; SVM: Support Vector Machine; XGB: XGBoost.

Fold 7: 0.7500; Fold 8: 0.7500; Fold 9: 1.0000; Fold 10: 0.8750), and the overall mean was calculated as 0.8875 ± 0.0875 . For the SVM, fold values ranged from 0.6250 to 1.0000 (Fold 1: 1.0000; Fold 2: 1.0000; Fold 3: 1.0000; Fold 4: 0.8750; Fold 5: 1.0000; Fold 6: 0.8750; Fold 7: 0.6250; Fold 8: 0.7500; Fold 9: 1.0000; Fold 10: 0.8750), yielding a mean accuracy of 0.9000 ± 0.1225 . The RF algorithm exhibited its lowest starting value in the first fold (0.6250). The subsequent folds were recorded as Fold 2: 1.0000; Fold 3: 1.0000; Fold 4: 1.0000; Fold 5: 0.8750; Fold 6: 0.8750; Fold 7: 0.8750; Fold 8: 0.7500; Fold 9: 1.0000; Fold 10: 0.8750, producing a mean of 0.8875 ± 0.1179 . The GB model—with fold values of 0.8750, 1.0000, 1.0000, 0.8750, 1.0000, 0.8750, 0.8750, 0.8750, and 0.8750—exhibited the most stable performance and reached a mean accuracy of 0.9125 ± 0.0800 . The fold distribution of XGB, highlighted in yellow in Table 2, contained the values 0.6250 (Fold 1), 1.0000 (Fold 2), 1.0000 (Fold 3), 1.0000 (Fold 4), 0.8750 (Fold 5), 0.8750 (Fold 6), 0.7500 (Fold 7), 0.8750 (Fold 8), 0.8750 (Fold 9), and 0.8750 (Fold 10). The mean CV accuracy of the model was calculated as 0.8750 ± 0.1118 . For the LGBM, fold values ranged from 0.7500 to 1.0000 (0.7500; 1.0000; 1.0000; 0.8750; 0.8750; 0.8750; 0.8750; 0.8750; 1.0000; 0.8750), with a mean accuracy of 0.8875 ± 0.0875 (Figure 3).

For the ADA, fold accuracies of 0.7500, 1.0000, 1.0000, 0.8750, 1.0000, 0.8750, 0.8750, 0.8750, 1.0000, and 0.8750 were obtained, with a mean of 0.9000 ± 0.0935 . The ET produced fold values of 0.8750, 1.0000, 1.0000, 0.8750, 0.8750, 0.8750, 0.8750, 0.8750, 1.0000, and 0.7500, yielding a mean of 0.8875 ± 0.1038 . Likewise, for the KNN, fold values were distributed over a notably wide range, from 0.5000 to 1.0000 (1.0000; 1.0000; 1.0000; 0.7500; 0.7500; 0.7500; 0.7500; 0.5000; 0.8750); the model reached a mean accuracy of 0.8500 ± 0.1561 and exhibited the highest standard deviation among all evaluated algorithms. The LDA algorithm produced fold values of 0.6250, 1.0000, 1.0000, 0.8750, 0.8750, 0.8750, 0.7500, 1.0000, 1.0000, and 0.7500, with a mean performance of 0.8750 ± 0.1250 . Examination of the inter-fold distribution pattern revealed that all ten algorithms achieved an accuracy of 1.0000 in at least one fold; in contrast, the lowest fold accuracy, 0.5000, was observed for KNN in Fold 9. When the standard deviations were ranked from lowest to highest, the order was as follows: GB (0.0800), LGBM (0.0875), LR (0.0875), ADA (0.0935), ET (0.1038), XGB (0.1118), RF (0.1179), SVM (0.1225), LDA (0.1250), and KNN (0.1561) (Table 3 and Figure 4).

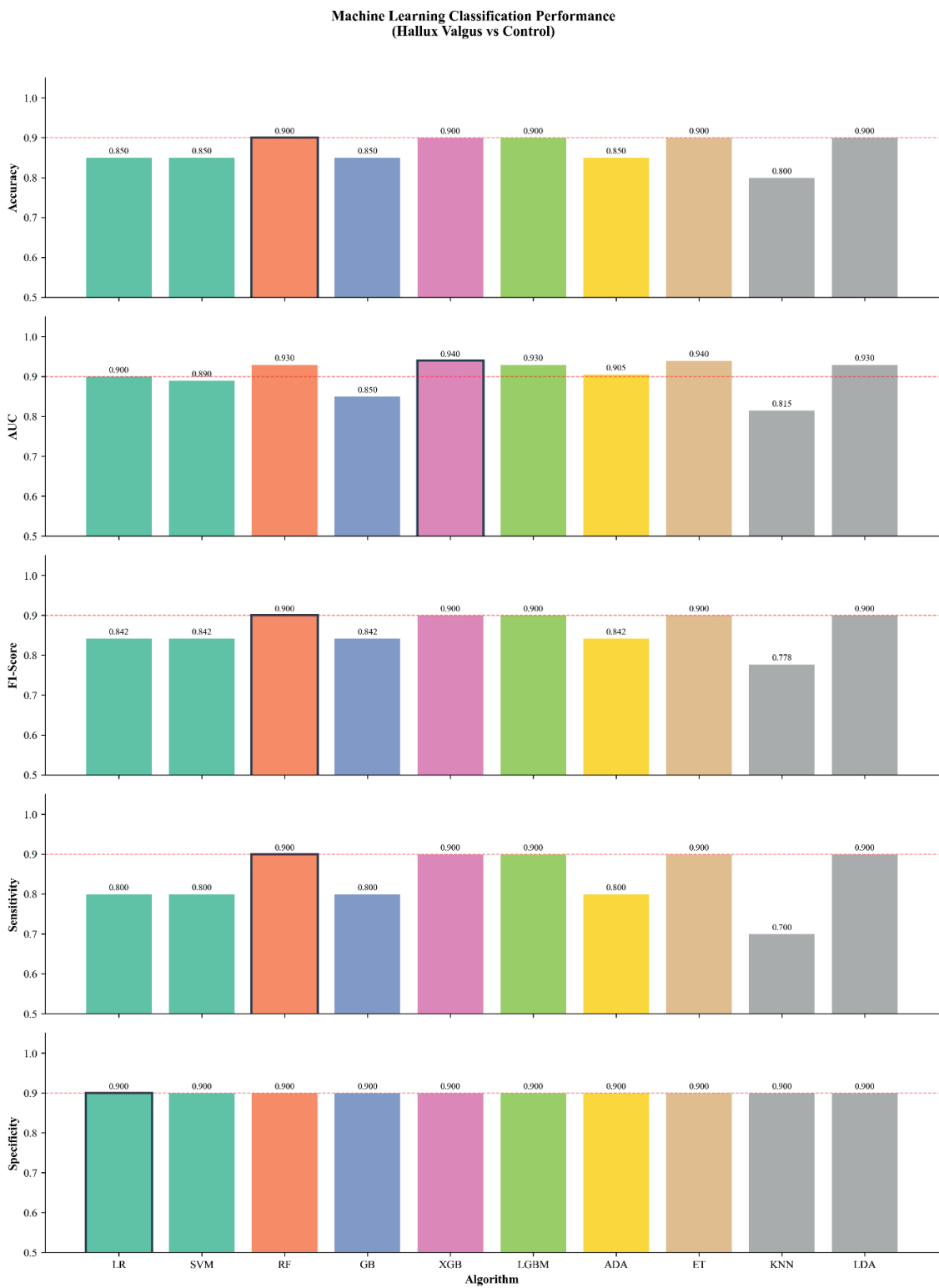
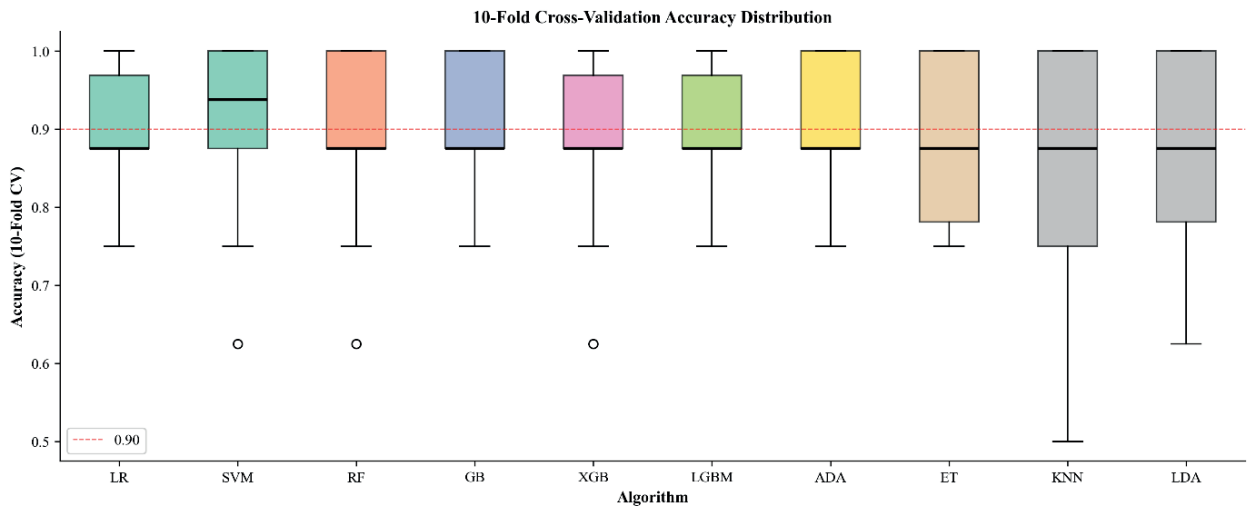


Figure 3. Machine learning algorithms classification performance.

Table 3. 10-fold CV results.

Algorithm	Fold 1	Fold 2	Fold 3	Fold 4	Fold 5	Fold 6	Fold 7	Fold 8	Fold 9	Fold 10	Mean $\pm$ SD
LR	0.8750	1.0000	1.0000	0.8750	0.8750	0.8750	0.7500	0.7500	1.0000	0.8750	0.8875 $\pm$ 0.0875
SVM	1.0000	1.0000	1.0000	0.8750	1.0000	0.8750	0.6250	0.7500	1.0000	0.8750	0.9000 $\pm$ 0.1225
RF	0.6250	1.0000	1.0000	1.0000	0.8750	0.8750	0.8750	0.7500	1.0000	0.8750	0.8875 $\pm$ 0.1179
GB	0.8750	1.0000	1.0000	1.0000	0.8750	1.0000	0.7500	0.8750	0.8750	0.8750	0.9125 $\pm$ 0.0800
XGB	0.6250	1.0000	1.0000	1.0000	0.8750	0.8750	0.7500	0.8750	0.8750	0.8750	0.8750 $\pm$ 0.1118
LGBM	0.7500	1.0000	1.0000	1.0000	0.8750	0.8750	0.8750	0.7500	0.8750	0.8750	0.8875 $\pm$ 0.0875
ADA	0.7500	1.0000	1.0000	1.0000	0.8750	0.8750	0.8750	0.7500	1.0000	0.8750	0.9000 $\pm$ 0.0935
ET	0.8750	1.0000	1.0000	1.0000	0.8750	0.8750	0.7500	0.7500	1.0000	0.7500	0.8875 $\pm$ 0.1038
KNN	0.8750	1.0000	1.0000	1.0000	0.7500	0.7500	0.7500	0.5000	1.0000	0.8750	0.8500 $\pm$ 0.1561
LDA	0.6250	1.0000	1.0000	1.0000	0.8750	0.8750	0.8750	0.7500	1.0000	0.7500	0.8750 $\pm$ 0.1250

ADA: AdaBoost; ET: Extra Trees; GB: Gradient Boosting; KNN: K-Nearest Neighbor; LDA: Linear Discriminant Analysis; LGBM: LightGBM; LR: Logistic Regression; RF: Random Forest; SVM: Support Vector Machine; XGB: XGBoost.

**Figure 4.** Accuracy distribution of the 10-fold cross-validation.

The relationship between sensitivity (true positive rate) and 1 – specificity (false positive rate) for the ten classifiers evaluated to discriminate between the HV and control groups is visualized in Figure 2 using Receiver Operating Characteristic (ROC) curves. The area under the curve AUC values indicated that the models had discriminative capacity significantly greater than the random-guess line (AUC = 0.50). According to the ROC analysis, the highest discriminative power was obtained by XGB and ET, with an AUC of 0.940; their curves were closest to the upper-left corner of the coordinate system, achieving an optimal balance between sensitivity and specificity. They were

followed by RF, LGBM, and LDA, each with an AUC of 0.930; these three algorithms exhibited closely aligned performance. The ADA (0.905), LR (0.900), and SVM (0.890) were ranked among the algorithms with high discriminative capacity. By contrast, GB displayed lower performance, with an AUC of 0.850, while KNN provided the weakest discriminative power among all models, with an AUC of 0.815. The ROC curve for KNN was noticeably closer to the random-guess line than the ROC curves of the other models. Overall, the AUC values ranged from 0.815 to 0.940, and nine models exceeded the 0.850 threshold, indicating clinically acceptable diagnostic performance (Figure 5).

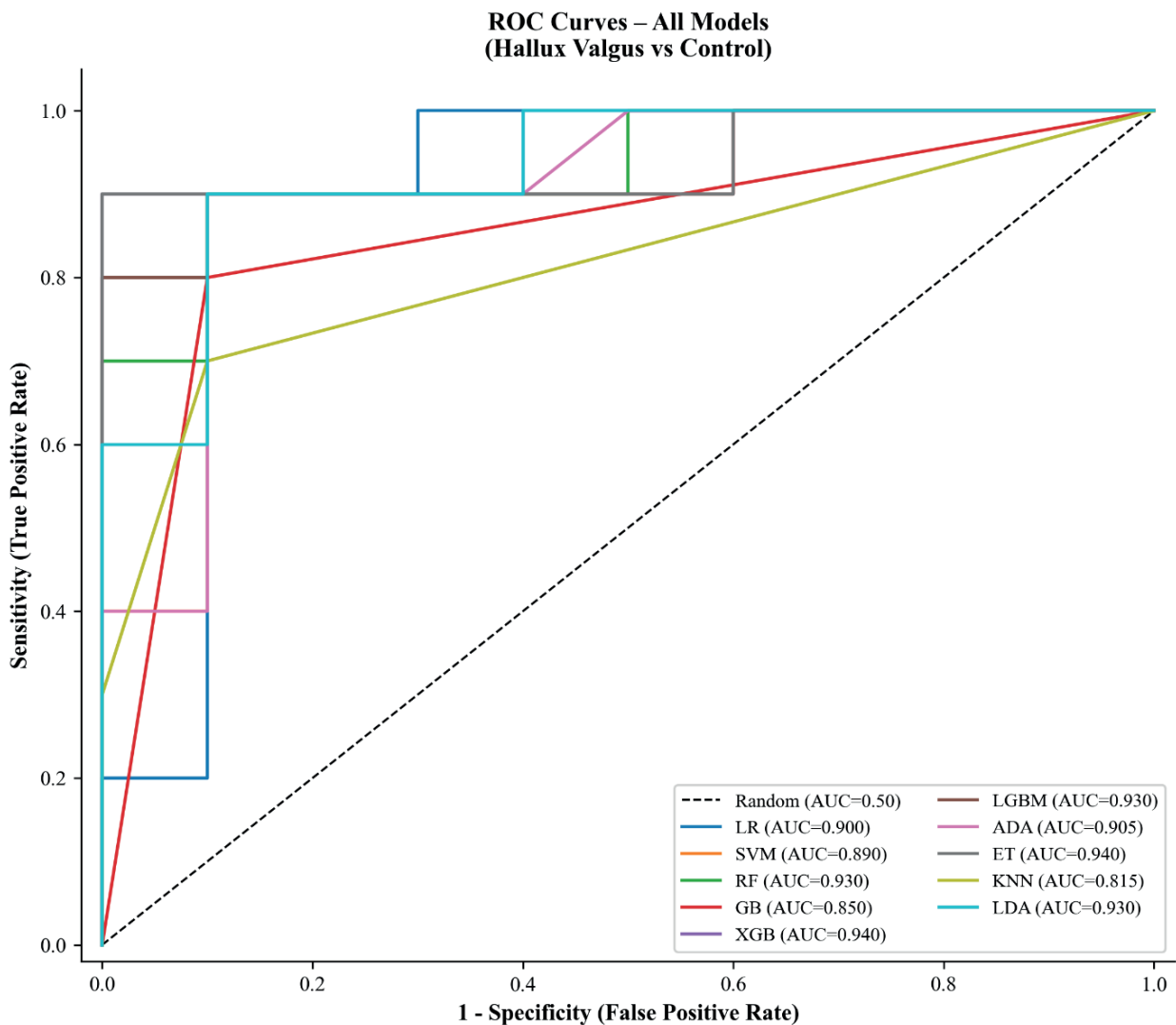


Figure 5. The ROC curve graphic of all the models.

ADA: AdaBoost; ET: Extra Trees; GB: Gradient Boosting; KNN: K-Nearest Neighbor; LDA: Linear Discriminant Analysis; LGBM: LightGBM; LR: Logistic Regression; RF: Random Forest; SVM: Support Vector Machine; XGB: XGBoost.

The relationship between the actual and predicted labels for the HV and control (CN) classes was presented in Figure 6 using normalized confusion matrices for each algorithm. The test set comprised ten cases for each class ($n = 10$ HV; $n = 10$ CN; total $n = 20$), and the proportions in each cell were calculated relative to the number of cases in the corresponding class. Regarding recognition of the control class, all algorithms exhibited a uniform performance pattern: across all ten models, 90% ($n = 9$) of the actual control cases were correctly classified as CN, while the

remaining 10% ($n = 1$) were misclassified as HV (false positive). However, marked differences among the models were observed in recognizing the HV class. RF, XGB, LGBM, ET, and LDA algorithms correctly classified 90% ($n = 9$) of the actual HV cases as HV and yielded false negatives in only 10% ($n = 1$); the accuracy of these five models was 90.0%. The LR, SVM, GB, and ADA algorithms achieved a correct identification rate of 80% ($n = 8$) for the HV class; the remaining 20% ($n = 2$) were misclassified as controls (false negatives). The overall accuracy of these four models was 85.0%. The

KNN algorithm exhibited the highest false-negative rate in the HV class: only 70% ($n = 7$) of the actual HV cases were correctly classified, while 30% ($n = 3$) were misclassified as controls. Consequently, KNN's overall accuracy was 80.0%, the lowest among all models.

The learning curve, obtained to evaluate the generalization behavior of the XGB algorithm as a function of training-set size, is shown in Figure 7; it compares the training score with the CV score. The analysis was conducted in five consecutive stages, during which the training-set size was increased from approximately 48 to 80 cases. The 95% confidence interval for the CV score was shown as a shaded band around the curve. The training score remained constant at 1.000 across all training-set sizes examined, indicating that the model achieved a perfect fit on the training data. The CV score, by contrast, displayed a marked upward trend as training-set size increased. Beginning at 0.73 for the smallest training-set size ($n \sim 48$), the CV score increased to 0.82 when the number of training samples reached 56, and to 0.89 when it reached 64. Expanding the training set to 72 cases brought the CV score to its peak of 0.93, at which point the gap between training and validation

scores was also minimized. When the training set was further enlarged to 80 cases, the CV score declined slightly to 0.90, while the confidence interval widened. The difference between the training and CV scores narrowed markedly as training set size increased: an initial separation of approximately 0.27 units decreased to 0.11 at 64 cases and to 0.07 at 72 cases.

Figure 8 shows the relative contributions of the radiographic angular variables used by the XGB algorithm to distinguish the HV and CN groups, presented in descending order as feature importance scores. Eight radiographic angular parameters were evaluated in the analysis, and a normalized importance score was calculated for each parameter. The variable with the highest contribution to the classification decision of the model was TMTA, with an importance score of 0.586; this value alone accounted for more than half of the total relative contribution of the eight variables. IPA was second to TMTA, with an importance score of 0.126, and made a marked contribution to the model's prediction process. The LTCA ranked third with an importance score of 0.106 and was highlighted in red in the figure. Variables providing moderate contributions to the model's prediction were

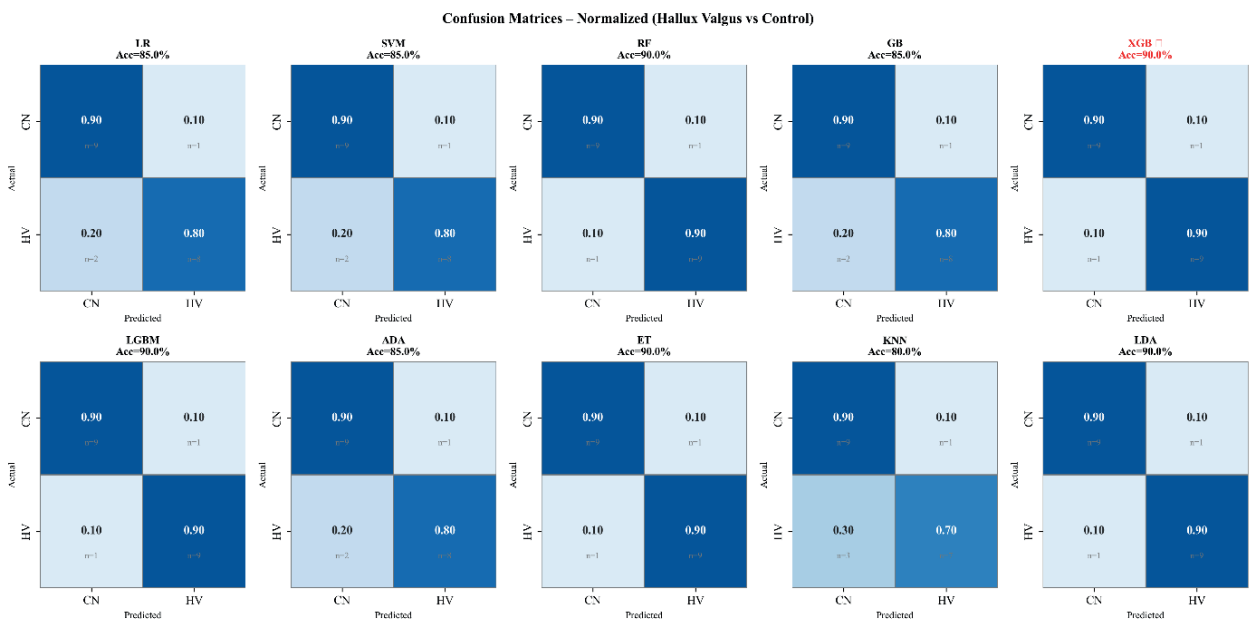


Figure 6. Confusion matrix tables of all the models.

ADA: AdaBoost; ET: Extra Trees; GB: Gradient Boosting; KNN: K-Nearest Neighbor; LDA: Linear Discriminant Analysis; LGBM: LightGBM; LR: Logistic Regression; RF: Random Forest; SVM: Support Vector Machine; XGB: XGBoost.

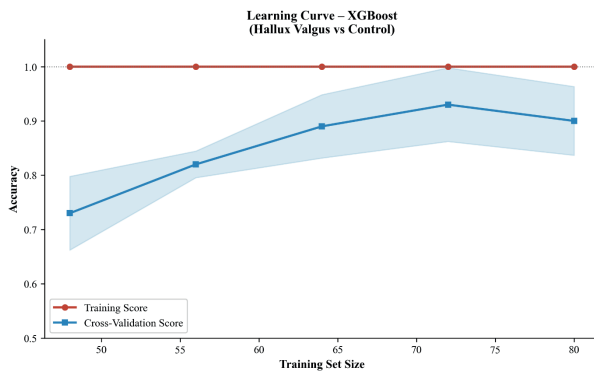


Figure 7. Training curve of the XGB algorithm.

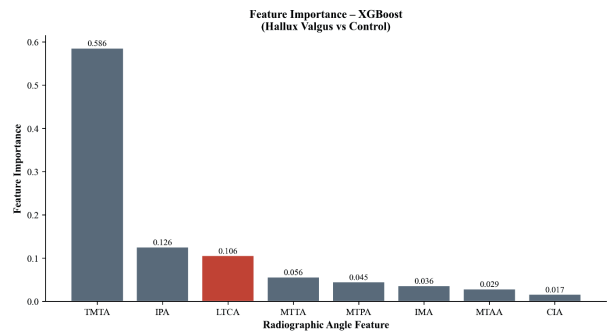


Figure 8. Feature importance ranking of the XGBoost algorithm.

MTTA (0.056) and MTPA (0.045). The parameters with relatively low contributions were IMA (0.036), MTAA (0.029), and CIA (0.017); these variables contributed least to the model's classification decision. Examination of the overall distribution showed that the cumulative importance of the top three variables (TMTA, IPA, LTCA) reached 0.818, indicating that the model's predictive power depended largely on these

three radiographic angles. By contrast, the cumulative contribution of the remaining five variables (MTTA, MTPA, IMA, MTAA, CIA) remained at 0.183.

The statistical results and descriptive values for all variables are shown in Table 4.

Table 4. Descriptive statistics and analytical results for the variables (n = 50 per group; CN vs HV).

Variable	CN			HV			Test Score	
	Mean $\pm$ SD / Median (min-max)	(Q1-Q3)	(95% CI Low-High)	Mean $\pm$ SD / Median (min-max)	(Q1-Q3)	(95% CI Low-High)	p-value	Sig.
IMA	21.71 $\pm$ 1.83	(20.50-23.07)	(21.19-22.23)	22.52 $\pm$ 3.32	(20.53-24.48)	(21.57-23.46)	0.1344	ns
IPA	36.48 (25.73-45.82)	(33.40-39.63)	(34.57-38.63)	45.41 (29.60-76.90)	(40.09-50.09)	(42.47-47.82)	<0.001	***
TMTA	7.60 (5.91-11.45)	(7.23-8.57)	(7.42-7.88)	12.54 (6.62-19.50)	(10.72-14.49)	(11.50-13.42)	<0.001	***
MTTA	7.93 (1.07-13.20)	(7.46-8.89)	(7.65-8.23)	7.13 (2.60-12.07)	(5.84-8.30)	(6.48-7.38)	0.0266	*
MTAA	17.86 $\pm$ 1.79	(16.72-19.01)	(17.35-18.37)	18.35 $\pm$ 2.28	(16.95-19.82)	(17.70-19.00)	0.2351	ns
MTPA	143.43 $\pm$ 9.20	(135.37-150.58)	(140.82-146.05)	143.25 $\pm$ 9.14	(137.44-147.63)	(140.66-145.85)	0.9230	ns
CIA	28.52 (22.31-37.65)	(26.97-30.33)	(27.54-29.31)	27.19 (10.85-47.46)	(25.21-29.23)	(25.82-28.40)	0.2632	ns
LTCA	50.96 $\pm$ 5.27	(47.30-54.27)	(49.46-52.46)	49.42 $\pm$ 6.98	(45.62-53.82)	(47.43-51.40)	0.2153	ns

Groups: CN = Control group; HV = Patient/Study group (n = 50 per group). Mean $\pm$ SD: mean $\pm$ standard deviation — measure of central tendency and dispersion of the data (parametric summary). Median (min-max): median value with the smallest and largest observed values — central tendency and range unaffected by extreme values. (Q1-Q3): interquartile range — boundaries of the central 50% of the data. (95% CI Low-High): 95% confidence interval for the mean — the interval within which the population mean lies with 95% probability. Test type: statistical approach selected according to the result of the normality test. Parametric $\rightarrow$ t-test (normal distribution); Non-parametric $\rightarrow$ Mann-Whitney U (non-normal distribution). All numerical values are rounded to two decimal places. p-value: statistical significance of the between-group difference based on one-way ANOVA. p < 0.05 was considered statistically significant. Significance codes: *** p < 0.001; ** p < 0.01; * p < 0.05; ns = not significant (p $\geq$ 0.05). Rows highlighted in red indicate statistically significant differences.

DISCUSSION

HV is a common foot deformity that significantly impairs activities of daily living, particularly among older adults. Studies employing artificial intelligence tools to diagnose or classify the disease by extracting reference points or morphometric measurements from foot radiographs, the conventional diagnostic method, are steadily increasing. Similarly, ten MLAs were used in this study to classify patients and controls based on angular measurements obtained from radiographic images. Overall, the study yielded accuracy rates ranging from 80% to 90%.

In a study aimed at automatically determining HV and IMA measurements from radiographic images, a deep neural network model was developed. By means of this model, the angles, consistent with the present study, were measured with an accuracy comparable to that of expert observers (12). Another study aimed at performing automatic IPA measurement on radiographs via bone segmentation employed a U-Net neural network. The AI-assisted measurements were found to coincide with the manual measurements, and the measurement results were consistent with the patient group of the present study (1). In a further study aimed at developing a deep learning model that automatically annotates reference points and computes the HV and IMA values, labels were placed on the radiographic images. The inter-rater agreement performance in HV and IMA measurements was found to be satisfactory, and the IMA values were comparable to those of the present study (19). In a separate study designed to measure HV angular values automatically using deep learning models and to compare these measurements with those made by surgeons, a UNet++-based DL model was developed. The HV and IMA measurements were found to be highly accurate, and the IMA values were consistent with those of the control and patient groups in the present study (20). Likewise, in another study presenting HV measurements from radiographic images via landmark detection with a deep convolutional neural network, 18 points were manually annotated. Although angular differences ranging from 3° to 5.08° were observed between predicted values and the reference standard, the IMA and IPA values were consistent with those of the patient group in this study (16).

Another paper aiming to compare deep learning and manual measurements of HV and IMA on radiographs developed an automated measurement system with low error rates. The IMA results were similar to those of the patient group in the present study (21). Yet a study that evaluated radiological findings using deep learning for HV diagnosis compared algorithmic performance with clinicians' manual performance. The findings which also formed the basis of patient control discrimination in the present study, demonstrated that the model accuracy for HV was sufficient (22). In further research designed to automatically classify the position of the hallux sesamoid on radiographs using a deep neural network, a kappa value above 0.6 was achieved. The developed model provided a high accuracy rate consistent with the present study and the related literature (23).

In a report aimed at developing predictive models for HV recurrence, patients who had completed treatment were classified as recurrent or non-recurrent based on one-year follow-up results. MLA models applied to pre- and postoperative radiographs achieved an accuracy of 75.6%. The angular measurement results, consistent with those of the present study, enable the identification of patients at risk of HV recurrence (5). An MLA study evaluating the relationship between structural deformity and foot pain in HV patients using radiographic findings found higher angular values in the painful group. According to the findings of the present study, the MTTA, MTAA, MTPA, and IPA variables, which are essential for HV diagnosis, were also reported to perform well in classifying painful versus non-painful feet (24). In a study designed to identify the presence of HV using MLAs applied to photographic images of participants' feet, two schemes were developed and evaluated with a convolutional neural network model. Consistent with the present study, the study yielded accuracy rates ranging from 62% to 96% (4). Finally, a subsequent study that evaluated the effects of age, sex, body mass index, and foot arch height on HV using a decision tree approach to construct a predictive model reported that these parameters influenced the presence of HV. According to the model's results, being over 50 years of age and female sex were associated with HV-related angles. In the present study, the same risk group can likewise be said to be more prevalent in the patient group (25).

CONCLUSION

Radiographic images, accepted as the standard method for diagnosing and monitoring foot deformities, are widely used. The present study is consistent with research employing artificial intelligence tools for HV diagnosis and classification. We believe that the study's findings will contribute to databases in this field and will guide developments in forefoot surgery and improve treatment options.

Ethical approval

This study has been approved by the Bolu Abant İzzet Baysal University Non-Interventional Clinical Research Ethics Committee (approval date 05/05/2026, number 2026/217).

Author contribution

Concept: MKu, AR; Design: MKu, GR; Data Collection or Processing: AR, MKa; Analysis or Interpretation: AR, İK; Literature Search: MKu; Writing: MKu, GR. All authors reviewed the results and approved the final version of the article.

Source of funding

The authors declare the study received no funding.

Conflict of interest

The authors declare that there is no conflict of interest.

Use of artificial intelligence

The authors declare that artificial intelligence was used in this study solely during the data analysis phase. Specifically, machine learning algorithms were employed to classify and evaluate the angular measurements derived from foot X-ray images for the diagnosis of hallux valgus. No generative artificial intelligence (generative AI) tools were used in the design of the study, the writing or editing of the manuscript, the generation or selection of references, or the interpretation of the results. The authors have reviewed all outputs and assume full responsibility for the entire content and findings of the manuscript.

REFERENCES

1. Kwolek K, Gądek A, Kwolek K, Lechowska-Liszka A, Malczak M, Liszka H. Artificial intelligence-based diagnosis of hallux valgus interphalangeus using anteroposterior foot radiographs. *World J Orthop.* 2025; 16(6): 103832. [\[Crossref\]](#)
2. Açıkgoz AK, Mutluay ŞD, Binokay F, Bozkir MG. Evaluation of the relationship between hallux valgus and foot measurements in radiographic images of adult female. *Surg Radiol Anat.* 2022; 44(9): 1281-8. [\[Crossref\]](#)
3. Nebhani N, Yatin Trivedy M, Heyes G, Molloy A, Mason L. Hallux valgus interphalangeus measurement: comparison of the 2 radiographic methods. *Foot Ankle Spec.* 2024; 17(4): 323-8. [\[Crossref\]](#)
4. Hida M, Eto S, Wada C, et al. Development of hallux valgus classification using digital foot images with machine learning. *Life (Basel).* 2023; 13(5): 1146. [\[Crossref\]](#)
5. Zhao R, Wang G, Li F, et al. Developing machine learning-based predictive models for hallux valgus recurrence based on measurements from radiographs. *Foot Ankle Int.* 2024; 45(9): 1000-8. [\[Crossref\]](#)
6. Kwolek K, Gądek A, Kwolek K, Kolecki R, Liszka H. Automated decision support for hallux valgus treatment options using anteroposterior foot radiographs. *World J Orthop.* 2023; 14(11): 800-12. [\[Crossref\]](#)
7. Forin Valvecchi T, Marcolli D, De Cesar Netto C. Advanced three-dimensional assessment and planning for hallux valgus. *Foot Ankle Clin.* 2025; 30(2): 349-62. [\[Crossref\]](#)
8. Spindler FT, Ettinger S, D. A. F. Scientific committee, Baumbach SF. Classification of hallux valgus deformity-is there a standard? *Arch Orthop Trauma Surg.* 2024; 144(11): 4737-43. [\[Crossref\]](#)
9. Chen YS, Liang CH, Shih HT, Tu KC, Tang SC, Wang SP. Correlation between hallux valgus severity and the prevalence of metatarsus adductus in hallux valgus. *J Foot Ankle Res.* 2025; 18(2): e70049. [\[Crossref\]](#)
10. Li R, Tan K, Xie Y, Wang F. Morphology variations with medial cuneiform in hallux valgus. *Anat Sci Int.* 2024; 99(1): 59-67. [\[Crossref\]](#)
11. Bu P, Li C, Pu L, Ma X, Meng X, Xu Y. Radiographic assessment of relationship between medial cuneiform obliquity and hallux valgus. *J Foot Ankle Surg.* 2023; 62(3): 583-9. [\[Crossref\]](#)
12. Takeda R, Mizuhara H, Uchio A, et al. Automatic estimation of hallux valgus angle using deep neural network with axis-based annotation. *Skeletal Radiol.* 2024; 53(11): 2357-66. [\[Crossref\]](#)
13. Matsumoto T, Higuchi J, Maenohara Y, et al. The discrepancy between radiographically-assessed and self-recognized hallux valgus in a large population-based cohort. *BMC Musculoskelet Disord.* 2022; 23(1): 31. [\[Crossref\]](#)

14. Kwolek K, Liszka H, Kwolek B, Gądek A. Measuring the angle of hallux valgus using segmentation of bones on X-ray images. In: Tetko I, Kůrková V, Karpov P, Theis F, editors. *Artificial Neural Networks and Machine Learning – ICANN 2019: Workshop and Special Sessions. ICANN 2019. Lecture Notes in Computer Science, Vol 11731*. Cham: Springer; 2019: 327-38. [\[Crossref\]](#)
15. Izquierdo P, Calderon A, Jenkins GL, Thorne S, Mathieson I. Automatic angle recognition in hallux valgus. In: *Proceedings of the UKSim-AMSS 22nd International Conference on Computer Modelling and Simulation (UKSim2020)*; 2020; Cambridge, United Kingdom. [\[Crossref\]](#)
16. Li T, Wang Y, Qu Y, Dong R, Kang M, Zhao J. Feasibility study of hallux valgus measurement with a deep convolutional neural network based on landmark detection. *Skeletal Radiol*. 2022; 51(6): 1235-47. [\[Crossref\]](#)
17. Ray A, Ray G, Kürtül İ, Şenol GT. Sex estimation from the variables of talocrural joint by using machine learning algorithms. *J Forensic Leg Med*. 2025; 114: 102912. [\[Crossref\]](#)
18. Taşkın Şenol G, Kürtül İ, Ray A, Ahmetoğlu G, Seçgin Y, Öner Z. Determination of gender by machine learning algorithms, through using craniocervical junction parameters and dimensions of the cervical spinal canal. *Osmangazi Journal of Medicine*. 2023; 45(5): 672-7. [\[Crossref\]](#)
19. Ma S, Wang H, Zhao W, et al. An interpretable deep learning model for hallux valgus prediction. *Comput Biol Med*. 2025; 185: 109468. [\[Crossref\]](#)
20. Wang Q, Wang JH, Ji DD, Zhang Y, Zhao HM. Automated measurement of hallux valgus angles using a deep learning model: validation and comparison with surgeons of varying expertise. *Sci Rep*. 2026; 16(1): 12280. [\[Crossref\]](#)
21. Wang Q, Wang JH, Ji DD, Liang JQ, Yang XQ, Zhao HM. Deep learning versus manual measurement of hallux valgus angle and intermetatarsal angle on weight-bearing X-rays in hallux valgus. *BMC Musculoskelet Disord*. 2025; 27(1): 58. [\[Crossref\]](#)
22. Hussain A, Lee C, Hu E, Amirouche F. Deep learning automation of radiographic patterns for hallux valgus diagnosis. *World J Orthop*. 2024; 15(2): 105-9. [\[Crossref\]](#)
23. Takeda R, Uchio A, Iidaka T, et al. Automatic Hardy and Clapham's classification of hallux sesamoid position on foot radiographs using deep neural network. *Foot Ankle Surg*. 2025; 31(3): 220-6. [\[Crossref\]](#)
24. Wang C, Li C, Zhang R, et al. Identification of radiographic characteristics associated with pain in hallux valgus patients: a preliminary machine learning study. *Front Public Health*. 2022; 10: 943026. [\[Crossref\]](#)
25. Jung DY, Jung SH, Gwak GT. Contributions of age, gender, body mass index, and normalized arch height to hallux valgus: a decision tree approach. *BMC Musculoskelet Disord*. 2023; 24(1): 278. [\[Crossref\]](#)