

Article

A Comparative Study of Machine Learning Classifiers for Enhancing Knee Osteoarthritis Diagnosis

Aquib Raza ^{1,†}, Thien-Luan Phan ^{1,2,†} , Hung-Chung Li ³ , Nguyen Van Hieu ², Tran Trung Nghia ^{4,*} 
and Congo Tak Shing Ching ^{1,5,6,*} 

¹ Graduate Institute of Biomedical Engineering, National Chung Hsing University, Taichung 402, Taiwan; g111068033@mail.nchu.edu.tw (A.R.); ptluan@hcmus.edu.vn (T.-L.P.)

² Department of Physics and Electronic Engineering, University of Science, Vietnam National University of Ho Chi Minh City, Ho Chi Minh City 70000, Vietnam; nvhieu@hcmus.edu.vn

³ Undergraduate Program of Intellectual Creativity Engineering, National Chung Hsing University, Taichung 402, Taiwan; hcli01@nchu.edu.tw

⁴ Laboratory of Laser Technology, Faculty of Applied Science, Ho Chi Minh City University of Technology (HCMUT), VNU-HCM, Ho Chi Minh City 72409, Vietnam

⁵ Department of Electrical Engineering, National Chi Nan University, Puli Township 545, Taiwan

⁶ International Doctoral Program in Agriculture, National Chung Hsing University, Taichung 402, Taiwan

* Correspondence: ttngchia@hcmut.edu.vn (T.T.N.); tsching@nchu.edu.tw (C.T.S.C.)

† These authors contributed equally to this work.

Abstract: Knee osteoarthritis (KOA) is a leading cause of disability, particularly affecting older adults due to the deterioration of articular cartilage within the knee joint. This condition is characterized by pain, stiffness, and impaired movement, posing a significant challenge in medical diagnostics and treatment planning, especially due to the current inability for early and accurate detection or monitoring of disease progression. This research introduces a multifaceted approach employing feature extraction and machine learning (ML) to improve the accuracy of diagnosing and classifying KOA stages from radiographic images. Utilizing a dataset of 3154 knee X-ray images, this study implemented feature extraction methods such as Histogram of Oriented Gradients (HOG) with Linear Discriminant Analysis (LDA) and Min–Max scaling to prepare the data for classification. The study evaluates six ML classifiers—K Nearest Neighbors classifier, Support Vector Machine (SVM), Gaussian Naive Bayes, Decision Tree, Random Forest, and XGBoost—optimized via GridSearchCV for hyperparameter tuning within a 10-fold Stratified K-Fold cross-validation framework. An ensemble model has also been made for the already high-accuracy models to explore the possibility of enhancing the accuracy and reducing the risk of overfitting. The XGBoost classifier and the ensemble model emerged as the most efficient for multiclass classification, with an accuracy of 98.90%, distinguishing between healthy and unhealthy knees. These results underscore the potential of integrating advanced ML methodologies for the nuanced and accurate diagnosis and classification of KOA, offering new avenues for clinical application and future research in medical imaging diagnostics.

Keywords: knee osteoarthritis; diagnosis; X-ray; machine learning



Citation: Raza, A.; Phan, T.-L.; Li, H.-C.; Hieu, N.V.; Nghia, T.T.; Ching, C.T.S. A Comparative Study of Machine Learning Classifiers for Enhancing Knee Osteoarthritis Diagnosis. *Information* **2024**, *15*, 183. <https://doi.org/10.3390/info15040183>

Academic Editor: Saurav Mallik

Received: 21 February 2024

Revised: 24 March 2024

Accepted: 25 March 2024

Published: 28 March 2024



Copyright: © 2024 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

Knee osteoarthritis (KOA), a prevalent joint disorder particularly affecting the tibiofemoral joint between the femur and tibia, is characterized by the degeneration of articular cartilage. Common in older adults, KOA induces increased joint friction and cartilage volume loss, leading to symptoms such as pain, stiffness, and restricted movement. In the United States, it is the most common joint condition, impacting 13% of women and 10% of men over 60 [1]. Knee joint functionality relies on the cartilage that cushions the femur and tibia bone, which diminishes with age or injury, and will affect knee mobility. Chondrocytes within this cartilage are crucial for load distribution, supporting the joint throughout an individual's life [2].

Osteoarthritis (OA) stands as a significant musculoskeletal disorder globally, with knee OA being one of the top fifty diseases worldwide, affecting approximately 250,000 people [1]. The Kellgren–Lawrence (KL) [3] grading system, recognized by the World Health Organization (WHO) since 1961, is widely used to assess KOA severity via radiographs, favored for their cost-effectiveness despite advances in imaging technologies [4]. Current OA management lacks curative treatments besides knee replacement [5], but behavioral interventions such as weight loss, exercise, and endurance training offer temporary symptom relief and may slow OA progression [6]. Globally, knee OA's prevalence is notable, with 10% of those over 15 and 22.9% over 40 affected. Its incidence rate is 203 per 10,000 person-years for those aged over 20. By 2020, it is estimated that 86.7 million people worldwide aged 20 and older will experience KOA [7], underscoring its significant health burden. Furthermore, according to findings from Global Burden of Disease 2010 [8], among the 291 disorders studied, knee OA was classified as the 11th leading cause of disability worldwide.

Computer-Assisted Detection (CAD) techniques are instrumental in diagnosing knee osteoarthritis, employing various methods such as radiography, Magnetic Resonance Imaging (MRI), gait analysis, and bioelectric impedance signals [9]. Radiography, particularly X-rays, is adept at highlighting osteophytes and joint space narrowing, indicative of KOA. MRI, on the other hand, excels in quantifying cartilage attributes including thickness, surface area, and texture, providing a more detailed assessment of the joint's condition. KOA is currently diagnosed by patient-reported symptoms in combination with X-ray evaluations. While radiography is celebrated for its accessibility, cost-effectiveness, and rapid results, it has notable limitations, particularly in detecting early-stage disease and subtle changes over time. Despite the drawbacks, it remains the gold standard for KOA diagnosis, offering valuable insights into alterations in human bones, such as cyst development, joint space decline, and subchondral bone sclerosis [10].

Recent developments in machine learning (ML) and deep learning (DL) algorithms have demonstrated significant potential for automatically extracting and analyzing features from clinical images for accurate disease detection. ML and DL excel in identifying intricate patterns in data, offering superior accuracy and earlier disease detection, with applications in semantic segmentation [11], medical imaging [12,13], monitoring ecosystem changes [14], and even weather forecasting [15].

This evolution in medical diagnostics promises improved patient care through timely interventions, highlighting the significant impact of ML and DL in enhancing diagnostic processes specifically. In the context of KOA detection [16–28], radiographic images have been successfully analyzed using a spectrum of ML and DL algorithms with promising results. For instance, L. Anifah's team tackled knee grading through image processing and self-organizing maps, observing varied accuracy across KL grades [16]. Another approach utilized semi-automated segmentation, focusing on Regions of Interest (ROIs) with Kellgren–Lawrence (KL) grades 0 and 2, to achieve 82.98% accuracy using Random Forest and Naïve Bayes algorithms [22]. R. Mahum et al. (2021) introduced a DL strategy for early KOA detection, employing hybrid feature models and multi-class classifiers to reach roughly 97% accuracy [24]. Similarly, T. Tariq (2023) developed a method combining pre-trained CNNs and transfer learning for KOA classification, achieving up to 90.8% accuracy in binary models [27]. These studies highlight the versatility and effectiveness of ML and DL in improving KOA diagnostic processes.

Moreover, the optimization of ML models through hyperparameter tuning has been identified as a critical step in achieving effective implementation [28]. This process involves adjusting various types of hyperparameters—discrete, categorical, or continuous—to enhance model efficacy [29]. GridSearch (GS), a longstanding method in machine learning, is widely used for hyperparameter optimization [30–35]. It searches through a predefined subset of the hyperparameter space, chosen manually. Known for its simplicity, ease of parallelization, and effectiveness in lower-dimensional spaces, GridSearch is preferred for its comprehensive approach to evaluating every possible combination of hyperparameter values, ensuring thorough exploration of model configurations [26].

This research introduces an innovative approach that combines machine learning (ML) with GridSearch optimization for classifying knee osteoarthritis (KOA) severity from knee X-ray images using Kellgren–Lawrence (KL) grades and optimized hyperparameters across six models using GridSearch to identify the best settings for accurate KOA detection, aiming to improve predictive performance while minimizing computational load.

The proposed method employs a Histogram of Oriented Gradients (HOG) and Linear Discriminant Analysis (LDA) in combination with Min–Max scaling for feature extraction, analyzing the effect of multiclass classifications on model performance. This approach is assessed through metrics like recall, accuracy, and precision.

Additionally, an ensemble model, combining all of the best tuning parameters of the six proposed models will also be explored. These efforts aim to advance diagnostic methods for KOA, underscoring the significant potential of ML in improving healthcare outcomes.

2. Materials and Methods

This section outlined the suggested process for KOA prediction and rating. Figure 1 depicts the flow diagram for the proposed technique. The KOA data undergo many preprocessing processes before feature extraction. After picking relevant features, they are fed into machine learning models. Finally, GridSearch optimizes parameters to improve model efficiency. Evaluation steps were made to ensure the performance of the models.

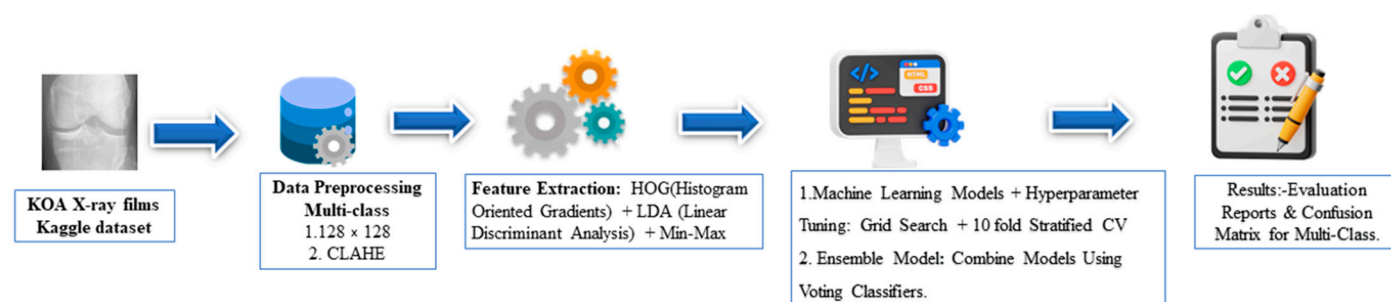


Figure 1. Flow diagram of the proposed model.

2.1. KOA X-ray Images Dataset

To carry out the experiments included in the suggested flow diagram, a publicly available dataset (Mendeley data) of 5778 X-ray images on knee osteoarthritis [36] was used. The data were distributed across various classes, including healthy, minimal, moderate, and severe. The distribution of cases within the KL scoring system for determining knee OA severity was at 2286, 1516, 757, and 173, respectively, for each class. The datasets were utilized given its comprehensive inclusion of knee X-ray images that are pivotal for both knee KL grading and joint identification. Notably, the original dataset also includes images under a fifth category, ‘doubtful’. However, this class was considered problematic for inclusion due to its ambiguous nature.

Given the disproportionately large number of healthy samples, this work aimed to balance the dataset by reducing the “healthy” class data, randomly selecting only 30% of the healthy images, equating to 708 images. The final number of images used in this study was 3154, of which 708 were healthy, 1516 were minimal, 757 were moderate, and 173 were severe. This adjustment was made to achieve a more uniform distribution of samples across the classes, ensuring a fair comparison, allowing the proposed model to learn the features of the unhealthy classes more effectively, improving the overall accuracy across all classes, and enhancing the validity of this work’s findings.

2.2. Data Pre-Processing

A key step in preparing the data for further processing in the proposed system involved preprocessing the images. This preprocessing aimed at not only converting the

format of the images but also enhancing the image quality to ensure optimal readiness for the subsequent analyses.

In the data preparation phase of the study, the focus was on readying the images for further analysis through two primary steps. The images were first scaled to a standard resolution of 128×128 pixels. This uniform resizing of input data is crucial as smaller images facilitate faster learning by machine learning algorithms.

Following this, the Contrast-Limited Adaptive Histogram Equalization (CLAHE) technique was employed to enhance the clarity of the images. CLAHE improves image contrast by adjusting brightness levels and ensuring optimal contrast distribution. It utilizes two key settings: the grid size, which determines how adjustments are distributed across the image, and the clip limit, which controls the extent of contrast enhancement. Optimal results were achieved with a clip limit set to 2.0 and a grid size of (8, 8), making the images more uniform and easier to analyze.

Enhancing image quality through CLAHE was identified as crucial, ensuring that the images were in the best possible condition for analysis by machine learning algorithms. This step underscores the importance of thorough data preparation in enabling accurate and efficient analysis in studies involving medical imaging.

2.3. Feature Extraction

Histogram of Oriented Gradients (HOG) is a popular feature classifier in image analysis and object identification [37,38]. It transforms object characteristics into numerical values by calculating the orientations of gradients within 8×8 -pixel cells. The range of gradient angles was divided into 8 bins, with cells grouped into 2×2 blocks for contrast normalization, effectively preserving texture and shape features for analysis and classification.

Linear Discriminant Analysis (LDA) was then used to reduce the dimensionality of the HOG features to three components ($n_components = 3$), and the utilization of three components was found to be best for computational efficiency while ensuring a balanced model complexity and predictive performance, which is important for real-time applications. LDA aims to enhance class separability and minimize variation within classes, leading to a lower-dimensional data representation that still contains essential discriminative information.

To ensure uniformity across features, Min–Max scaling was applied to the LDA-transformed features, adjusting them to a specific range, typically between 0 and 1. This scaling is crucial for many machine learning algorithms, preventing features with larger magnitudes from dominating those with smaller ones. It standardizes the data and facilitates faster convergence of optimization techniques during model training. Min–Max scaling is defined by the formula:

$$X_{\text{scaled}} = \frac{X - X_{\min}}{X_{\max} - X_{\min}} \quad (1)$$

where X indicates a feature value transformed by LDA. $X_{\max}$ and $X_{\min}$ are the maximum and minimum values for the feature across all samples in the dataset, respectively. X_{scaled} defined by Equation (1) is the scaled value of the feature, ensuring all features are in the same range, downscaling the calculation.

2.4. Model Selection and Hyperparameter Tuning

In the study, six well-established machine learning algorithms were utilized to predict knee osteoarthritis (KOA) outcomes. The chosen models were K Nearest Neighbors (KNN), Support Vector Machine (SVM), Gaussian NaiveBayes (GaussianNB), Decision Tree (DT), Random Forest (RF), and Extreme Gradient Boosting (XGBoost). These models were selected for their diverse functionalities, hyperparameters, and proven records in KOA classification. The tuning process plays a crucial role in differentiating the performance of these six models, influencing their comparative evaluation.

Hyperparameter tuning was implemented to enhance the models' efficiency, aiming to mitigate overfitting and bias. For each model, a set of hyperparameters was preselected for optimization, ranging from the basic necessary ones to the more optimized value found in the literature [30–35]. The training and evaluation of these models were conducted using Python 3.11.4, employing specific frameworks tailored to machine learning tasks. This approach ensured that each model was fine-tuned to its optimal settings, allowing for a fair and effective comparison across the different algorithms in predicting KOA outcomes.

GridSearch (GS) is the leading method for exploring hyperparameter configurations, conducting an exhaustive search across a grid of hyperparameter combinations [39]. This approach evaluates predefined hyperparameter values using the Cartesian product method, allowing for thorough parameter space exploration [40].

The proposed method employs GridSearch cross-validation with Stratified K-Fold for 10-fold cross-validation, systematically exploring each model's hyperparameter space. GridSearch cross-validation methodically tests different hyperparameter combinations to determine the best ones from the preset tuning value (Table 1) for a model. Stratified K-Fold is a variation of k-fold cross-validation where each fold is made by preserving the percentage of samples for each class, and is particularly useful when dealing with an imbalanced dataset, i.e., when the number of samples in each class varies widely. In total, 90% of each class was preserved for training while 10% was used for testing. This ensures that each class is adequately represented in each fold, leading to more reliable and unbiased evaluation metrics for the model.

Table 1. Hyperparameter tuning chosen for each classifier.

Classifier	Hyperparameters	Tuning Value
K Nearest Neighbors classifier (KNN)	n_neighbors	{3, 5, 7, 9, 11, 13}
Support Vector Machine (SVM)	C	{1, 10, 50, 70, 100}
	gamma	{'scale', 'auto'}
	kernel	{'linear', 'poly', 'rbf', 'sigmoid'}
	class_weight	{None, 'balanced'}
GaussianNB	var_smoothing	{ 10^{-11} , 10^{-10} , 10^{-9} , 10^{-8} }
Decision Tree classifier	max_depth	{None, 3, 5, 7, 10}
Random Forest classifier	n_estimators	{50, 100, 200, 300, 500}
	warm_start	{True, False}
	n_jobs	{None, 1, 2, 3}
XGBoost classifier	max_depth	{3, 5, 7}
	learning_rate	{0.1, 0.01, 0.001}
	n_estimators	{50, 100, 200}
	gamma	{0, 1, 2}
	subsample	{0.8, 0.9, 1.0}

2.5. Performance Evaluations

Analyzing the performance of machine learning models required a thorough and diversified approach to determine how effectively these classifiers operate. This work outlines a detailed strategy for evaluating classifiers, integrating both statistical data and visual aids into the analysis. At the core of this approach are several fundamental strategies including cross-validation to assess a model's generalizability to unseen data, the use of confusion matrices for visualizing algorithm performance, and the application of precision-recall metrics alongside Receiver Operating Characteristic (ROC) curves analysis and Area Under the Curve (AUC) score computation.

The evaluation strategy begins with cross-validation accuracy, where the dataset is segmented into subsets or "folds" for training and validation in a cycle that ensures each fold acts as the validation set once. This method provides an average accuracy that reflects

the model's overall performance, aiding in overfitting prevention and offering a solid performance metric.

The variance in cross-validation scores is then assessed to determine the model's responsiveness to the specific subsets of training data. A high variance could indicate overfitting, whereas a lower variance suggests a model's stability and its ability to generalize better to new data.

The confusion matrix was applied in the evaluation step in order to enhance the evaluation of the model's performance, presenting true positives, true negatives, false positives, and false negatives. This approach helps identify biases or specific error types that straightforward accuracy metrics may not detect.

The classification report goes further, providing essential metrics like precision score defined by Equation (2), recall score (Equation (3)), F1 score (Equation (4)), and accuracy score (Equation (5)) support for each class, offering a deeper insight into the model's performance across different categories, an aspect especially critical in datasets with imbalanced classes. The report score is depicted as follows:

$$\text{Precision score} = \frac{TP}{TP + FP} \quad (2)$$

$$\text{Recall score} = \frac{TP}{TP + FN} \quad (3)$$

$$\text{F1 score} = 2 \times \frac{\text{Precision score} \times \text{Recall score}}{\text{Precision score} + \text{Recall score}} \quad (4)$$

$$\text{Accuracy score} = \frac{TP + TN}{TP + TN + FP + FN} \quad (5)$$

Finally, the evaluation incorporates the ROC curve and AUC score. The ROC curve plots the true positive rate against the false positive rate at various thresholds, facilitating the selection of the most effective model. The AUC score, quantifying the area beneath the entire ROC curve, serves as an overall performance measure across all possible classification thresholds, with higher scores denoting a model's excellence.

Together, these strategies form a robust framework for understanding and evaluating the operational effectiveness of machine learning models, ensuring a comprehensive view of their predictive capabilities.

2.6. Ensemble Model

Additionally, the implementation of an ensemble model using Scikit-learn's voting classifier was made to enhance predictive performance on the dataset. This ensemble combines outputs from various optimally tuned machine learning models—K Nearest Neighbors, SVM, GaussianNB, Decision Tree classifier, Random Forest classifier, and XGBoost classifier—using a 'soft' voting mechanism. This method averages the probabilities from all models for predictions, offering a nuanced decision-making process that leverages the strengths of each model.

To evaluate the ensemble's effectiveness, cross-validation predictions for the entire dataset are generated using `cross_val_predict` with a Stratified K-Fold strategy, ensuring balanced sampling. The accuracy of these predictions is assessed with the `accuracy_score` function, providing a clear metric for overall performance. Additionally, a classification report details precision, recall, and F1 scores for each class, offering insights into the model's capabilities across different categories.

A confusion matrix is computed and visualized with `ConfusionMatrixDisplay` to show the ensemble model's accuracy and misclassifications between classes. This serves as a diagnostic tool for understanding where the ensemble excels or falls short, guiding improvements to the modeling approach. This ensemble method represents a strategic effort to utilize the collective power of multiple models for superior predictive performance.

2.7. Computational Settings

The development and evaluation of the machine learning models for classifying the stages of knee osteoarthritis took place within a Jupyter Notebook environment, leveraging Python 3.11.4. These computational experiments utilized a high-performance setup featuring an Intel® Core™ i7-10700 CPU at 2.90 GHz, complemented by 32 GB of RAM. This configuration ensured the efficient management of the dataset and computational demands. Graphical processing tasks, including model training and image processing, were handled by an Nvidia GeForce RTX 3070 video card. Such a hardware arrangement provided ample computational power for extensive model training, hyperparameter optimization, and validation efforts.

3. Results and Discussions

This section elaborates on the results achieved using the proposed method for detecting knee osteoarthritis (KOA).

3.1. Feature Extraction Result

Figure 2 illustrates the results of feature extraction following the preprocessing of images using Contrast-Limited Adaptive Histogram Equalization (CLAHE) followed by feature extraction techniques HOG and LDA. The scatter plot shows four distinct clusters corresponding to the different stages of knee osteoarthritis: healthy (red), minimal (green), moderate (blue), and severe (yellow). The axes, 'LDA Component 1', 'LDA Component 2', and 'LDA Component 3' represent the three most discriminative features extracted by LDA that best separate the classes. The extracted feature is then rescaled using Min-Max to ensure uniformity across features. Each point on the plot likely represents an X-ray image, transformed into this two-dimensional space.

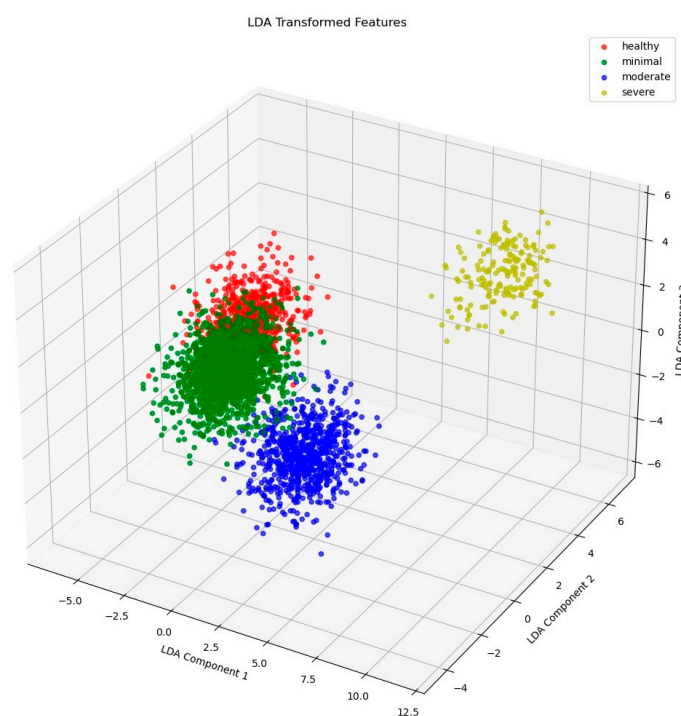


Figure 2. HOG and LDA transformed feature.

The scatter plot suggests that the Histogram of Oriented Gradients (HOG) feature extraction, followed by Linear Discriminant Analysis (LDA), has effectively differentiated the stages of knee osteoarthritis. The stark separation between the 'severe' class and other categories suggests that the machine learning model can easily distinguish the advanced stage of the disease. However, the noticeable overlap among the 'healthy', 'minimal', and 'moderate' classes suggests that the model may have difficulty distinguishing between these early stages of the disease.

and ‘moderate’ stages hints at possible challenges in classification. This overlap may require a more nuanced approach in model training and could benefit from more complex classification algorithms or additional features that capture the subtleties between these intermediate stages.

3.2. Confusion Matrix and Classification Report of the Cross-Validation

The 10-fold cross-validation was performed with the data set, using 90% of the available data as training data, and 10% as testing data in combination with GridSearch to find the optimized tuning parameter. Figure 3 displays the confusion matrix for six distinct classifiers, revealing varying degrees of accuracy and performance nuances.

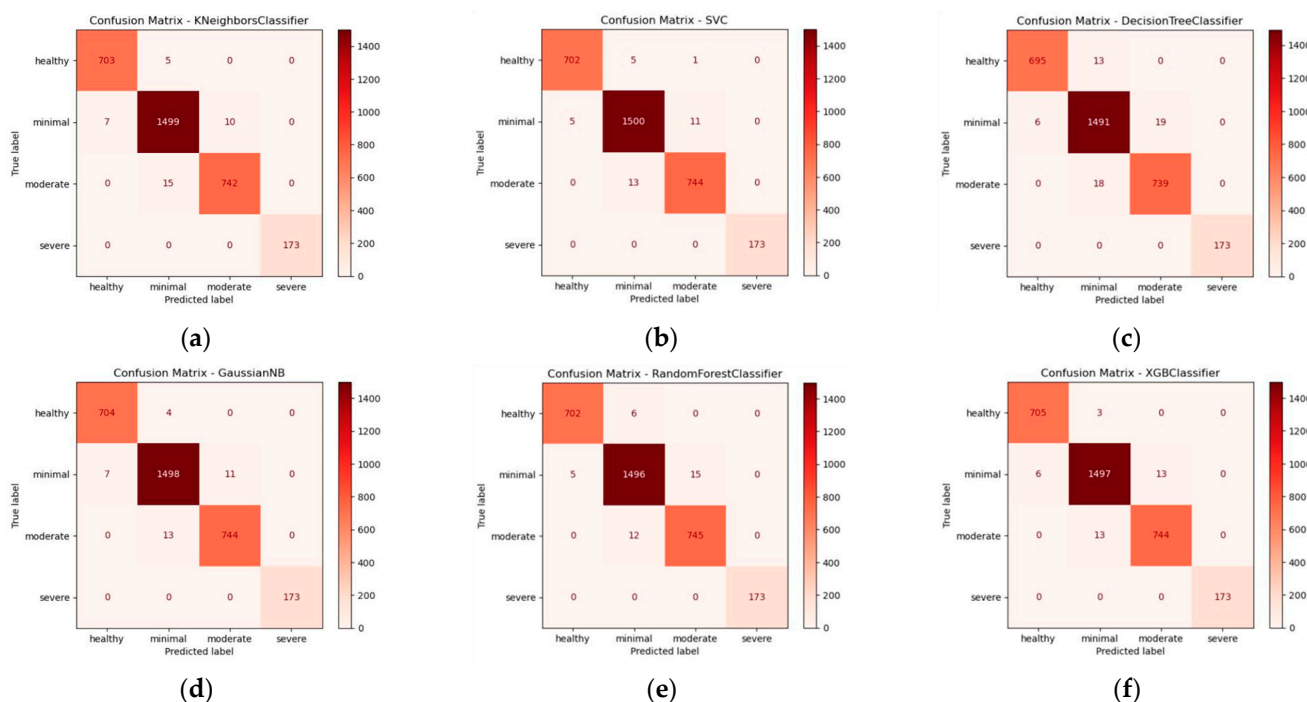


Figure 3. Confusion matrix of (a) KNN model, (b) SVM classifier, (c) Decision Tree classifier, (d) GaussianNB, (e) Random Forest classifier, and (f) XGBoost classifier for the classification of 4 categories of KOA.

KNN classifier: Achieved an impressive accuracy of approximately 98.83% with low variance, showcasing consistent performance across cross-validation folds. This suggests robustness and efficacy, likely due to its proficiency in capturing the local structure of the image feature space. GridSearch tuning revealed 3 neighbors as the optimal number, highlighting its importance in model generalization.

SVM model: Slightly outperformed the KNN classifier, recording an accuracy of about 98.89% with low variance, indicative of excellent generalizability. The use of a linear kernel with $C = 100$ suggests an optimal balance between decision boundary clarity and misclassification penalty, particularly after applying HOG and LDA techniques.

Decision Tree classifier: Achieved an accuracy of about 98.19%, with a slightly higher variance, hinting at potential overfitting. An optimal max depth of 7 was determined, indicating a balance between model complexity and the ability to distinguish between classes.

Gaussian Naive Bayes classifier: Matched the SVM model’s accuracy at 98.89% with a slightly higher variance, performing well possibly due to its effectiveness in handling uneven datasets. An optimized var_smoothing parameter at 10^{-9} plays a crucial role in ensuring numerical stability and variance estimation.

Random Forest classifier: Showcased an accuracy of roughly 98.67%, underscoring the value of ensemble methods in improving generalization. Employing 200 trees with

warm_start enabled, this method effectively reduces overfitting risk while enhancing class-wide performance.

XGBoost classifier: Emerged as the top performer with the highest accuracy of approximately 98.90% and shows the lowest variance among the models. Its success is attributed to the use of advanced regularization and gradient boosting mechanisms, with optimal settings that include a learning rate of 0.1, max depth of 5, 100 boosting rounds, and a subsampling rate of 1.0.

The classification report (Table 2) and confusion matrix (Figure 3) across all classifiers show high precision, recall, and F1 scores, demonstrating their capability to differentiate between the stages of knee osteoarthritis effectively. Table 2 also sums up the optimized hyperparameter value of each classifier which provides the highest model performance. Despite minor misclassifications, primarily between adjacent categories, these results highlight the classifiers' overall proficiency and the challenge of distinguishing inherently similar stages.

Table 2. Classification reports of the 6 classifiers for 4 categories of KOA.

Model	Optimized Hyperparameter Value	Healthy (Precision, Recall, F1-Score)	Minimal (Precision, Recall, F1-Score)	Moderate (Precision, Recall, F1-Score)	Severe (Precision, Recall, F1-Score)	Cross Validation Accuracy (%)
K Nearest Neighbors (KNN)	n_neighbors: 3	0.99, 0.99, 0.99	0.99, 0.99, 0.99	0.99, 0.98, 0.98	1.00, 1.00, 1.00	98.83
SVM	C: 100, kernel: rbf	0.99, 0.99, 0.99	0.99, 0.99, 0.99	0.98, 0.98, 0.98	1.00, 1.00, 1.00	98.89
GaussianNB	var_smoothing: 10^{-9}	0.99, 0.99, 0.99	0.99, 0.99, 0.99	0.99, 0.98, 0.98	1.00, 1.00, 1.00	98.89
Decision Tree	max_depth: 7	0.99, 0.98, 0.99	0.98, 0.98, 0.98	0.97, 0.98, 0.98	1.00, 1.00, 1.00	98.19
Random Forest	n_estimators: 200, warm_start: false	0.99, 0.99, 0.99	0.99, 0.99, 0.99	0.98, 0.98, 0.98	1.00, 1.00, 1.00	98.67
XGBoost	gamma: 1, learning_rate: 0.1, max_depth: 5, n_estimators: 100, subsample: 1.0	0.99, 1.00, 0.99	0.99, 0.99, 0.99	0.98, 0.98, 0.98	1.00, 1.00, 1.00	98.90

3.3. Learning Curve, ROC Curve, and AUC Score Report of the Multi-Model Classifier

Learning curves were utilized for model evaluations. The analysis of learning curves for the classifiers reveals distinct insights into their performance and generalization capabilities. KNN learning curves (Figure 4a) show high and consistent training and cross-validation scores, indicating that the model is performing well with no signs of overfitting or underfitting. SVM and GaussianNB ((Figure 4b,d) learning curves show high training and cross-validation scores, suggesting a stable model and excellent generalization. The learning curve of the Decision Tree classifier (Figure 4c) shows that high scores are maintained across training and validation, with a small gap, indicating good performance but potential for overfitting compared to others. A slight gap between the very high training and cross-validation scores can be seen from the Random Forest classifier and XGBoost classifier learning curves (Figure 4e,f), which also indicates that the models are well-fitted and generalize exceptionally well. The lack of disparity between the scores indicates that the models are neither overfitting nor underfitting, which is ideal for predictive tasks.

The learning curves indicate strong generalization for all classifiers, with high, consistent validation scores suggesting low bias and variance. The small score gaps hint at robust model training. Given their performance, these models seem ready for deployment, assuming validation was thorough and the data were truly unseen. The uniform success across various classifiers, from GaussianNB to Random Forest and XGBoost classifier, implies a straightforward task that does not require intricate models, potentially easing the balance between complexity and efficiency.

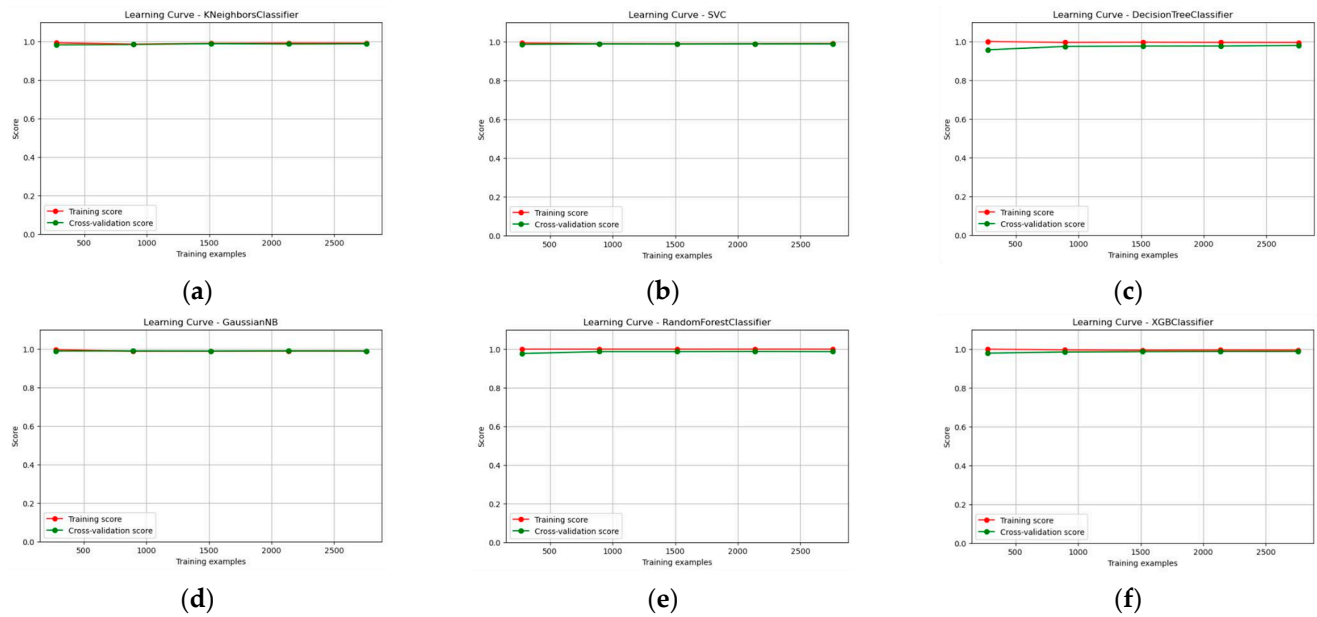


Figure 4. Learning curves for each classifier. (a) KNN model, (b) SVM classifier, (c) Decision Tree classifier, (d) GaussianNB, (e) Random Forest classifier, and (f) XGBoost classifier.

However, the almost perfect scores could also be indicative of an easy-to-learn dataset or possibly a data leakage issue. If the task is genuinely this well performed by all models, it may be worth investigating if the models are indeed learning the underlying patterns or simply memorizing the data.

Each ROC curve (Figure 5) plots the true positive rate against the false positive rate at various threshold settings, demonstrating the trade-offs between sensitivity and specificity. An AUC of 1.0 for all classes across classifiers indicates perfect classification with no overlap between the positive and negative class distributions. The Decision Tree's ROC curve (Figure 5c) shows nearly perfect performance with an AUC close to 1.00, but there is a slight distance from the top-left corner for the minimal and moderate classes, indicating very minor misclassification. KNN, SVM, GaussianNB, Random Forest, and XGBoost show an identical ROC curve occupying the top-left corner, which represents an AUC of approximately 1.00 for all classes, signaling no misclassification errors, and that the ensemble method is highly effective with no trade-off between sensitivity and specificity.

These ROC curves indicate the models are robust and highly capable of distinguishing between the various classes of knee osteoarthritis. However, such results may also necessitate a review of the experimental setup to ensure that the results are genuine and not due to an artifact of the data or evaluation process.

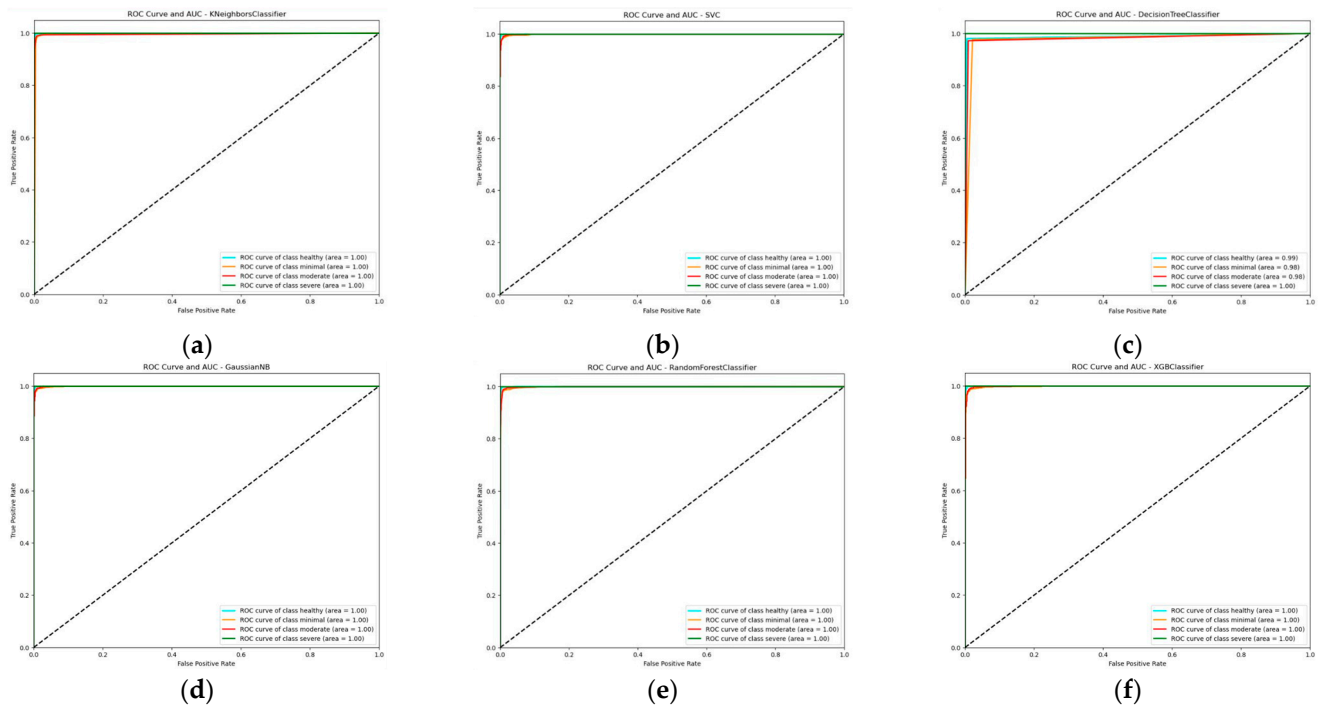


Figure 5. ROC and AUC curves for the various models. (a) KNN model, (b) SVM classifier, (c) Decision Tree classifier, (d) GaussianNB, (e) Random Forest classifier, and (f) XGBoost classifier.

3.4. Ensemble Model

An ensemble model has also been explored, combining the uses of each and all of the best parameter tuning of the six classifiers. The confusion matrix (Figure 6) visualizes the model's performance, with the majority of predictions concentrated along the diagonal, indicating correct classifications. There are a few off-diagonal numbers representing misclassifications, but these are very low with 4 instances of 'healthy' misclassified as 'minimal', 5 instances of 'minimal' misclassified as 'healthy' and 13 as 'moderate', and 13 instances of 'moderate' misclassified as 'minimal'. No instances were misclassified as 'severe', and all 'severe' cases were correctly classified.

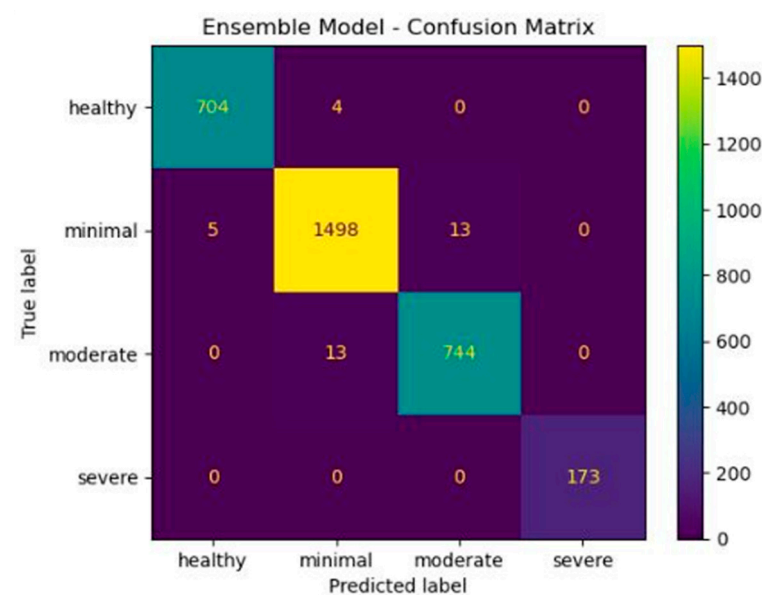


Figure 6. Confusion matrix of the ensemble model.

The classification report shows the precision, recall, and F1 score for four categories: healthy, minimal, moderate, and severe. All categories have very high metrics, nearly perfect, with precision, recall, and F1 scores all above 0.98. The ‘severe’ category has perfect scores of 1.00 across all three metrics. The model also achieves a near-perfect overall accuracy of approximately 98.90%.

The ensemble model displays a similar performance compared to XGBoost on the knee osteoarthritis classification task. The high values in precision, recall, and F1 score suggest that the model is both accurate and consistent in its predictions across all classes. The perfect scores in the ‘severe’ category indicate that the model is particularly effective at identifying the most advanced cases of osteoarthritis, which is critical from a clinical perspective.

The very few misclassifications between ‘healthy’, ‘minimal’, and ‘moderate’ suggest some overlap in the features of these classes, which is to be expected in medical imaging where disease progression can be gradual. The same can also be observed from the feature extraction.

The overall high accuracy and the detailed breakdown of performance metrics across classes show that the ensemble method is highly effective for this dataset. However, the near-perfect performance should be scrutinized to ensure that it is not the result of overfitting, data leakage, or an overly simplistic task. It is also important to validate these findings on an independent test set, particularly one that reflects real-world class imbalances and variations.

3.5. Comparison with Available Literature

A brief comparison with other methodologies for KOA detection using radiography images including feature extraction and machine learning classifier was made and summed up in Table 3. The proposed technique shows significant improvement in model accuracy, leading to a high potential for clinical application.

Table 3. KOA detection comparison of the literature.

References	Methodology	Accuracy
[17]	SVM with Structural 2-dimensional Principle Component Analysis	94.33%
[21]	SVM with HOG feature extraction	~95%
[22]	RF and NB with Independent Component Analysis	82.98%
[27]	LR with HOG + Haralick Features	84.50%
This study	XGBoost Classifier with CLAHE, HOG + LDA feature extraction, Min–Max scaling	98.90%

4. Conclusions

This study has made a significant contribution to the field of medical imaging and diagnostics by rigorously evaluating machine learning (ML) for the diagnosis of knee osteoarthritis (KOA).

We preprocessed the images with CLAHE and then extracted features using a combination of Histogram of Oriented Gradients (HOG), Linear Discriminant Analysis (LDA), and Min–Max scaling, which is critical in enhancing the feature space for machine learning models, enabling more precise and distinguishable classifications. HOG captures essential edge and shape information from images, which, when combined with LDA’s ability to maximize class separability, greatly improves the model’s ability to identify nuanced differences between classes. Min–Max scaling ensures that these extracted features contribute equally to the model’s performance, preventing features with larger scales from dominating the learning process and improving the overall diagnostic accuracy.

As far as the team knowledge, the combination of HOG, LDA, and Min–Max scaling for feature extraction is yet to be performed in the previous literature and the effectiveness

of the method has been proven by the clear distinction between the categories, providing a basis for the success of the machine learning model.

The study has highlighted the immense potential of computational approaches to enhance diagnostic precision. In particular, models like XGBoost and the ensemble model have demonstrated high accuracy rates of 98.90% in multi-class classification tasks.

It has also underscored the importance of fine-tuning hyperparameters and the strategic extraction of feature extraction methods to maximize model performance. These insights are pivotal for the development of diagnostic tools that are both precise and reliable. As the landscape of medical imaging evolves, leveraging ML and DL promises to transform the diagnosis and management of various health conditions, including KOA.

Looking ahead, future research should aim at enhancing these models further, incorporating broader datasets, and extending these computational methods across different diagnostic modalities such as MRI, CT scans, and ultrasound. Integrating varied data sources could offer a more holistic analysis of knee health. The ongoing progress in ML and DL within medical imaging is poised to deliver considerable benefits in patient treatment and healthcare outcomes, signaling a new era in clinical diagnostics.

Author Contributions: Conceptualization, C.T.S.C. and T.T.N.; methodology, C.T.S.C. and H.-C.L.; validation, A.R., T.-L.P. and H.-C.L.; formal analysis, A.R., N.V.H. and H.-C.L.; investigation, A.R.; data curation, A.R.; writing—original draft preparation, A.R. and T.-L.P.; writing—review and editing, T.-L.P. and C.T.S.C.; visualization, A.R., N.V.H. and T.T.N. All authors have read and agreed to the published version of the manuscript.

Funding: This research is partially supported by grants from National Science and Technology Council, Taiwan (NSTC 112-2221-E-005-042- and NSTC 111-2221-E-005-018-).

Data Availability Statement: The original dataset presented in the study are openly available at [36].

Acknowledgments: The authors would like to give acknowledgement to Dao Nguyen Phuong Linh M.D., Master of Pediatrics (Ho Chi Minh city 700000, Vietnam) for her valuable insights into the problem of knee osteoarthritis.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Hsu, H.; Siwicz, R.M. Knee Osteoarthritis. In *StatPearls*; StatPearls Publishing: Treasure Island, FL, USA, 2024; Available online: <http://www.ncbi.nlm.nih.gov/books/NBK507884/> (accessed on 1 February 2024).
2. Hayashi, D.; Roemer, F.W.; Jarraya, M.; Guermazi, A. Imaging in Osteoarthritis. *Radiol. Clin. N. Am.* **2017**, *55*, 1085–1102. [CrossRef]
3. Schiphof, D.; Boers, M.; Bierma-Zeinstra, S.M.A. Differences in descriptions of Kellgren and Lawrence grades of knee osteoarthritis. *Ann. Rheum. Dis.* **2008**, *67*, 1034–1036. [CrossRef] [PubMed]
4. Kellgren, J.H.; Lawrence, J.S. Radiological Assessment of Osteo-Arthrosis. *Ann. Rheum. Dis.* **1957**, *16*, 494–502. [CrossRef]
5. Chen, P.; Gao, L.; Shi, X.; Allen, K.; Yang, L. Fully automatic knee osteoarthritis severity grading using deep neural networks with a novel ordinal loss. *Comput. Med. Imaging Graph.* **2019**, *75*, 84–92. [CrossRef] [PubMed]
6. Tiulpin, A.; Thevenot, J.; Rahtu, E.; Lehenkari, P.; Saarakkala, S. Automatic Knee Osteoarthritis Diagnosis from Plain Radiographs: A Deep Learning-Based Approach. *Sci. Rep.* **2018**, *8*, 1727. [CrossRef] [PubMed]
7. Cui, A.; Li, H.; Wang, D.; Zhong, J.; Chen, Y.; Lu, H. Global, regional prevalence, incidence and risk factors of knee osteoarthritis in population-based studies. *eClinicalMedicine* **2020**, *29–30*, 100587. [CrossRef] [PubMed]
8. Cross, M.; Smith, E.; Hoy, D.; Nolte, S.; Ackerman, I.; Fransen, M.; Bridgett, L.; Williams, S.; Guillemin, F.; Hill, C.L.; et al. The global burden of hip and knee osteoarthritis: Estimates from the Global Burden of Disease 2010 study. *Ann. Rheum. Dis.* **2014**, *73*, 1323–1330. [CrossRef] [PubMed]
9. Tiulpin, A.; Saarakkala, S. Automatic Grading of Individual Knee Osteoarthritis Features in Plain Radiographs Using Deep Convolutional Neural Networks. *Diagnostics* **2020**, *10*, 932. [CrossRef] [PubMed]
10. Swamy, M.S.M.; Holi, M.S. Knee joint cartilage visualization and quantification in normal and osteoarthritis. In Proceedings of the 2010 International Conference on Systems in Medicine and Biology, Kharagpur, India, 16–18 December 2010; IEEE: New York, NY, USA, 2010; pp. 138–142. [CrossRef]
11. Li, L. Deep Residual Autoencoder with Multiscaling for Semantic Segmentation of Land-Use Images. *Remote Sens.* **2019**, *11*, 2142. [CrossRef]

12. Saygili, A. A new approach for computer-aided detection of coronavirus (COVID-19) from CT and X-ray images using machine learning methods. *Appl. Soft Comput.* **2021**, *105*, 107323. [\[CrossRef\]](#)
13. Ha, M.-K.; Phan, T.-L.; Nguyen, D.H.H.; Quan, N.H.; Ha-Phan, N.-Q.; Ching, C.T.S.; Hieu, N.V. Comparative Analysis of Audio Processing Techniques on Doppler Radar Signature of Human Walking Motion Using CNN Models. *Sensors* **2023**, *23*, 8743. [\[CrossRef\]](#) [\[PubMed\]](#)
14. Huang, X.; Han, X.; Ma, S.; Lin, T.; Gong, J. Monitoring ecosystem service change in the City of Shenzhen by the use of high-resolution remotely sensed imagery and deep learning. *Land Degrad. Dev.* **2019**, *30*, 1490–1501. [\[CrossRef\]](#)
15. Hapsari, Y.; Syamsuryadi. Weather Classification Based on Hybrid Cloud Image Using Principal Component Analysis (PCA) and Linear Discriminant Analysis (LDA). *J. Phys. Conf. Ser.* **2019**, *1167*, 012064. [\[CrossRef\]](#)
16. Anifah, L.; Purnama, I.K.E.; Hariadi, M.; Purnomo, M.H. Osteoarthritis Classification Using Self Organizing Map Based on Gabor Kernel and Contrast-Limited Adaptive Histogram Equalization. *Open Biomed. Eng. J.* **2013**, *7*, 18–28. [\[CrossRef\]](#)
17. Wahyuningrum, R.T.; Anifah, L.; Purnama, I.K.E.; Purnomo, M.H. A novel hybrid of S2DPCA and SVM for knee osteoarthritis classification. In Proceedings of the 2016 IEEE International Conference on Computational Intelligence and Virtual Environments for Measurement Systems and Applications (CIVEMSA), Budapest, Hungary, 27–28 June 2016; IEEE: New York, NY, USA, 2016; pp. 1–5. [\[CrossRef\]](#)
18. Mohammed, A.S.; Hasanaath, A.A.; Latif, G.; Bashar, A. Knee Osteoarthritis Detection and Severity Classification Using Residual Neural Networks on Preprocessed X-ray Images. *Diagnostics* **2023**, *13*, 1380. [\[CrossRef\]](#)
19. Kotti, M.; Duffell, L.D.; Faisal, A.A.; McGregor, A.H. Detecting knee osteoarthritis and its discriminating parameters using random forests. *Med. Eng. Phys.* **2017**, *43*, 19–29. [\[CrossRef\]](#)
20. Kokkoti, C.; Moustakidis, S.; Papageorgiou, E.; Giakas, G.; Tsaopoulos, D.E. Machine learning in knee osteoarthritis: A review. *Osteoarthr. Cartil. Open* **2020**, *2*, 100069. [\[CrossRef\]](#) [\[PubMed\]](#)
21. Gornale, S.S.; Patravali, P.U.; Manza, R.R. Detection of Osteoarthritis Using Knee X-ray Image Analyses: A Machine Vision based Approach. *Int. J. Comput. Appl.* **2016**, *145*, 20–26.
22. Brahim, A.; Jennane, R.; Riad, R.; Janvier, T.; Khedher, L.; Toumi, H.; Lespessailles, E. A decision support tool for early detection of knee OsteoArthritis using X-ray imaging and machine learning: Data from the OsteoArthritis Initiative. *Comput. Med. Imaging Graph.* **2019**, *73*, 11–18. [\[CrossRef\]](#)
23. Mehta, S.; Gaur, A.; Sarathi, M.P. A Simplified Method of Detection and Predicting the Severity of Knee Osteoarthritis. In Proceedings of the 2023 14th International Conference on Computing Communication and Networking Technologies (ICCCNT), Delhi, India, 6–8 July 2023; IEEE: New York, NY, USA, 2023; pp. 1–7. [\[CrossRef\]](#)
24. Mahum, R.; Rehman, S.U.; Meraj, T.; Rauf, H.T.; Irtaza, A.; El-Sherbeeney, A.M.; El-Meligy, M.A. A Novel Hybrid Approach Based on Deep CNN Features to Detect Knee Osteoarthritis. *Sensors* **2021**, *21*, 6189. [\[CrossRef\]](#)
25. Bayramoglu, N.; Tiulpin, A.; Hirvasniemi, J.; Nieminen, M.T.; Saarakkala, S. Adaptive segmentation of knee radiographs for selecting the optimal ROI in texture analysis. *Osteoarthr. Cartil.* **2020**, *28*, 941–952. [\[CrossRef\]](#) [\[PubMed\]](#)
26. Belete, D.M.; Huchaiah, M.D. Grid search in hyperparameter optimization of machine learning models for prediction of HIV/AIDS test results. *Int. J. Comput. Appl.* **2022**, *44*, 875–886. [\[CrossRef\]](#)
27. Tariq, T.; Suhail, Z.; Nawaz, Z. Machine Learning Approaches for the Classification of Knee Osteoarthritis. In Proceedings of the 2023 3rd International Conference on Electrical, Computer, Communications and Mechatronics Engineering (ICECCME), Tenerife, Canary Islands, Spain, 19–21 July 2023; IEEE: New York, NY, USA, 2023; pp. 1–6. [\[CrossRef\]](#)
28. Thornton, C.; Hutter, F.; Hoos, H.H.; Leyton-Brown, K. Auto-WEKA: Combined selection and hyperparameter optimization of classification algorithms. In Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Chicago, IL, USA, 11–14 August 2013; ACM: New York, NY, USA, 2013; pp. 847–855. [\[CrossRef\]](#)
29. DeCastro-García, N.; Castañeda, Á.L.M.; García, D.E.; Carriegos, M.V. Effect of the Sampling of a Dataset in the Hyperparameter Optimization Phase over the Efficiency of a Machine Learning Algorithm. *Complexity* **2019**, *2019*, 6278908. [\[CrossRef\]](#)
30. Zhang, S.; Li, X.; Zong, M.; Zhu, X.; Wang, R. Efficient kNN Classification with Different Numbers of Nearest Neighbors. *IEEE Trans. Neural Netw. Learn. Syst.* **2018**, *29*, 1774–1785. [\[CrossRef\]](#) [\[PubMed\]](#)
31. Chapelle, O.; Vapnik, V.; Bousquet, O.; Mukherjee, S. Choosing Multiple Parameters for Support Vector Machines. *Mach. Learn.* **2002**, *46*, 131–159. [\[CrossRef\]](#)
32. Probst, P.; Wright, M.; Boulesteix, A.-L. Hyperparameters and Tuning Strategies for Random Forest. *arXiv* **2018**, arXiv:1804.03515. [\[CrossRef\]](#)
33. Quinlan, J.R. Induction of decision trees. *Mach. Learn.* **1986**, *1*, 81–106. [\[CrossRef\]](#)
34. Ntakolia, C.; Kokkoti, C.; Moustakidis, S.; Tsaopoulos, D. A machine learning pipeline for predicting joint space narrowing in knee osteoarthritis patients. In Proceedings of the 2020 IEEE 20th International Conference on Bioinformatics and Bioengineering (BIBE), Cincinnati, OH, USA, 26–28 October 2020; pp. 934–941. [\[CrossRef\]](#)
35. Chen, T.; Guestrin, C. XGBoost: A Scalable Tree Boosting System. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, 13–17 August 2016; ACM: New York, NY, USA, 2016; pp. 785–794. [\[CrossRef\]](#)
36. Chen, P. Knee Osteoarthritis Severity Grading Dataset. *Mendeley* **2018**. [\[CrossRef\]](#)

37. Dalal, N.; Triggs, B. Histograms of Oriented Gradients for Human Detection. In Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05), San Diego, CA, USA, 20–25 June 2005; IEEE: New York, NY, USA, 2005; pp. 886–893. [\[CrossRef\]](#)
38. Gornale, S.S.; Patravali, P.U.; Marathe, K.S.; Hiremath, P.S. Determination of Osteoarthritis Using Histogram of Oriented Gradients and Multiclass SVM. *Int. J. Image Graph. Signal Process.* **2017**, *9*, 41–49. [\[CrossRef\]](#)
39. Zöller, M.-A.; Huber, M.F. Benchmark and Survey of Automated Machine Learning Frameworks. *arXiv* **2019**, arXiv:1904.12054. [\[CrossRef\]](#)
40. Hutter, F.; Kotthoff, L.; Vanschoren, J. (Eds.) Automated Machine Learning: Methods, Systems, Challenges. In *The Springer Series on Challenges in Machine Learning*; Springer International Publishing: Cham, Switzerland, 2019. [\[CrossRef\]](#)

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.