**Integrating Deep Learning and Machine Learning for Enhanced Diagnosis of Hemophilic Arthropathy in Ankle X-ray Imaging**

**Abstract**

Hemophilia, caused by a deficiency in clotting factors VIII and IX, primarily affects males and leads to severe joint damage known as hemophilic arthropathy. This condition mainly impacts joints such as ankles, knees, and elbows due to frequent bleeding episodes. While radiography is a common diagnostic tool, integrating Artificial Intelligence (AI) has shown significant potential to improve accuracy and efficiency. This study evaluates the effectiveness of combining deep learning (DL) and machine learning (ML) algorithms to classify ankle X-ray images as 'Healthy' or 'Ill,' focusing on hemophilic arthropathy.

The Inception V3 architecture was employed for extracting deep features, while classifiers like SVM, KNN, and XGBoost were used for categorization. Among these, the SVM model achieved the highest accuracy of 97.78% and an Area Under the Curve (AUC) score of 0.99, demonstrating its strong diagnostic capability. The KNN and XGBoost classifiers also performed well with 95.56% accuracy, though their AUC scores were slightly lower.

This study highlights the role of AI in improving diagnostic processes, reducing the time required for diagnosis, and enhancing accuracy, particularly in resource-limited settings. Future work should aim to validate these findings on larger datasets, extend applications to other joint diseases, and integrate advanced imaging techniques. Leveraging AI in clinical practice promises a new era of precision medicine, enhancing outcomes for hemophilia patients and beyond.

**Keywords**:  
Deep Learning, Machine Learning, Artificial Intelligence, Hemophilia, Hemophilic Arthropathy, Ankle X-ray Classification, Inception V3, Support Vector Machine (SVM), K-Nearest Neighbors (KNN), XGBoost, Precision Medicine.

**1. Introduction**

Hemophilia is a rare but serious bleeding disorder caused by the deficiency of clotting factors VIII and IX, with Hemophilia A and Hemophilia B being the most common types. These conditions predominantly affect males due to their X-linked inheritance pattern [1]. Hemophilia is characterized by recurrent episodes of joint bleeding, particularly in the ankles, knees, and elbows, leading to severe hemophilic arthropathy, a debilitating condition [2]. Hemophilia A is the most prevalent form, accounting for 80-85% of all hemophilia cases worldwide [3]. According to the World Federation of Hemophilia's 2020 report, an estimated 815,100 individuals globally have hemophilia, but only 347,026 of these cases are diagnosed, with 276,900 being classified as severe cases [4]. The prevalence rate for Hemophilia A is approximately 1 in 5,000 males, while Hemophilia B affects around 1 in 30,000 males [5]. Notably, over 90% of severe hemophilia cases develop joint diseases, which contribute significantly to both morbidity and the financial burden on healthcare systems [6].

Intra-articular bleeding, which occurs in 80% of hemophilia cases, is a defining feature of the disorder. This bleeding causes joint inflammation and degeneration, leading to complications such as cartilage damage, synovium thickening, and bone deformities [7]. The ankle joint, which is the third most frequently affected after the knee and elbow, is particularly vulnerable due to its wide range of motion and the stress placed on it during physical activity [8]. As the disease progresses, treatment options such as physical therapy, orthoses, and advanced surgical procedures like arthrodesis or total ankle replacement are considered [10]. However, if left untreated, even mild bleeding episodes can progress into debilitating Hemophilic Arthropathy (HA), characterized by joint damage, severe pain, and the eventual need for joint replacement [11].

Joint damage in hemophilia typically begins in early childhood, when physical activity and movement trigger bleeding episodes [12]. Despite the use of clotting factor replacement therapies, arthropathy remains persistent, especially in older patients. Hemoglobin breakdown products, particularly iron deposits, accumulate in the joint, causing progressive tissue damage. While the exact biological mechanisms of this progression remain unclear, it is evident that early intervention is crucial to prevent long-term damage [13]. Thus, improved diagnostic tools and methods are essential for enhancing care and outcomes for individuals with hemophilia.

Radiography continues to be the most commonly used diagnostic tool in hemophilia care due to its affordability and widespread availability. However, despite the availability of more advanced imaging modalities like MRI and CT, radiographs have limitations. They are unable to detect approximately 23% of ankle fractures, a significant challenge in ensuring accurate diagnosis and treatment [14]. The early stages of hemophilic arthropathy are characterized by soft tissue swelling, while advanced stages show more severe signs, such as bone cysts, cartilage loss, and irregular joint profiles [17]. The Pettersson scoring system, which assigns a cumulative score of up to 13, is the standard for assessing joint damage through X-ray imaging [18].

The use of artificial intelligence (AI), particularly deep learning (DL) and machine learning (ML) techniques, presents an exciting opportunity for improving orthopedic diagnostics, including the analysis of X-rays. AI has the potential to automate X-ray analysis, providing more accurate and efficient assessments. While traditional ML approaches rely on manually engineered features and algorithms like Principal Component Analysis (PCA) and random forests, DL methods use Convolutional Neural Networks (CNNs) to directly process image data, enhancing diagnostic accuracy [19-24]. Pre-trained models such as Inception V3 can further enhance diagnostic capabilities through transfer learning, enabling more efficient and effective analysis [26]. While DL methods typically require substantial computational resources, they have demonstrated significant success in identifying musculoskeletal conditions without the need for radiologist input [27-30].

Several studies using DL models like Inception V3, ResNet, and Xception have shown high accuracy in classifying ankle fractures, with some achieving sensitivity and specificity rates above 98% [31-37]. For example, a study using Inception V3 achieved 98.7% sensitivity and 98.6% specificity in detecting ankle fractures, demonstrating the efficacy of multi-view imaging strategies compared to single-view approaches [37]. However, while many studies have focused on knee osteoarthritis or fractures, there is a gap in the application of DL techniques to classify hemophilic arthropathy using the Pettersson scoring system. This research aims to fill that gap by leveraging DL for feature extraction, PCA for dimensionality reduction, and ML classifiers such as Support Vector Machine (SVM), K-Nearest Neighbors (KNN), and XGBoost to differentiate between healthy and HA-affected ankles. By focusing on binary classification, this approach simplifies the diagnostic process, reduces costs, and improves early disease management, ultimately leading to better patient outcomes.

**2. Materials and Methods:**

This section outlines the methodology for classifying X-ray images into "healthy" or "ill" categories by combining deep learning and traditional machine learning, as shown in Figure 1. The process begins with extensive preprocessing, including resizing, normalization, and data augmentation, to standardize the images and ensure consistency. The dataset is then split into training and testing subsets, and Synthetic Minority Over-sampling Technique (SMOTE) is applied to address class imbalances in the training set, improving model generalization.

Next, the balanced training data is fed into a customized InceptionV3 neural network, optimized for binary classification. Early stopping is used to prevent overfitting during training, ensuring the model doesn’t degrade performance on the validation set. After training, feature extraction is performed, and Principal Component Analysis (PCA) reduces the feature space while retaining key information. The reduced features are then classified using advanced machine learning models.

The methodology is evaluated using classification reports and confusion matrices from both the training and testing datasets, providing insights into performance and areas for improvement. This iterative process ensures maximum model accuracy and reliability, making the system a robust tool for classifying medical X-ray images and advancing healthcare diagnostics.

****

**Figure 1:** Flow chart for the suggested model.

**2.1 Dataset Preparation:**

In this study, X-ray images from patients were sourced from a medical hospital, focusing on individuals diagnosed with hemophilic arthropathy and comparing them with healthy controls. Ethical approval for the study was obtained from the Institutional Review Board (IRB) under Approval No. CMUH113-REC1-002, which waived the requirement for informed and written consent. This allowed for the evaluation of X-ray films using Pettersson scores, a method that aids in assessing disease severity and progression.

The dataset consisted of X-ray images depicting both anterior-posterior (AP) and lateral views of the ankles. These images were analyzed by a team of senior medical professionals, including an orthopaedist specializing in foot and ankle disorders. Each X-ray was manually labeled based on the diagnosis, resulting in two groups: 111 individuals diagnosed with arthropathy ("ill") and an equal number of healthy individuals. To maintain patient confidentiality and adhere to ethical guidelines, no personal identifying information such as age, gender, or medical history was included in the dataset. The data collection and preparation processes were conducted in compliance with strict ethical standards, ensuring the integrity of patient data and providing valuable insights into medical conditions through X-ray analysis.

**2.2 Data Pre-processing:**

The dataset for classifying X-ray images into "healthy" and "ill" categories consisted of 222 images. Each image was preprocessed by resizing to 256×256 pixels, normalizing pixel values to a range of 0 to 1, and stored in arrays for easy handling. Stratified sampling was used to divide the dataset into training (80%, 177 images) and testing (20%, 45 images) subsets, maintaining the distribution of labels. To address class imbalance, SMOTE generated synthetic examples for the underrepresented class, balancing the dataset. Data augmentation techniques, including rotation, shifts, shear, zoom, and flipping, were applied to increase variability and improve generalization, reducing overfitting. After SMOTE and augmentation, the training set grew to 540 images, enhancing the model’s robustness for real-world clinical applications.

**Feature Extraction:**

**InceptionV3 Architecture:**

InceptionV3 was chosen for its high performance in medical image classification. Originally developed for ImageNet, it was adapted for binary classification by removing the top layer (include\_top=False). The model was initialized with pre-trained ImageNet weights, leveraging learned features that aid in medical imaging. The input was configured to accept 256×256 RGB images.

**Custom Layer Integration:**

The custom layer integration begins with a Global Average Pooling 2D layer, which reduces the dimensionality of the InceptionV3 base model's output by condensing each feature map into a single average value, enhancing simplicity and reducing overfitting risks. A dense layer with 756 units follows, employing LeakyReLU activation to mitigate the "dying ReLU" issue by allowing a small positive gradient for non-active neurons, ensuring continuous learning of complex patterns. To improve generalization, a dropout layer with a 50% rate introduces regularization by randomly deactivating half of the neurons during training, compelling the network to learn robust features. The architecture concludes with a sigmoid output layer, generating a probability score that determines the likelihood of the X-ray exhibiting pathological features categorized as "ill."

**Training Strategy:**

Early Stopping: Stops training after five epochs with no improvement in validation loss, restoring weights from the best epoch.

Training Execution: Training was done over 50 epochs with a batch size of 32 on both original and augmented images to ensure robustness. Figure 2 shows the model structure for feature extraction.

|  |
| --- |
|  |

**Figure 2:** Flow Diagram for the model used for feature extraction.

**2.3 Dimensionality Reduction with PCA:**

**Component Selection:** After feature extraction with InceptionV3, Principal Component Analysis (PCA) was applied to reduce the dimensionality of the extracted features. The PCA model retained a maximum of 500 components or fewer, depending on the feature set size. This was necessary as InceptionV3 generates a large number of features, some of which may be redundant, leading to challenges in binary classification due to noise or redundancy.

**Purpose of PCA:** PCA was used to simplify the feature set by retaining the most informative components, focusing on those with the greatest variance. This reduces computational complexity, alleviates overfitting, and enhances the performance and generalization capabilities of subsequent classifiers.

**PCA Transformation:** PCA isolates principal components that account for most of the variance, reducing the dataset’s dimensionality and mitigating overfitting. By retaining only the most critical features, PCA enhances model robustness and reliability.

**Integration with Classifiers:** After PCA transformation, the simplified dataset was fed into classifiers like Support Vector Machines (SVM), K-Nearest Neighbors (KNN), and XGBoost. The reduced complexity allowed these models to operate more efficiently and effectively, improving their ability to discern meaningful patterns.

**Overall Importance of PCA:** PCA bridges the gap between feature extraction and classification, optimizing the feature set and enhancing classifier performance. This process improves generalization to unseen data, contributing to the accuracy and robustness of the classification system.

* 1. **Classification via Machine Learning Classifiers :**

Three prominent machine learning algorithms—Support Vector Machine (SVM), K-Nearest Neighbors (KNN), and Extreme Gradient Boosting (XGBoost)—were employed for forecasting results due to their reliability in image classification. Hyperparameter tuning was crucial to improve model performance, reduce bias, and prevent overfitting, with optimized values selected based on literature.

The SVM model used a sigmoid kernel to handle non-linear relationships in medical datasets, with a regularization parameter (C) set to 1 and gamma adjusted based on feature count to fine-tune the decision boundary. The KNN classifier was set to 11 neighbors with a uniform weighting scheme, ensuring equal voting power among neighbors and preventing bias.

The XGBoost classifier, optimized for binary classification, utilized a binary logistic regression framework with 200 estimators, a max depth of 7, and a learning rate of 0.01. To avoid overfitting, subsample and colsample\_bytree were set to 0.9 and 0.8, respectively, and gamma was set to 2 to ensure only significant splits. Hyperparameter tuning ensured optimal configurations for unbiased performance comparison across models.

* 1. **Model Evaluation:**

Model evaluation is a critical stage in machine learning, used to assess how well trained models predict unknown data. For the SVM, KNN, and XGBoost classifiers, evaluation involved several metrics and visual aids, such as the confusion matrix, which highlights true positives (TP), false positives (FP), false negatives (FN), and true negatives (TN). This matrix helps identify biases or errors that may be missed by accuracy alone.

The classification report further extends the evaluation with metrics like precision, recall, F1 score, and accuracy. These metrics offer a more detailed understanding of each model’s performance, highlighting strengths and weaknesses across different classes and improving the interpretation of the model's predictive ability.

|  |  |
| --- | --- |
|  | (1) |
|  | (2) |
|  | (3) |
|  | (4) |

Finally, the evaluation incorporates the AUC score and the ROC curve. The ROC curve plots the true positive rate against the false positive rate at different threshold settings, aiding in identifying the most effective model. Higher values on this curve indicate better model performance. The AUC score quantifies the area under the entire ROC curve and is a comprehensive performance measure across all possible classification thresholds.

Together, these evaluation techniques establish a robust framework for understanding and assessing the performance of machine learning models, ensuring a detailed appreciation of their predictive capabilities.

* 1. **Computational Environments :**

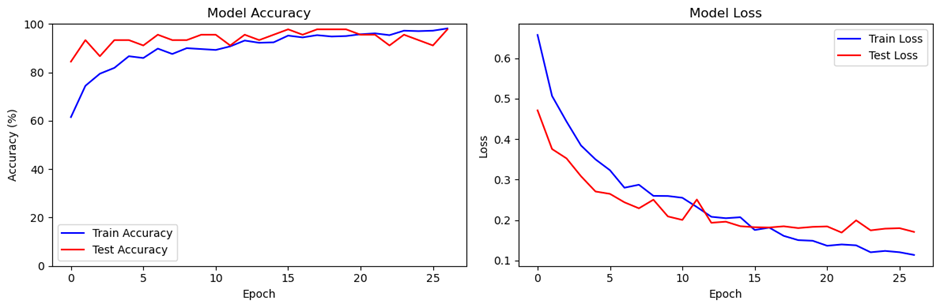
The Jupyter Notebook environment, leveraging Python 3.11.4, served as the platform for developing and evaluating machine learning classification models. The computational tasks were efficiently managed using a high-performance setup, which included an Intel® Core™ i7-10700 CPU operating at 2.90 GHz, 32 GB of RAM, and an Nvidia GeForce RTX 3070 graphics card for GPU-intensive tasks like image processing and model training. This robust configuration ensured smooth handling of datasets, facilitated model training, supported hyperparameter optimization, and enabled thorough validation.

**3. Results and Discussion:**

The experimental findings for each model are presented in this section.

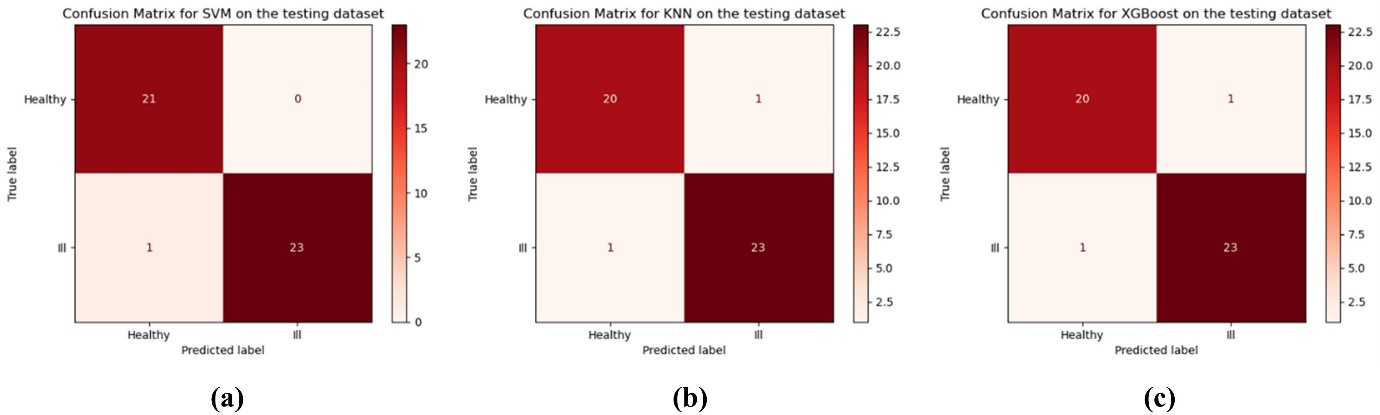
**3.1 Accuracy and Loss Over Epochs:** This section presents the experimental results of the proposed Inception V3 architecture, trained from scratch. The **Model Accuracy Graph** features the vertical axis representing accuracy as a percentage (0–100%) and the horizontal axis marking training progress over 25 epochs. The blue line, indicating the model's training accuracy, shows a sharp improvement early on, rising from approximately 60% to over 80% within the initial epochs and plateauing near 95%, signaling the model's peak performance on the training data. Similarly, the red line, representing test dataset accuracy, mirrors the training accuracy with minor variations. The minimal gap and parallel trend between the two lines highlight effective generalization, with no significant overfitting or underfitting observed.

In **Figure 3**, the **Model Loss Graph** complements the accuracy graph by visualizing prediction error, where lower values indicate better performance. The x-axis again represents the epoch count. The blue line, corresponding to training dataset loss, starts high and drops sharply, reflecting rapid learning before stabilizing in later epochs. The red line, representing test dataset loss, declines similarly but remains slightly higher than the training loss throughout. The close alignment and parallel trend of these lines confirm balanced performance on both training and test data, indicating the model generalizes effectively without overfitting.



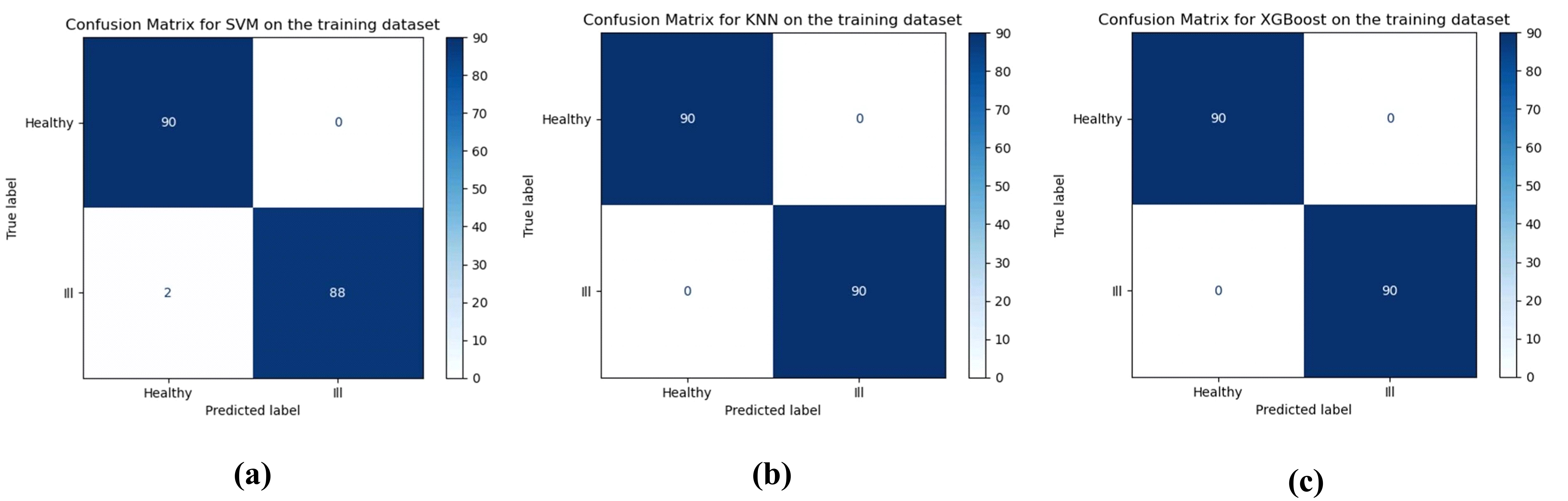
**Figure 3:** Illustrations of model performance, with the left image displaying the training and validation accuracy and the right image showing the loss for each epochs.

**3.2 Confusion Matrices and Classification Reports:** The confusion matrices for the testing dataset indicate a high number of correct predictions across three classifiers: SVM, KNN, and XGBoost, tested on a dataset with two classes, "Healthy" and "Ill." All models exhibit a similar prediction pattern, achieving high accuracy overall. The SVM classifier stands out with the highest number of correct predictions for the "Healthy" class, totaling 21 instances. KNN and XGBoost are slightly behind, each with one fewer correct prediction for "Healthy," but they achieve the same accuracy for "Ill" and maintain identical overall error rates. Figure 4 illustrates the confusion matrices for each classifier, highlighting minor variations in their performance and accuracy. These matrices underscore the consistency among the models while revealing subtle differences in their ability to classify the two classes accurately.



**Figure 4:** Confusion matrices of (a) SVM, (b) KNN, (c) XGB for the binary classification of HA on testing dataset.

The performance of the SVM, KNN, and XGBoost classifiers was evaluated on the training dataset, with results summarized in confusion matrices. The SVM classifier accurately identified all "Healthy" cases but misclassified two "Ill" cases. In contrast, both KNN and XGBoost achieved perfect accuracy, correctly classifying all cases. While this flawless performance may appear ideal, it raises concerns about potential overfitting, as such a tight fit to the training data might lead to poor generalization in real-world scenarios. Validation on an unseen test dataset is crucial to ensure the models perform reliably beyond the training set. Interestingly, the slight inaccuracy in SVM's predictions may suggest a better ability to generalize compared to the error-free results of KNN and XGBoost. Figure 5 displays the confusion matrices for each classifier, highlighting these performance differences on the training data.



**Figure 5:** Confusion matrices of (a) SVM, (b) KNN, (c) XGB for the binary classification of HA on the training dataset.

Each model was configured with tailored hyperparameters to optimize performance for binary classification. The SVM classifier used a sigmoid kernel, a C parameter of 1, gamma set to 'auto,' and enabled probability estimates, ensuring suitability for applications like healthcare. A consistent random state was maintained for reproducibility. The KNN classifier was optimized with 11 neighbors, uniform weights, and the 'auto' algorithm for selecting the best computational approach. The XGBoost model was configured with 200 estimators, a learning rate of 0.01, a max depth of 7, and adjusted subsample and colsample\_bytree parameters, along with a specific gamma value for a tailored fit to the training data.

The SVM classifier excelled on the testing dataset, achieving precision between 0.95 and 1.00 and an overall accuracy of 97.78%. It demonstrated perfect recall for "Healthy" instances and high recall for "Ill" instances, making it a reliable choice for applications like preventive medicine. Its testing performance, combined with an impressive 98.89% accuracy on the training dataset, highlights its robustness as a diagnostic tool.

In contrast, while KNN and XGBoost achieved flawless classification on the training dataset, their performance metrics dropped slightly on the testing data, with average sensitivity and specificity scores around 95.83% and 95.24%, and an accuracy of 95.56%. This discrepancy suggests possible overfitting, limiting their ability to generalize to new data. The SVM's slightly lower sensitivity during training, compared to its excellent testing performance, indicates a more generalized approach that avoids overfitting and enhances reliability on unseen datasets.

The SVM classifier's superior testing performance underscores its potential for real-world deployment, where consistent accuracy across diverse datasets is critical. Its ability to construct an optimized hyperplane for binary classification gives it an edge over KNN and XGBoost. Table 1 provides a detailed comparison of the classification reports for SVM, KNN, and XGBoost, summarizing their performance on both training and testing datasets.

Table 1: Classification reports of 3 machine mearning classifiers.

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Classifier** | **Dataset** | **Precision (Healthy)** | **Recall (Healthy)** | **F1-score (Healthy)** | **Precision (Ill)** | **Recall (Ill)** | **F1-score (Ill)** | **Accuracy** |
| SVM | Testing | 0.95 | 1.00 | 0.98 | 1.00 | 0.96 | 0.98 | 0.98 |
| SVM | Training | 0.98 | 1.00 | 0.99 | 1.00 | 0.98 | 0.99 | 0.99 |
| KNN | Testing | 0.95 | 0.95 | 0.95 | 0.96 | 0.96 | 0.96 | 0.96 |
| KNN | Training | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 |
| XGBoost | Testing | 0.95 | 0.95 | 0.95 | 0.96 | 0.96 | 0.96 | 0.96 |
| XGBoost | Training | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 |

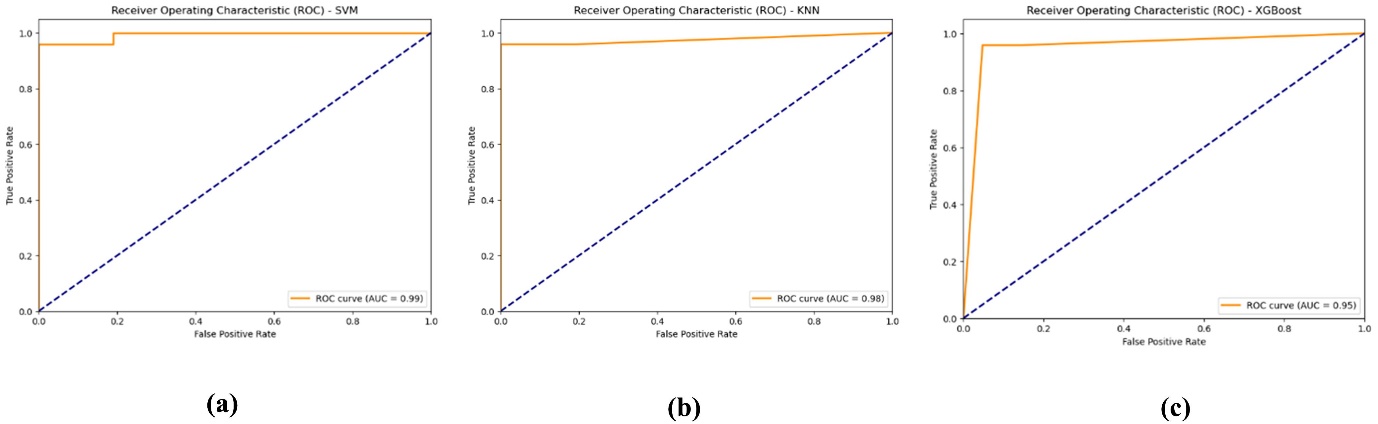
**3.3 ROC Curve, and AUC Score Report of the Machine learning classifier:**

The performance of three classifiers—XGBoost, KNN, and SVM—was thoroughly evaluated using Receiver Operating Characteristic (ROC) curves. The ROC curve is an essential tool that visualizes a classifier's ability to differentiate between classes across various threshold levels, plotting the True Positive Rate (TPR) against the False Positive Rate (FPR). The Area Under the Curve (AUC) serves as a summary metric, quantifying the model's overall ability to discriminate between positive and negative classes at all thresholds. Higher AUC values indicate better model performance, with a value of 1.0 representing perfect classification and 0.5 indicating no discriminative ability.

The XGBoost classifier achieved an AUC of 0.95, demonstrating strong discriminative ability by correctly identifying the "Healthy" and "Ill" classes with a 95% probability. This result indicates excellent model performance and reflects its high accuracy in predictive capabilities.

The KNN classifier, based on proximity in the feature space, scored an AUC of 0.98. This near-perfect value indicates a 98% probability of correctly classifying a randomly selected positive instance over a randomly selected negative one, showcasing the KNN model’s high separability and its effectiveness in distinguishing between the two classes.

The SVM classifier outperformed both KNN and XGBoost, achieving an AUC of 0.99. With a 99% probability of correctly discriminating between classes, the SVM displayed superior classification performance. This outstanding AUC highlights the model's exceptional sensitivity and specificity, making it a robust choice for applications where classification precision is critical, such as medical diagnostics, where misclassification could have severe consequences.

****

**Figure 6** **:** ROC and AUC curves for different classifiers (a) SVM, (b) KNN, (c) XGThe ROC curves for the XGBoost, KNN, and SVM classifiers present a compelling narrative of their robustness and acumen in binary classification.

**4. Conclusions**

In conclusion, our study has made significant progress in applying deep learning and machine learning algorithms to automate the diagnostic process of hemophilic arthropathy using X-ray imaging. The research successfully integrated deep learning, leveraging the powerful Inception V3 architecture for feature extraction, combined with dimension reduction techniques like PCA, and machine learning classifiers such as SVM, KNN, and XGBoost for the binary classification of X-ray images into 'Healthy' and 'Ill' categories.

The SVM classifier stood out with remarkable performance on the test data, achieving an accuracy of 97.78% and an AUC of 0.99, demonstrating exceptional discriminative power and confirming its clinical potential. While the KNN and XGBoost classifiers slightly lagged, with accuracy at 95.56% on the test data, they still proved reliable for classification tasks. Both models also showed strong discriminative abilities, reflected in their high AUC values. However, the performance gap between their training and test results indicates a possible overfitting concern that requires further investigation.

The potential of these computational models is promising; however, translating them from research to clinical practice requires addressing several challenges. These include expanding and diversifying datasets, validating models across multiple clinical settings, and seamlessly integrating AI tools into the diagnostic workflow. The study emphasized the importance of both algorithmic accuracy and model interpretability, which are crucial for clinician acceptance and ethical AI deployment.

While the initial outcomes are encouraging, the true success will depend on how well these models perform in real-world clinical environments and their impact on patient outcomes. Future research should focus on scaling these models and diversifying datasets to include a wider range of joint disorders. Further enhancement of computational methods to incorporate other diagnostic imaging modalities, such as ultrasound, CT, and MRI scans, will be vital in creating a more comprehensive diagnostic framework.

Rigorous validation across diverse clinical settings is essential to ensure these models are not only technically sound but also practically applicable in real-world scenarios. Expanding the focus beyond hemophilic arthropathy to include a broader range of joint diseases will be a critical step in realizing the full potential of AI in medical diagnostics. As machine learning and deep learning continue to advance in medical imaging, they hold the potential to revolutionize clinical diagnostics, improving patient care and treatment outcomes significantly. This marks the beginning of a new era in clinical diagnostics, driven by AI’s ability to enhance the precision and effectiveness of patient care.

**5.** **References**

[1] T. Greco, C. Polichetti, A. Cannella, V. La Vergata, G. Maccauro, and C. Perisano, “Ankle hemophilic arthropathy: literature review,” *Am. J. Blood Res.*, vol. 11, no. 3, pp. 206–216, 2021.

[2] D. J. Bowen, “Haemophilia A and haemophilia B: molecular insights,” *Mol. Pathol.*, vol. 55, no. 2, pp. 127–144, Apr. 2002, doi: 10.1136/mp.55.2.127.

[3] J. S. Stonebraker, P. H. B. Bolton‐Maggs, J. Michael Soucie, I. Walker, and M. Brooker, “A study of variations in the reported haemophilia B prevalence around the world,” *Haemophilia*, vol. 18, no. 3, May 2012, doi: 10.1111/j.1365-2516.2011.02588.x.

[4] “Annual Global Survey,” WFH - World Federation of Hemophilia. Accessed: Jan. 04, 2024. [Online]. Available: https://wfh.org/research-and-data-collection/annual-global-survey/

[5] P. H. Bolton-Maggs and K. J. Pasi, “Haemophilias A and B,” *The Lancet*, vol. 361, no. 9371, pp. 1801–1809, May 2003, doi: 10.1016/S0140-6736(03)13405-8.

[6] M. W. Hilgartner, “Current treatment of hemophilic arthropathy:,” *Curr. Opin. Pediatr.*, vol. 14, no. 1, pp. 46–49, Feb. 2002, doi: 10.1097/00008480-200202000-00008.

[7] L. Solimeno and G. Pasta, “Knee and Ankle Arthroplasty in Hemophilia,” *J. Clin. Med.*, vol. 6, no. 11, p. 107, Nov. 2017, doi: 10.3390/jcm6110107.

[8] A. Aronstam, S. G. Rainsford, and M. J. Painter, “Patterns of bleeding in adolescents with severe haemophilia A.,” *BMJ*, vol. 1, no. 6161, pp. 469–470, Feb. 1979, doi: 10.1136/bmj.1.6161.469.

[9] J. Ahn, M. C. Yoo, J. Seo, M. Park, and B. O. Jeong, “Comparison of Total Ankle Arthroplasty and Ankle Arthrodesis in End-Stage Hemophilic Arthropathy,” *Foot Ankle Int.*, vol. 41, no. 8, pp. 937–944, Aug. 2020, doi: 10.1177/1071100720929002.

[10] E. C. Rodriguez-Merchan, “Management of Hemophilic Arthropathy of the Ankle,” *Cardiovasc. Hematol. Disord.-Drug Targets*, vol. 17, no. 2, Dec. 2017, doi: 10.2174/1871529X17666170908110758.

[11] R. Gualtierotti, L. P. Solimeno, and F. Peyvandi, “Hemophilic arthropathy: Current knowledge and future perspectives,” *J. Thromb. Haemost.*, vol. 19, no. 9, pp. 2112–2121, Sep. 2021, doi: 10.1111/jth.15444.

[12] A. Muda and F. Martino, “Hemophiliac Arthropathy,” in *Musculoskeletal Ultrasound in Orthopedic and Rheumatic disease in Adults*, F. Martino, E. Silvestri, and D. Orlandi, Eds., Cham: Springer International Publishing, 2022, pp. 149–153. doi: 10.1007/978-3-030-91202-4\_18.

[13] T. Wyseure, L. O. Mosnier, and A. Von Drygalski, “Advances and challenges in hemophilic arthropathy,” *Semin. Hematol.*, vol. 53, no. 1, pp. 10–19, Jan. 2016, doi: 10.1053/j.seminhematol.2015.10.005.

[14] J. K. Willett, “Imaging in trauma in limited-resource settings: A literature review,” *Afr. J. Emerg. Med.*, vol. 9, pp. S21–S27, 2019, doi: 10.1016/j.afjem.2018.07.007.

[15] W. Huysse *et al.*, “Morphometric analysis of the incisura fibularis in patients with unstable high ankle sprains,” *Skeletal Radiol.*, vol. 50, no. 6, pp. 1141–1150, Jun. 2021, doi: 10.1007/s00256-020-03649-9.

[16] P.-J. Moonen, L. Mercelina, W. Boer, and T. Fret, “Diagnostic error in the Emergency Department: follow up of patients with minor trauma in the outpatient clinic,” *Scand. J. Trauma Resusc. Emerg. Med.*, vol. 25, no. 1, p. 13, Dec. 2017, doi: 10.1186/s13049-017-0361-5.

[17] D. Plut *et al.*, “Diagnostic accuracy of haemophilia early arthropathy detection with ultrasound (HEAD-US): a comparative magnetic resonance imaging (MRI) study,” *Radiol. Oncol.*, vol. 53, no. 2, pp. 178–186, Jun. 2019, doi: 10.2478/raon-2019-0027.

[18] H. Pettersson, A. Ahlberg, and I. M. Nilsson, “A radiologic classification of hemophilic arthropathy,” *Clin. Orthop.*, no. 149, pp. 153–159, Jun. 1980.

[19] H. A. Mann, G. S. Biring, M. Z. B. Choudhury, C. A. Lee, and N. J. Goddard, “Ankle arthropathy in the haemophilic patient: a description of a novel ankle arthrodesis technique,” *Haemophilia*, vol. 15, no. 2, pp. 458–463, Mar. 2009, doi: 10.1111/j.1365-2516.2008.01651.x.

[20] M. Kotti, L. D. Duffell, A. A. Faisal, and A. H. McGregor, “Detecting knee osteoarthritis and its discriminating parameters using random forests,” *Med. Eng. Phys.*, vol. 43, pp. 19–29, May 2017, doi: 10.1016/j.medengphy.2017.02.004.

[21] R. T. Wahyuningrum, L. Anifah, I. K. E. Purnama, and M. H. Purnomo, “A novel hybrid of S2DPCA and SVM for knee osteoarthritis classification,” in *2016 IEEE International Conference on Computational Intelligence and Virtual Environments for Measurement Systems and Applications (CIVEMSA)*, Budapest, Hungary: IEEE, Jun. 2016, pp. 1–5. doi: 10.1109/CIVEMSA.2016.7524317.

[22] V. Arvind, D. A. London, C. Cirino, A. Keswani, and P. J. Cagle, “Comparison of machine learning techniques to predict unplanned readmission following total shoulder arthroplasty,” *J. Shoulder Elbow Surg.*, vol. 30, no. 2, pp. e50–e59, Feb. 2021, doi: 10.1016/j.jse.2020.05.013.

[23] A. Bohr and K. Memarzadeh, “The rise of artificial intelligence in healthcare applications,” in *Artificial Intelligence in Healthcare*, Elsevier, 2020, pp. 25–60. doi: 10.1016/B978-0-12-818438-7.00002-2.

[24] J. Prijs *et al.*, “Development and external validation of automated detection, classification, and localization of ankle fractures: inside the black box of a convolutional neural network (CNN),” *Eur. J. Trauma Emerg. Surg.*, vol. 49, no. 2, pp. 1057–1069, Apr. 2023, doi: 10.1007/s00068-022-02136-1.

[25] A. S. Panayides *et al.*, “AI in Medical Imaging Informatics: Current Challenges and Future Directions,” *IEEE J. Biomed. Health Inform.*, vol. 24, no. 7, pp. 1837–1857, Jul. 2020, doi: 10.1109/JBHI.2020.2991043.

[26] P. Lakhani and B. Sundaram, “Deep Learning at Chest Radiography: Automated Classification of Pulmonary Tuberculosis by Using Convolutional Neural Networks,” *Radiology*, vol. 284, no. 2, pp. 574–582, Aug. 2017, doi: 10.1148/radiol.2017162326.

[27] R. Mahum *et al.*, “A Novel Hybrid Approach Based on Deep CNN Features to Detect Knee Osteoarthritis,” *Sensors*, vol. 21, no. 18, p. 6189, Sep. 2021, doi: 10.3390/s21186189.

[28] S. M. Ahmed and R. J. Mstafa, “Identifying Severity Grading of Knee Osteoarthritis from X-ray Images Using an Efficient Mixture of Deep Learning and Machine Learning Models,” *Diagnostics*, vol. 12, no. 12, p. 2939, Nov. 2022, doi: 10.3390/diagnostics12122939.

[29] S. Ashkani-Esfahani *et al.*, “Detection of ankle fractures using deep learning algorithms,” *Foot Ankle Surg.*, vol. 28, no. 8, pp. 1259–1265, Dec. 2022, doi: 10.1016/j.fas.2022.05.005.

[30] A. Raza, T.-L. Phan, H.-C. Li, N. V. Hieu, T. T. Nghia, and C. T. S. Ching, “A Comparative Study of Machine Learning Classifiers for Enhancing Knee Osteoarthritis Diagnosis,” *Information*, vol. 15, no. 4, p. 183, Mar. 2024, doi: 10.3390/info15040183.

[31] H. N. Huynh, Q. T. N. Diep, M. B. Pham, A. T. Tran, and T. N. Tran, “Analysis and detection of COVID-19 cases on chest X-ray images using a novel architecture self-development deep-learning,” in *2021 IEEE International Biomedical Instrumentation and Technology Conference (IBITeC)*, Yogyakarta, Indonesia: IEEE, Oct. 2021, pp. 59–64. doi: 10.1109/IBITeC53045.2021.9649263.

[32] D. Wang, J. Mo, G. Zhou, L. Xu, and Y. Liu, “An efficient mixture of deep and machine learning models for COVID-19 diagnosis in chest X-ray images,” *PLOS ONE*, vol. 15, no. 11, p. e0242535, Nov. 2020, doi: 10.1371/journal.pone.0242535.

[33] R. Lindsey *et al.*, “Deep neural network improves fracture detection by clinicians,” *Proc. Natl. Acad. Sci.*, vol. 115, no. 45, pp. 11591–11596, Nov. 2018, doi: 10.1073/pnas.1806905115.

[34] M. Varma *et al.*, “Automated abnormality detection in lower extremity radiographs using deep learning,” *Nat. Mach. Intell.*, vol. 1, no. 12, pp. 578–583, Dec. 2019, doi: 10.1038/s42256-019-0126-0.

[35] D. Pinto Dos Santos *et al.*, “Structured report data can be used to develop deep learning algorithms: a proof of concept in ankle radiographs,” *Insights Imaging*, vol. 10, no. 1, p. 93, Dec. 2019, doi: 10.1186/s13244-019-0777-8.

[36] G. Kitamura, C. Y. Chung, and B. E. Moore, “Ankle Fracture Detection Utilizing a Convolutional Neural Network Ensemble Implemented with a Small Sample, De Novo Training, and Multiview Incorporation,” *J. Digit. Imaging*, vol. 32, no. 4, pp. 672–677, Aug. 2019, doi: 10.1007/s10278-018-0167-7.

[37] J.-H. Kim, Y.-C. Mo, S.-M. Choi, Y. Hyun, and J. W. Lee, “Detecting Ankle Fractures in Plain Radiographs Using Deep Learning with Accurately Labeled Datasets Aided by Computed Tomography: A Retrospective Observational Study,” *Appl. Sci.*, vol. 11, no. 19, p. 8791, Sep. 2021, doi: 10.3390/app11198791.

[38] A. Fernandez, S. Garcia, F. Herrera, and N. V. Chawla, “SMOTE for Learning from Imbalanced Data: Progress and Challenges, Marking the 15-year Anniversary,” *J. Artif. Intell. Res.*, vol. 61, pp. 863–905, Apr. 2018, doi: 10.1613/jair.1.11192.

[39] L. Perez and J. Wang, “The Effectiveness of Data Augmentation in Image Classification using Deep Learning,” 2017, doi: 10.48550/ARXIV.1712.04621.

[40] I. T. Jolliffe and J. Cadima, “Principal component analysis: a review and recent developments,” *Philos. Trans. R. Soc. Math. Phys. Eng. Sci.*, vol. 374, no. 2065, p. 20150202, Apr. 2016, doi: 10.1098/rsta.2015.0202.

[41] O. Chapelle, V. Vapnik, O. Bousquet, and S. Mukherjee, “Choosing Multiple Parameters for Support Vector Machines,” *Mach. Learn.*, vol. 46, no. 1/3, pp. 131–159, 2002, doi: 10.1023/A:1012450327387.

[42] S. Zhang, X. Li, M. Zong, X. Zhu, and R. Wang, “Efficient kNN Classification With Different Numbers of Nearest Neighbors,” *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 29, no. 5, pp. 1774–1785, May 2018, doi: 10.1109/TNNLS.2017.2673241.

[43] T. Chen and C. Guestrin, “XGBoost: A Scalable Tree Boosting System,” in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, San Francisco California USA: ACM, Aug. 2016, pp. 785–794. doi: 10.1145/2939672.2939785.