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

by Aquib Raza1, Mohammad Rameez4, Hung-Chung Li3,

1 College of Electrical Engineering and Computer Science, National Taiwan University, 10673, Taipei City, Taiwan, Institute of Information Science, Artificial Intelligence of Things (AIoT) Taiwan International Graduate Program (TIGP), Academia Sinica, Nankang, Taipei 11529, Taiwan (d13949005@ntu.edu.tw)

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

4 Department of Computer Science, University of York, Heslington, York YO10 5DD, United Kingdom. (Mr1606@york.ac.uk)

**Abstract**

Hemophilia, characterized by a deficiency in coagulation factors VIII and IX, manifests predominantly in males and leads to severe hemophilic arthropathy, particularly affecting joints such as ankles, knees, and elbows due to recurrent bleeding. Traditional diagnostic approaches rely heavily on radiography, yet advancements in medical imaging have highlighted the potential of integrating Artificial Intelligence (AI) to enhance diagnostic accuracy and efficiency. This study explores the efficacy of combining deep learning (DL) and machine learning (ML) algorithms for the automated classification of X-ray images into 'Healthy' and 'Ill' categories, specifically addressing hemophilic arthropathy on ankle images.

Utilizing the Inception V3 architecture for deep feature extraction and SVM, KNN, and XGBoost for classification, our research meticulously evaluates the performance of these models on a curated dataset. The SVM classifier demonstrated exceptional accuracy (97.78%) and an Area Under the Curve (AUC) of 0.99 on testing data, proving its superior discriminative ability and potential for clinical utility. The KNN and XGBoost classifiers also showed commendable performance with accuracies of 95.56%, though with slightly lower AUC values, indicating robust but slightly less precise capabilities.

Our findings suggest that AI can significantly improve the diagnostic process for hemophilic arthropathy by reducing the time to diagnosis and potentially increasing the accuracy of assessments, especially in resource-limited settings. Future research should validate and scale these models across more extensive datasets and diverse clinical environments, explore their applicability to other joint diseases, and enhance computational methods to include various imaging modalities. Integrating AI into clinical practice promises a new era of precision medicine, offering profound benefits for patient care and treatment outcomes in hemophilia and beyond.

**Keywords:** Deep learning, Ankle, knee, X-ray Classification imaging, Artificial Intelligence, Hemophilia, Hemophilic arthropathy, Machine Learning.

**1. Introduction**

"Hemophilia" is a bleeding disorder caused by a lack of coagulation factors. Because of a lack of coagulation factors VIII and IX, the most frequent kinds were Hemophilia A and B, affecting most men [1]. The X-linked, recessive, inherited blood disorders are hemophilia A along with hemophilia B [2]. Hemophilia A is the most common, accounting for 80-85% of all cases [3]. As reported in the Annual Global Report 2020 by the World Federation of Hemophilia (WFH), the estimated number of individuals with hemophilia globally is 815,100; among that, only 347,026 are diagnosed, with 276,900 instances having severe hemophilia [4]. It is reported that one in 5,000 guys worldwide has hemophilia A and one in 30,000 males has hemophilia B [5]. Over 90% of individuals with severe hemophilia have joint disease, which is the cause of the most morbidity and expense among hemophilia patients [6]. Intra-articular bleeding represents the most frequent sign of hemophilia, occurring in 80% of cases [7]. Degenerative conditions arise due to repeated hemarthrosis, initially causing aches and pains. Inflammation, retraction, and synovium hypertrophy are abnormalities of the joint capsule. Subchondral cysts, osteophytes, and chondral erosion comprise most of the articular surface involvement. All of these changes result in reduced dorsal and plantar flexion, misalignment of the talocrural articular surface, and a permanent varus or valgus equinus foot in the ankle and foot. the ankle is considered to be the third most prevalent location of hemarthrosis after the knee and elbow, which have a prevalence of 24% and 22%, respectively [8]. These joints are more susceptible because of their large range of motion and the strain physical activity places on them [9]. Ankle arthropathy can be treated in a variety of ways, from conservative measures like physical therapy and orthoses to more drastic measures like arthrodesis as well as total ankle replacement surgery [10]. Serious hemophilia most commonly manifests clinically as recurrent joint bleeding (hemarthrosis). Even mild cases of hemarthrosis can progress to Hemophilic Arthropathy (HA), which is a debilitating disorder marked by joint alterations, excruciating pain, diminished quality of life, and the ultimate need for joint replacement if left neglected [11]. From the earliest years of life, hemorrhagic instances can happen, especially when a kid learns to walk (which highlights the role of the body's motion in triggering the bleeding) [12]. It is a form of arthritis characterized by cartilage, bone loss, and synovial enlargement that arises from bleeding in the joints. Even with clotting factor replacement, arthropathy can still occur and is particularly crippling in the elderly population. Hemoglobin release along with iron deposition in the joint cause pathobiological tissue changes, although the exact sequence of events and molecular processes causing joint degradation are not fully known. There aren't many choices for treatment outside replacing clotting factors. Enhancements in hemophilia therapy require improvements in diagnostic instruments [13]. Even with the growing accessibility of contemporary imaging modalities such as MRI (magnetic resonance imaging) and computed tomography (CT), traditional radiography remains essential for the first evaluation of these individuals. Because radiographs are inexpensive, quick, easy to get, and low radiation, they have been the standard for first assessment—especially in settings with limited resources [14]. Despite these obvious advantages, approximately 23% of ankle fractures are overlooked during the first radiographic evaluation for a variety of reasons, such as insufficient training or experience, outdated imaging technology, anatomical variations, or superimposition of bony structures in two-dimensional views on radiographs [15, 16]. Soft-tissue swelling characterizes hemophilic arthropathy in its early stages radiologically, while the later stages are characterized by juxta-articular osteoporosis, lesions of the bone with an overgrowth of the epiphysis, along with cartilage damage, which manifests radiologically as narrowing of the joint space and subsequently as total loss of cartilage space, along with the formation of subchondral bones cysts as well as, bones erosions, and irregularities in the joint profile [17]. The popular Petterson scoring method is still the most utilized one for evaluating the degree of joint injury using plain radiography. In 1980, the Pettersson score was described. The Pettersson scoring is a cumulative scale. Every anomaly is assigned a score ranging from zero to one or as 2, with variations resulting from various phases of the illness being included. An individual joint might have a maximum score of 13, which denotes a completely ruined joint [18]. Joint examination heavily relies on radiological imaging, particularly X-rays. Anteroposterior along with lateral view standard x-rays of the ankle can be used to assess several changes, including osteoporosis, epiphysis enlargement, irregular subchondral surface, narrowing of the joint space, subchondral cyst, erosion of the joint margins, incongruence between the surfaces of the joints, and both posterior and anterior osteophytes. It can also detect the overall deformity and flattery of the talus dome caused by avascular necrosis, as well as the valgus tilt brought on by medial tibial overgrowth [19]. However, limited access to radiologists, particularly in rural areas, leads to delays in diagnosis. Implementing automated tools for faster analysis could significantly enhance the efficiency of diagnosing and tracking the progression of hemophilic arthropathy.

Consequently, several computer-aided diagnosis (CAD)-based medical imaging techniques have been put forth in the literature to identify and assess knee OA [20, 21]. AI or its subsets, ML, or DL, are rapidly being investigated for uses in medicine and orthopedic surgery [22-24]. In a nutshell, AI and its subcategories include the implementation of machinery to replicate human intellect. Algorithms or models that learn and interpret complex interactions from data sets can be created. These models may be used for various reasons, such as automating radiography image processing, predicting surgical results, or forecasting athlete injuries.

In machine learning before the deep learning period, tasks were defined by the input data, and features that were manually created to solve invariances and minimize dimensionality were frequently used. Principal Component Analysis (PCA), the K-means algorithm, and other imaging approaches were used in conjunction with various transformations, including Fourier, Cosine, and Wavelet. To find the right features for later learnable decision algorithms, feature engineering was essential and required a lot of work. Classification techniques that were often used were support vector machines as well as random forests, including probabilistic boosting trees providing an alternative [25].

DL, a subset of machine learning, employs deep convolutional neural networks (DCNNs) for interpreting clinical images. Unlike traditional methods requiring manual feature assessment, DCNNs use native image inputs such as pixels, Hounsfield differences, distances, brightness, and contrast for image evaluation [26]. Larger training datasets enhance DCNN accuracy, and pre-trained models, optimized for object recognition, can be utilized through transfer learning. This approach applied to radiographs, enhances DL algorithm performance in accurately and rapidly detecting musculoskeletal pathologies.

Recently, there has been a surge in the application of DL or ML approaches in medical imaging to address classification [27-30] identification [31-33] and other related problems without necessitating the knowledge of radiologists.

More precisely, DL-based detection algorithms have been developed and effectively used to quantify the severity of knee osteoarthritis [34]. Furthermore, they perform astonishingly effectively in analyzing X-ray imaging in the biomedical field since they do not require human feature engineering. This occurs implicitly through the training stage by improving their internal parameters to match the relevant data. In contrast, all typical ML techniques require the provided data to be changed using a specific feature engineering field and learning process to yield the intended results. DL methods frequently need significantly more processing power and resources than traditional ML techniques. Additionally, if little data is utilized, this leads to an overfitting. Moreover, several DL forms, such as Inception, Resnet, and Xception, or Transfer learning-based CNN, perform extraordinarily in the field of computer vision—even surpassing human performance [32]. In previous studies like Santos et al.'s the data from structured reports of X-ray diagrams of ankle fractures were utilized. There were 157 patients in the data collection; 129 had fractures, and 28 did not. The model's accuracy was 77%, and its AUC was 0.85 [35]. Kitamura et al, conducted an in-depth study on a bigger dataset of 596 pictures (ankle as well as without fractures were evenly distributed) using five different CNN models; the accuracy reached 81% [36]. Kim et al with deep learning utilized Inception V3 to detect ankle fractures in the computed tomography (CT) and X-ray of 1226 individuals. The dataset consisted of 1040 fracture patients and 186 normal instances, which were handled via data augmentation and under-sampling. The model attained an accuracy of up to 83% for AP and 90% for lateral X-rays, yielding an AUC of 0.91 as well as 0.95, accordingly [37]. Simarlry Ashkani-Esfahani et al using Radiographs from 1,050 individuals with ankle fractures and 1,050 with healthy ankles were used in research to train DCNNs, especially Inception V3 and ResNet-50 models, using the Danis-Weber classification. Anteroposterior and lateral radiographs were used in the study to train DCNNs, and their performance was compared between the two views. Inception V3 achieved up to 98.7% sensitivity as well as 98.6% specificity, demonstrating how the three-view strategy greatly exceeded the single-view on accuracy, F-score, and AUC.

From the research reviewed, it is evident that several studies have proposed effective methods for detecting knee osteoarthritis (KOA) and ankle fractures in X-ray images or using custom datasets. Many researchers have utilized feature extraction and classification techniques, as well as morphological processing of images, to achieve favorable results. DL and ML methods for diagnosing various bone disorders have demonstrated impressive results, particularly in the binary-class classification of KOA. However, they have not been applied to categorizing hemophilic arthropathy (HA) in ankles using Pettersson scores. Therefore, it is challenging to develop an efficient tool or technique for categorizing HA ankles. To determine the efficacy of classification, a combined learning approach is essential. This research provides ways that use DL for feature extraction along with Dimension reduction techniques like PCA and for training 3 ML algorithms were used to determine a subject's class designation from an HA hemophilic x-ray picture based on Petterson scoring. By focusing on binary classification, this method differentiates between healthy individuals and those with Hemophilic Arthropathy (HA) in the ankle. This streamlined approach enhances detection efficiency, accuracy, and cost-effectiveness, crucial for early treatment initiation. Simplifying the diagnostic process to distinguish only between healthy and HA-affected ankles not only accelerates diagnosis but also reduces costs, aiding in early disease management and improving patient outcomes.

1. **Materials and Methods:**

This section details our proposed methodology our methodology for classifying X-ray images into "healthy" or "ill" categories integrates deep learning with traditional machine learning techniques, structured and depicted in a systematic flowchart, as shown in Figure 1. The process begins with comprehensive preprocessing steps: resizing, normalization, and data augmentation of the X-ray images. These steps standardize the input data, making it suitable for complex model processing and ensuring consistency across the dataset. Once preprocessing is complete, the dataset is split into training and testing subsets. To address potential class imbalances and ensure equitable representation of both categories, we apply the Synthetic Minority Over-sampling Technique (SMOTE) to the training set. This approach helps to mitigate any bias toward the majority class and enhances the model’s ability to generalize across conditions.

The balanced training data is then input into a customized InceptionV3 neural network, which has been specifically modified by adding layers designed for effective binary classification. The study focus on reducing overfitting through the implementation of early stopping mechanisms during the model training phase. Following training, the model is saved for subsequent use.

Feature extraction is performed next, utilizing the trained model to extract valuable features from both the training and testing datasets. These features are then subjected to dimensionality reduction via PCA, simplifying the complex high-dimensional data into a more manageable form without significant loss of information. The reduced features are classified using advanced machine learning classifiers.

Our methodology’s effectiveness is rigorously evaluated using detailed classification reports and confusion matrices from both the training and testing datasets. This structured and iterative process not only facilitates thorough assessment but also boosts the model's accuracy and reliability, making our system a robust and efficient tool for medical X-ray image classification.

****

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

* 1. **Dataset preparation:**

In the research, X-ray images from patients were obtained at a medical hospital (), focusing on individuals diagnosed with hemophilic arthropathy and comparing them with healthy controls. The study was conducted under ethical standards approved by the Institutional Review Board (IRB), which waived the requirement for informed and written consent (IRB Approval No. CMUH113-REC1-002). This facilitated the evaluation of X-ray films using Pettersson scores, a methodological approach that enhances our understanding of disease severity and progression.

The dataset comprised X-ray images showing both anterior-posterior (AP) and lateral views of the ankles. These images were carefully analyzed by a team of senior medical professionals, including an orthopaedist with a specialization in foot and ankle disorders. To ensure objectivity and accuracy, each X-ray was manually labeled based on the diagnosis, resulting in two groups: 111 individuals classified as having arthropathy ("ill") and an equal number identified as healthy. To protect patient privacy and adhere to ethical guidelines, no personal identifying information such as age, gender, or medical history was retained in our dataset. Our data collection and preparation processes are designed to uphold the highest standards of confidentiality and integrity of patient data. This rigorous approach ensures that our research is conducted ethically while providing valuable insights into medical conditions through the analysis of X-ray images.

**2.2 Data pre-processing:**

X-ray image dataset for a classification model aimed at distinguishing "healthy" from "ill" categories, using a blend of deep learning and machine learning techniques. The dataset, comprising a total of 222 X-ray images, was sourced from two distinctly labeled directories. Each image underwent a series of preprocessing steps including loading, resizing to a uniform dimension of 256256 pixels, and normalization of pixel values to a range between 0 and 1. These steps were critical in ensuring consistency across the dataset and enhancing neural network performance by promoting stable and quick convergence during training.

To organize the dataset, the processed images and their corresponding labels were stored in arrays, which facilitated straightforward handling during the model training and testing phases. The study implemented stratified sampling to divide the dataset into training (80%) with 177 images and testing (20%) 45 image sets, ensuring that each subset accurately represented the overall distribution of the dataset.

Addressing class imbalance, a significant concern in medical image analysis was tackled using the Synthetic Minority Over-sampling Technique (SMOTE) [38]. This method augmented the underrepresented class in the training dataset by creating statistically unique yet similar entries, helping to balance the dataset and prevent model bias towards the majority class.

Further enhancing the dataset's robustness, a variety of data augmentation techniques were applied using "ImageDataGenerator" in the Keras module [39]. Techniques such as rotations (up to 20 degrees), width and height shifts (up to 20%), shear transformations, zooming (up to 20%), and horizontal flipping were used. These augmentations dynamically introduced realistic variations during the training process, significantly enriching the training set with a diverse array of image conditions. This not only helped in preventing overfitting but also significantly improved the model's ability to generalize from training data to unseen real-time data.

Following the application of SMOTE and data augmentation techniques, the number of images in the training set expanded substantially to 540 images. This increase was crucial for developing a robust and accurate classification model capable of handling the diverse imaging scenarios encountered in clinical settings. Through this structured and iterative approach, our study ensured the comprehensive preparation of X-ray images, ultimately enhancing the model's predictive accuracy and robustness in a clinical environment.

**Feature Extraction:**

InceptionV3 Architecture: An InceptionV3 model was selected, renowned for its high performance in image classification tasks due to its intricate architecture that involves multiple convolutional stages and inception modules demonstrating a high success rate in medical image analysis [29, 37]. This model was originally developed for the ImageNet competition, which involves classifying images into one thousand categories. For our application, the model was adapted to the binary classification task by removing the top layer (include\_top=False) which is specifically designed for the 1000-class output of ImageNet. This allows the model to be repurposed for our binary classification needs.

Initialization and Input Configuration: The model was initialized with weights pre-trained on the ImageNet dataset, allowing us to utilize a broad and diverse base of visual features learned from a large dataset of generic images. This pre-training aids significantly in feature extraction, especially in medical imaging, where certain visual patterns may be subtle yet diagnostically significant. The input configuration was set to accept images of 256×256 pixels with three color channels (RGB), which is a common standard for processing medical images and allows the network to operate efficiently while retaining crucial image details.

**Custom Layer Integration:**

Global Average Pooling 2D: Positioned immediately after the base InceptionV3 outputs, this layer simplifies the output by reducing each of the feature maps to a single average value, effectively decreasing the feature dimensionality and reducing the model's susceptibility to overfitting.

Dense Layer with Advanced Activation: A dense layer of 756 units is included to enable the learning of complex and nuanced patterns from the reduced feature set. The layer uses LeakyReLU activation, which allows a small, positive gradient when the activation is otherwise zero, preventing the "dying ReLU" problem and ensuring continuous learning throughout the network.

Dropout for Regularization: A dropout layer with a 50% rate randomly disables half of the neuron connections in the dense layer during each training pass. This stochastic element introduces noise into the training process, forcing the network to learn more robust features that generalize better to unseen data, thus preventing overfitting.

Sigmoid Output Layer: The architecture is finalized with a sigmoid activation function in the output layer, which provides a probability score indicating the likelihood of the X-ray depicting pathological features categorized as "ill".

**Model Compilation and Advanced Training Techniques:**

Model Compilation: The network is compiled using the Adam optimizer, noted for its adaptive learning rate capabilities, which helps in fine-tuning the model more effectively. Binary Cross entropy is chosen as the loss function because it is particularly suited for binary outputs where the prediction is a probability value.

**Training Strategy:**

Early Stopping Mechanism: Implemented to monitor the validation loss during training, this callback function halts training if no improvement is observed in the loss for five consecutive epochs. It also restores the weights from the epoch with the lowest validation loss, optimizing the model's performance without overfitting.

Training Execution: Training is conducted over 50 epochs with a batch size of 32. The model is trained on a dataset enriched with both original and augmented images to ensure the model is robust and can generalize well across varied imaging conditions encountered in clinical settings. Figure 2 demonstrates the structure of the model that is utilized for feature extraction.

|  |
| --- |
|  |

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

**2.3 Dimensionality Reduction with PCA:**

Following feature extraction, PCA was adopted to reduce the dimensionality of our features:

Component Selection: The PCA model is configured to retain a maximum of 500 components, or fewer if the feature set is smaller, thus ensuring we stay within the practical limits of our feature space. Feature extraction is initially handled by the InceptionV3 architecture, which is adept at identifying complex patterns within the images. This process, while thorough, tends to produce a high volume of features, many of which may be closely interrelated or superfluous for the task of binary classification. To address this issue, PCA is employed post-feature extraction. PCA streamlines the feature set by condensing the high-dimensional data into a more manageable form, focusing on retaining features that account for the most variance. This reduction not only simplifies the dataset but also alleviates the computational demands on subsequent classifiers, potentially enhancing their performance and generalization capabilities.

Following the application of PCA, the transformed and simplified dataset is then fed into machine learning classifiers such as SVM, KNN, and XGBoost. These classifiers benefit from the reduced complexity of the data, allowing them to operate more efficiently and effectively. By decreasing the dimensionality, PCA facilitates a more focused analysis by the classifiers, improving their ability to discern and react to meaningful patterns without being overwhelmed by noise and redundancy. Thus, PCA is a critical intermediary, optimizing the extracted features to ensure the classifiers can perform optimally within the pipeline.

PCA Transformation: This technique simplifies the feature set by focusing on the principal components that account for the majority of the variance within the data. This reduction not only curtails computational complexity but also aids in mitigating overfitting by distilling the data into its most informative aspects [40].

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

Three prominent machine-learning algorithms were employed in the study to forecast results. The models that were selected were Support Vector Machine (SVM), K Nearest Neighbors (KNN), and Extreme Gradient Boosting (XGBoost). These models were chosen based on their wide range of functionality and reliable track records in Image classification. The tuning process is essential in distinguishing the efficiency of these three models and affecting how they should be compared.

Hyperparameter adjustment improved the models' performance and reduced bias and overfitting. A predetermined range of hyperparameters, ranging from the bare minimum required for each model to the optimal value discovered in the literature, was chosen for optimization [41-43]. First, the SVM employs a sigmoid kernel suitable for managing the non-linear relationships often present in intricate medical datasets. The parameter for regularization C is set to 1, balancing classification accuracy against model complexity. The gamma parameter is automatically adjusted based on the number of features, which helps finely tune the decision boundary without overly emphasizing individual samples.

Second, the KNN classifier is configured with 11 neighbors. This choice ensures that the classification decisions are based on a sufficient yet manageable number of nearest samples, fostering computational efficiency and decision reliability. All neighbors have equal voting power, marked by the 'uniform' weight set, which prevents biases toward more densely populated regions of the data space, which is crucial in diverse scenarios.

Lastly, the XGBoost classifier, renowned for its performance in binary classification tasks, is optimized for medical diagnostics. It operates under a binary logistic regression framework with 200 estimators, providing a robust ensemble method capable of capturing complex patterns and interactions in the data. A max depth of 7 and a learning rate of 0.01 are chosen to refine the learning process incrementally, enhancing the model’s ability to generalize well on new, unseen data. The subsample and colsample\_bytree parameters are set to 0.9 and 0.8, respectively, ensuring that each tree in the model is built from a subset of data and features, thus reducing overfitting risks. Additionally, a gamma value of 2 encourages the model to make only those splits that offer substantial improvement, a conservative approach that is valuable in medical applications where predictive stability is crucial. Python 3.11.4 was used to train and assess these models, and frameworks designed specifically for machine learning tasks were used. By ensuring that every model was adjusted to its ideal parameters, this method made it possible to compare the various methods for forecasting outcomes fairly and efficiently.

* 1. **Model Evaluation:**

The process of evaluating trained models to determine how well they predict unknown data is known as model evaluation, and it is a crucial stage in machine learning. The assessment of the SVM, KNN, and XGBoost classifiers involves a variety of metrics and visual aids, providing a comprehensive understanding of each machine learning model's predictive ability. The confusion matrix was utilized during the assessment phase to enhance the evaluation of the model's performance. This matrix delineates true positives (TP), false positives (FP), and false negatives (FN), true negatives (TN), providing a clear visual representation of the model's predictive capabilities. This approach is particularly valuable for identifying specific biases or types of errors that a mere accuracy metric might overlook.

The classification report extends this analysis by incorporating essential metrics for each class, such as precision (defined by Equation (1)), recall (defined by Equation (2)), the F1 score (defined by Equation (3)), and the overall accuracy score (defined by Equation (4)). These metrics provide more detailed insight into the model's effectiveness across different classifications, offering a granular view of its predictive accuracy and reliability. This detailed scoring in the report helps in understanding the nuanced performance of the model in various scenarios.

|  |  |
| --- | --- |
|  | (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 was utilized to create and assess machine learning models for classification using Python 3.11.4. A high-performance configuration was employed for these computational studies, featuring an Intel® CoreTM i7-10700 CPU running at 2.90 GHz and 32 GB of RAM. This setup effectively handled the dataset and computing needs. Additionally, an Nvidia GeForce RTX 3070 graphics card was utilized for graphic processing activities such as image processing and model training. Sufficient processing power was available with this hardware configuration to support intensive model training, hyperparameter tweaking, and validation.

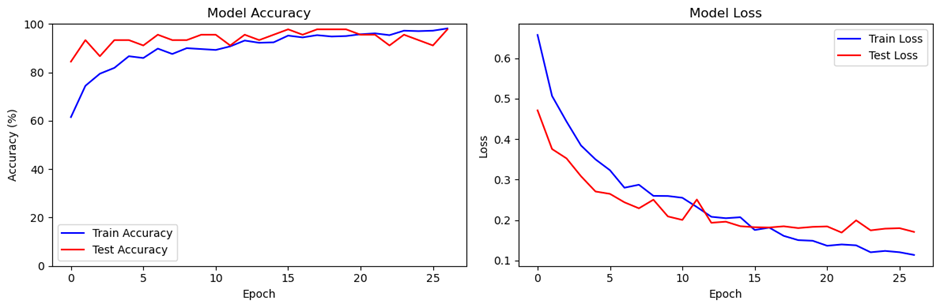
1. **Results and Discussion:**

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

**3.1** **Accuracy and Loss Over Epochs:**

This section compiles the experimental findings of the proposed Inception V3 architectures, trained from scratch. In the Model Accuracy Graph, the vertical axis quantifies the model's accuracy as a percentage, ranging from 0 to 100%, while the horizontal axis tracks the training progress over 25 epochs. The blue line, representing the model's performance on the training dataset, demonstrates a significant increase in accuracy from the onset, swiftly ascending from around 60% to beyond 80% in the initial few epochs. This trajectory levels off near the 95% mark, suggesting that the model is reaching its performance capacity on the training data. The red line, which traces the model's accuracy on the unseen test dataset, closely aligns with the training accuracy, albeit with expected variations. The two lines' small gap and parallel trend suggest effective generalization with no significant overfitting or underfitting

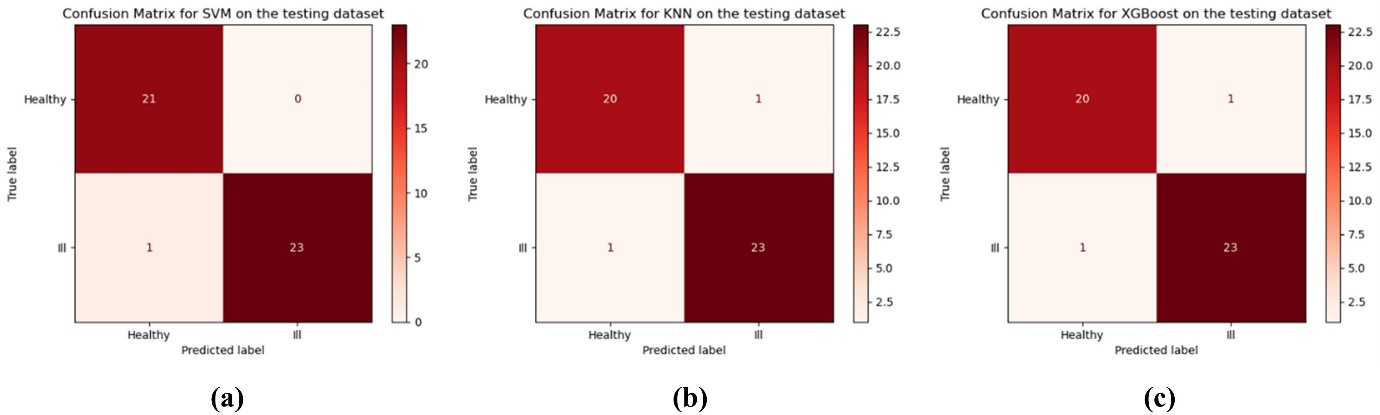
In Figure 3, the model loss graph complements the accuracy graph by illustrating the model's loss, a representation of prediction error, where lower values are preferable. On this graph, the epoch count again serves as the x-axis. The blue line depicts the loss on the training dataset, starting at a higher point and sharply decreasing, indicating rapid learning before plateauing as the epochs progress. The red line, mapping the loss on the test dataset, follows the decline of the training loss but consistently stays slightly higher throughout the training duration. The close and parallel paths of these lines through the graph reinforce that the model maintains a balanced performance on both seen and unseen data, showing no signs of overfitting and confirming its ability to generalize well.



**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.

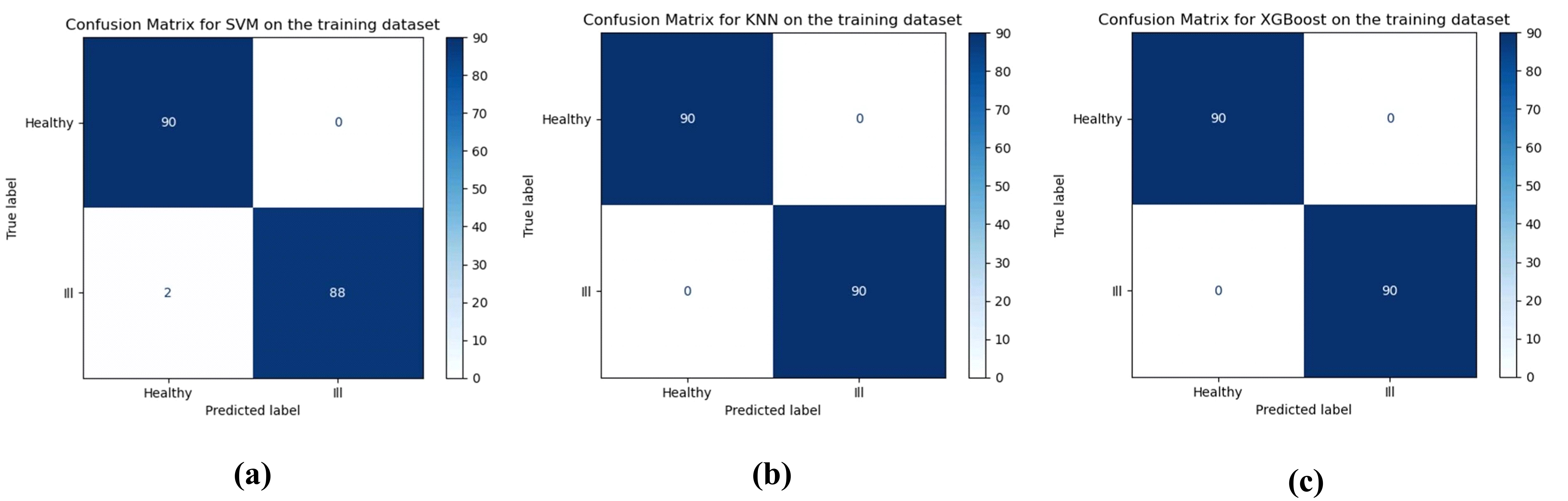
**3.2 Confusion Matrices and Classification Reports:**

The confusion matrix for the testing dataset reveals a high number of correct predictions; The confusion matrices represent the results of testing three different classifiers: SVM, KNN, and XGBoost on a dataset with two classes labeled "Healthy" and "Ill". All three models show a similar pattern of predictions with very high accuracy. The SVM model has the highest number of correct predictions for "Healthy" with 21 instances. KNN and XGBoost have one less correct prediction for "Healthy" but make the same number of correct predictions for "Ill" and the same number of incorrect predictions overall. The confusion matrix for each of the three different classifiers on the test data set is shown in Figure 4, which also shows small variations in performance and accuracy.



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

The SVM, KNN, and XGBoost classifiers were assessed on a training dataset, with their performance captured in confusion matrices. The SVM classifier accurately identified all 'Healthy' cases but misclassified two 'Ill' cases. Conversely, both KNN and XGBoost achieved 100% accuracy on the training data, identifying all cases correctly. While this may suggest excellent performance, it also raises the possibility of overfitting. The absence of any error in KNN and XGBoost could indicate they need to be more closely fitted to the training data, which might translate poorly to real-world predictive accuracy. Hence, validation with an unseen test dataset is essential to ensure the models generalize well beyond the training data. The slight imperfection in the SVM's performance could, in fact, be a sign of better generalization potential. The confusion matrix for each of the three different classifiers on the train data set is shown in Figure 5.



**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 specific hyperparameters to optimize their performance for binary classification. The SVM classifier was employed with a sigmoid kernel, C parameter set to 1, and gamma defined as 'auto', ensuring probability estimates are available, an important feature for certain applications like healthcare, with a consistent random state for reproducibility. The KNN classifier was tailored with 11 neighbors and uniform weights, utilizing the 'auto' algorithm to select the most appropriate computational approach. Meanwhile, the XGBoost model was set up with 200 estimators, a learning rate of 0.01, a max depth of 7, and adjusted subsample and colsample\_bytree parameters, along with a specified gamma value, ensuring a robust and tailored fit to the training data.

The SVM classifier's performance on the testing dataset was exceptional, demonstrating precision between 0.95 and 1.00, and an overall accuracy of 97.78%. With perfect recall for 'Healthy' instances and a high recall for 'Ill' instances, the SVM proved highly reliable, potentially serving as an effective tool in preventive medicine. This model's capability to generalize was further indicated by its impressive 98.89% accuracy on training data, suggesting robustness as a diagnostic tool.

Conversely, while the KNN and XGBoost models achieved flawless classification on the training data, they exhibited a slight decrease in performance metrics on the testing dataset, with average sensitivity and specificity scores around 95.83% and 95.24%, respectively, and an accuracy of 95.56%. The contrast between the training perfection and the reduced testing efficacy points to a possible overfitting issue, a common concern that could limit the models' predictive accuracy on new data. Compared to its testing performance, the SVM's slightly reduced sensitivity during training may imply a more generalized approach that successfully avoids overfitting, thus ensuring better performance when applied to unseen datasets.

The slight edge of the SVM classifier in testing conditions suggests it may be better suited for deployment in real-world applications, where consistent accuracy across diverse datasets is vital. Its marginally superior performance can be attributed to the SVM's ability to construct an optimized hyperplane that effectively differentiates between the classes in a binary classification task. Table 1 lists the classification reports for the SVM, KNN, and XGBoost classifiers and clearly compares their performance on both the 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 |

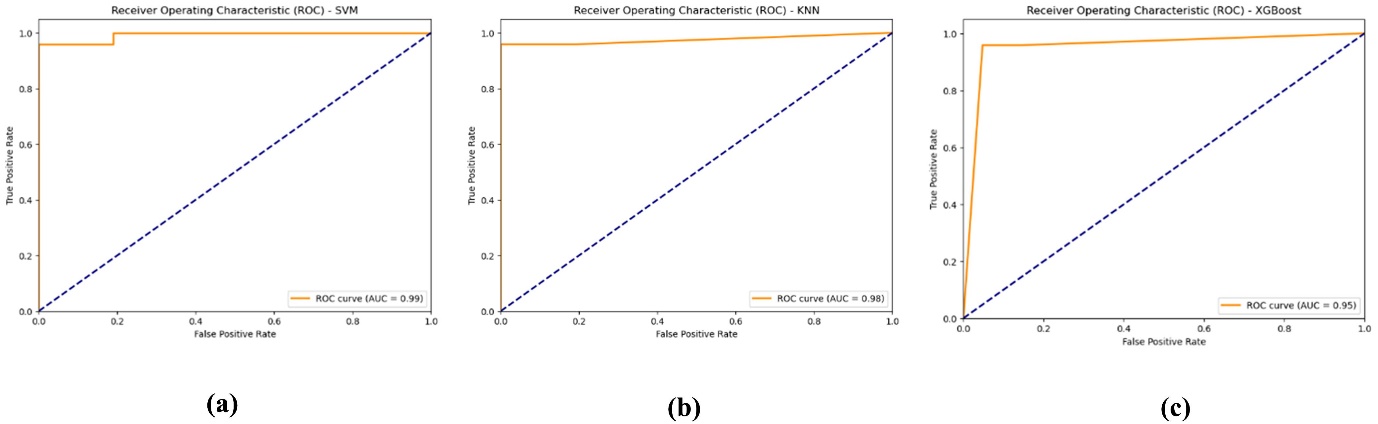
**3.3 ROC Curve, and AUC Score Report of the Machine learning classifier:**

The performance of three classifiers—XGBoost, KNN, and SVM—was rigorously assessed using Receiver Operating Characteristic (ROC) curves. The ROC curve is a crucial tool that illustrates a classifier's ability to distinguish between classes at various threshold levels, plotting the True Positive Rate (TPR) against the False Positive Rate (FPR). The AUC is a summary metric that quantifies the overall ability of the model to discriminate between positive and negative classes across all thresholds. Higher AUC values are indicative of better model performance, with a value of 1.0 representing perfect classification and 0.5 denoting no discriminative ability.

The AUC for the XGBoost classifier was computed at 0.95, suggesting a strong discriminative ability where the model correctly identifies the 'Healthy' and 'Ill' classes with a 95% probability. This performance is considered excellent and reflects a high level of accuracy in the model's predictive capability.

KNN, a model based on proximity in feature space, achieved an AUC of 0.98. This near-perfect score indicates a 98% probability that the KNN classifier will correctly classify a randomly chosen positive instance higher than a randomly chosen negative instance. Such a high AUC value denotes a very good measure of separability, demonstrating that KNN is highly effective in distinguishing between the two classes.

Outperforming the other two classifiers, the SVM reached an AUC of 0.99, placing it at the forefront of classification capability in this study. With a 99% probability of correctly discriminating between classes, the SVM classifier exhibits superior performance, denoting an outstanding level of sensitivity and specificity. This implies that the SVM is exceptionally accurate in classifying individuals correctly, making it a robust choice for applications where the precision of the classification is critical, such as medical diagnostics where the cost of misdiagnosis is high.

****

**Figure 6** **:** ROC and AUC curves for different classifiers (a) SVM, (b) KNN, (c) XGB.

The ROC curves for the XGBoost, KNN, and SVM classifiers present a compelling narrative of their robustness and acumen in binary classification.

1. **Conclusions**

In conclusion, our investigation has made considerable headway in applying deep learning and machine learning algorithms to automate the diagnostic process of hemophilic arthropathy using X-ray imaging. The study meticulously executed the integration of deep learning, Employing the sophisticated Inception V3 architecture for robust feature extraction with dimension reduction PCA with machine learning classifiers like SVM, KNN, and XGBoost for binary classification of X-ray images into 'Healthy' and 'Ill' categories.

The SVM classifier has particularly distinguished itself with a stellar performance on testing data, boasting an accuracy of 97.78% and an AUC of 0.99, indicating exceptional discriminative power and establishing its potential for clinical utility. While slightly trailing, the KNN and XGBoost classifiers still showcase admirable accuracy at 95.56% on testing data, substantiating their utility as reliable classification tools in their own right. KNN and XGBoost also demonstrated strong discriminative abilities with high AUC values. Yet, the disparity between their training and testing performances suggests that overfitting could be a concern that warrants further investigation.

These computational models' promise is significant; however, transitioning from a research setting into clinical practice involves overcoming certain hurdles. These include expanding and diversifying the datasets, validating models across multiple clinical sites, and integrating AI tools seamlessly into the diagnostic workflow. The study also underscored the necessity of algorithmic accuracy and the interpretability of the models, which is critical for clinician acceptance and ethical AI practices.

While the initial results are promising, the true measure of success will come from the models' performance in real-world clinical settings and their impact on patient outcomes. Future research initiatives must pivot towards scaling up these models and diversifying the datasets to encompass a more extensive spectrum of joint disorders. Enhancing computational methodologies to integrate a broader array of diagnostic imaging tools, including ultrasound, CT, and MRI scans, will be pivotal in crafting a more comprehensive diagnostic framework.

It is essential to rigorously validate these models across a broader array of clinical settings, fine-tuning them in alignment with the insights drawn from clinical practice. This will ensure the models are not only technically sound but also practically relevant and ready for real-world application.

Exploration beyond the confines of hemophilic arthropathy to include a broader spectrum of joint diseases is a crucial step toward realizing the full potential of AI in medical diagnostics. As machine learning and deep learning continue to make significant strides within the domain of medical imaging, they are set to revolutionize clinical diagnostics, promising substantial improvements in patient care and treatment outcomes. The cusp of this new era in clinical diagnostics is upon us, driven by AI's potential to enhance the precision and effectiveness of patient care.

1. **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.