Journal of Technology Research (JTR)
Volume 3, Special Issue, (2025), pp 481-493 ISSN 3005-639X
The 7th Conference of Engineering Sciences and Technology (CEST-2025)

Automated Detection of Bone Fractures in X-ray Images Using Deep Learning and Ensemble Learning

Basma Balam*[1], Atef Eldenfria[1]

[1]Department of Computer Science, Faculty of Information Technology, Misurata University, Misurata, Libya.

*Corresponding author email: basmafarajbalam@gmail.com

Received: 18-09-2025 | Accepted: 27-11-2025 | Available online: 25-12-2025 | DOI:10.26629/jtr.2025.45

ABSTRACT

Bone fractures are among the most common injuries worldwide and pose a significant challenge for accurate diagnosis, with error rates reaching up to 10%, potentially leading to health complications and delayed treatment. This study aims to develop and evaluate deep learning models for enhancing the accuracy and efficiency of fracture detection, utilizing three primary frameworks: conventional Convolutional Neural Networks (CNNs), the VGG19 architecture, and DenseNet121, with transfer learning leveraging CheXNet-pretrained weights optimized for medical imaging. Preprocessing techniques and hyperparameter optimization using the Hyperband algorithm were applied, and ensemble learning through soft voting was employed to integrate model outputs. The models were trained and evaluated on the FracAtlas dataset, which comprises over 4,000 X-ray images. Results indicated that the conventional CNN achieved an accuracy of 82.89%, although fracture recall was limited to 13%. Both VGG19 and DenseNet121 improved performance balance, achieving area under the curve (AUC) values of 0.79 and 0.81, respectively. The ensemble learning model achieved a performance close to that of the individual models. These findings demonstrate that deep learning can effectively support fracture diagnosis, particularly when incorporating transfer learning. However, challenges such as data imbalance and clinical case variability continue to affect model performance. This study represents a step toward the development of more reliable clinical decision support systems for fracture detection.

Keywords: Bone fractures, X-ray imaging, deep learning, medical diagnosis.

الكشف الآلي عن كسور العظام بالأشعة السينية باستخدام تقنيات التعلم العميق والتعلم الجماعي

بسمة بلعم[1]، عاطف الدنفرية[1]

[1]قسم علوم حاسوب، كلية تقنية المعلومات، جامعة مصراتة، مصراتة، ليبيا.

ملخـــــص البحـــــث

كسور العظام من أكثر الإصابات شيوعا على مستوى العالم، وتشكل تحديا في التشخيص الدقيق، حيث إن نسبة الخطأ قد تصل إلى 10% مما يؤدي إلى مضاعفات صحية وتأخير العلاج، تهدف هذه الدراسة إلى تطوير وتقييم نماذج تعلم عميق لتحسين دقة وكفاءة

كشف كسور العظام بالاعتماد على ثلاثة أطر رئيسية هي الشبكات العصبية التلافيفية التقليدية (Convolutional Neural Networks, CNNs) وبنية (VGG19) وبنية (DenseNet121)، مع الاستفادة من التعلم بالنقل (Transfer learning) باستخدام أوزان (CheXNet) المهيأة للصور الطبية، وكذلك جرى تطبيق تقنيات المعالجة المسبقة وضبط المعلمات (Hyperparameter) باستخدام خوارزمية (Hyperband)، كما استخدمنا التعلم الجماعي (Ensemble Learning) بطريقة (Soft Voting) لدمج مخرجات النماذج، وتم تدريب النماذج وتقييمها على قاعدة بيانات (FracAtlas) التي تضم أكثر من أربعة آلاف صور، وأظهرت النتائج أن النموذج التقليدي (CNN) حقق دقة بلغت 82.89% إلا أن استدعاء حالات الكسر كان ضعيفا بنسبة 13%، بينما حسنت نماذج (VGG19) و(DenseNet121) من التوازن في الأداء؛ حيث بلغت قيم منحنى (Area Under the Curve AUC) 0.79 و0.81 على التوالي، أما نموذج التعلم الجماعي فقد حقق أداء مقاربا، تشير هذه النتائج إلى أن التعلم العميق قادر على دعم تشخيص كسور العظام، خاصة مع استخدام نماذج (Transfer learning)، إلا أن تحديات مثل: عدم توازن البيانات، وتنوع الحالات السريرية تؤثر على الأداء بشكل كبير، وبذلك تمثل هذه الدراسة خطوة نحو بناء أنظمة دعم قرار سريري أكثر ثقة.

الكلمات الدالة: كسور العظام، الأشعة السينية، التعلم العميق، التشخيص الطبي.

1. المقدمة

يتكون الهيكل العظمي للإنسان البالغ من 206 عظمة[1]، ورغم قوة العظام وصلابتها، فإنها تظل عرضة للإصابات المختلفة، وعلى رأسها الكسور، ويعرف الكسر بأنه انقطاع كامل أو جزئي في استمرارية العظم؛ يؤدي إلى فقدان الاستقرار الميكانيكي، وقد ينشأ عن أسباب متعددة مثل: السقوط، أو حوادث السيارات، أو الإصابات الرياضية، أو الأمراض المزمنة كمرض هشاشة العظام، وفي لحظة واحدة، قد يؤدي حادث بسيط إلى كسر يغير مسار حياة المريض[2]؛ الأمر الذي يبرز الأهمية الطبية والاجتماعية لهذه الإصابات.

تشير الإحصائيات إلى أن معدل حدوث الكسور مرتفع عالميا، إذ سجل نحو 178 مليون كسر جديد في عام 2019م بزيادة بلغت 33.4% مقارنة بعام 1990م[3]، وفي المملكة المتحدة مثلا يتراوح معدل الإصابات بالكسور بين 733 و4,017 لكل 100,000 شخص سنويا، بينما تتراوح نسبة الخطأ في التشخيص بين 3% و10%[4]، وقد يقود التشخيص الخاطئ أو المتأخر إلى مضاعفات خطيرة تشمل الإعاقة الدائمة أو حتى الوفاة[5]، ورغم التطور الكبير في تقنيات التصوير الطبي، لا تزال صور الأشعة السينية محدودة القدرة على إظهار التفاصيل الدقيقة[3]، مما يجعل دقة التشخيص رهينة بخبرة اختصاصي الأشعة، ومع الارتفاع المتزايد في الطلب عليهم تظهر الإحصائيات انه حتى مع توفر اختصاصي الأشعة أداء اختصاصي الأشعة يقل بسبب تأثير التعب بحلول نهاية يوم العمل[4].

لذلك انطلاقا من ذلك، تتمحور المشكلة البحثية حول الحاجة إلى تقييم كفاءة نماذج التعلم العميق في الكشف عن كسور العظام باستخدام صور الأشعة السينية، في ظل التحديات المرتبطة بالتشخيص التقليدي؛ ومن هنا تتحدد أسئلة البحث:

- ما مدى دقة النماذج المختلفة للشبكات العصبية التلافيفية (DenseNet121, VGG19, CNN) في الكشف عن كسور العظام؟
- كيف يختلف أداء هذه النماذج من حيث المقاييس الأساسية (Accuracy، Precision، Recall، F1-Score)؟
- ما أثر تقنيات التعلم الجماعي (Ensemble Learning) على تحسين الأداء مقارنة باستخدامها بشكل منفرد؟

وبناء على هذه التساؤلات، تطرح فرضيات البحث على النحو الآتي:

- يمكن للنماذج العميقة (CNNs, VGG19, DenseNet121) تحقيق دقة عالية في الكشف عن الكسور باستخدام صور الأشعة السينية.
- تؤدي النماذج المتقدمة (VGG19 وDenseNet121) إلى نتائج أفضل من الشبكات التقليدية (CNN).
- يسهم دمج النماذج عبر تقنيات التعلم الجماعي (Ensemble Learning) في رفع مستوى الأداء الكلي وتحقيق نتائج أكثر استقرارا.

وللإجابة عن هذه التساؤلات، واثبات فرضياتنا اعتمدنا على منهجية تجريبية تضمنت تطوير ثلاثة نماذج تعلم عميق (CNNs, VGG19, DenseNet121) وتقييم أدائها، مع تطبيق أسلوب التعلم الجماعي بطريقة (Soft Voting) لتعزيز النتائج، وتم تدريب النماذج باستخدام صور أشعة سينية للعظام، مع الاستعانة بخوارزميات الضبط الدقيق (Fine-Tuning)، واختيار المعلمات الفائقة المثلى باستخدام خوارزمية (Hyperband).

وبذلك تسعى هذه الدراسة إلى الربط بين الإمكانات النظرية للتعلم العميق وتطبيقاته العملية في الطب السريري، من خلال تقديم تقييم مقارن شامل لأداء النماذج، واستكشاف حدودها؛ مما يعزز جودة الرعاية الصحية ويقلل من المخاطر المرتبطة بالتشخيص الخاطئ أو المتأخر.

2.1 الدراسات السابقة

ركزت التطورات الحديثة في التعلم العميق لاكتشاف كسور العظام على الاستفادة من هياكل الشبكات العصبية التلافيفية (CNN) وتحسين أساليب المعالجة المسبقة، بالإضافة إلى استكشاف تقنيات التعلم الجماعي لتعزيز الدقة السريرية.

في دراسة أجراها Jakub Ola et al (2017) استخدم خمس شبكات عميقة مثل:(VGG)، و(BVLC Reference CaffeNet network) على مجموعة بيانات مكونة من 256,458 صورة أشعة سينية، محققين دقة لا تقل عن 90%، وكان نموذج (VGG) الأفضل بدقة 83%؛[6] ومع ذلك كان نهجهم محدودا بالاعتماد على الصور السريرية الفردية، مما دفع الباحثين لاحقا لاستخدام التصوير متعدد الزوايا وتحسين النماذج.

لذلك حل مشكلة الصور الفردية Yutoku Yamada et al (2020) من خلال دمج ثلاثة أوضاع تصوير لكسور عظم الفخذ والحدبة، حيث استخدم 1,703 صورة شعاعية امامية، و1,220 صورة شعاعية جانبية، واستخدم مزيج من الموضعين، والتي اخذت صورة واحدة فقط من كل مريض لتجنب الافراط في التجهيز (Overfitting)، محققه دقة مماثلة أو أفضل من دقة الجراحين، حيث بلغت (Precision) و (Recall) نسبة 0.98 لكل منهما [7].

وأظهرت دراسات لاحقة مثل Kemal Üreten et al (2022) فعالية بنية (ResNet50)، و(GoogleNet)، و(VGG16) خصوصا في الكشف عن كسور الرسغ، حيث حققت (VGG16) دقة 93.3% و84% للمجموعتين المختبرتين [8]، وعلى النقيض، أظهرت دراسة Shinawar Naeem et al (2023) تفوق نماذج (ResNet) على (VGG) في تصنيف كسور أصابع اليد، حيث بلغت دقة 81.9% لنموذج (ResNet) مقارنة بدقة 78.5% لنموذج (VGG) [9]، وتشير هذه الدراسات إلى أن نماذج (CNN) المختلفة ومواضع الكسر قد تعطي نتائج متباينة.

كما أبرزت الدراسات أهمية المعالجة المسبقة وتحسين الأداء، فاستخدم Salih Beyaz et al (2020) الخوارزميات الوراثية لتحسين معلمات (CNN) لكسور عنق الفخذ، وعلى الرغم من مجموعة البيانات الغير متوازنة كانت ابرز نتائج الدراسة 83%(Sensitivity)، و 73% (Specificity) ومع تضمين (GA) زاد هذا

المعدل بنسبة 1.6%، كما قامت الدراسة بتحويل الصور إلى تدرجات رمادية، وقص الصور[10]، وأكد اكتشاف كسر الضلع بواسطة Tien Huang et al (2023) أيضا على المعالجة المسبقة، باستخدام (AlexNet)، و(DenseNet) على مجموعة بيانات من 2,000صورة أشعة سينية للوصول إلى دقة (Accuracy) تبلغ حوالي 92%على الرغم من تحيزات مجموعة البيانات[11]، بالتالي أظهرت النتائج أن التجهيز المسبق للصور وتحسين الشبكات يزيد من الدقة، خصوصا عند التعامل مع بيانات غير متوازنة أو تحتوي على ضوضاء.

واعتمد الباحثون أيضا التعلم الجماعي (Ensemble Learning) استخدم Fatih Usal et al (2021) تقنيات التعلم الجماعي (Ensemble Learning) على مجموعة بيانات (MURA)، حيث جمعوا نماذج مثل: (DenseNet)، و(ResNet)، و(VGG) مع خطوات المعالجة المسبقة[12]، وبالمثل قام Ayesha Tahir et al 2024) بتحسين التعلم الجماعي بشكل أكبر من خلال الجمع بين (MobileNetV2، VGG16، InceptionV3، ResNet50) نتج عن هذا ان نموذج المجموعة يتفوق مقارنة بنماذج التعلم العميق المنفردة 92.96% و 91.62% و 92.14% من (Accuracy)، و(Recall)، و(F1 Score) على التوالي لنموذج المجموعة[13].

على الرغم من التقدم الكبير في اكتشاف كسور العظام باستخدام شبكات (CNN) والتعلم الجماعي، فإن الدراسات السابقة غالبا ما اقتصرت على نوع واحد من الصور السريرية، أو لم تستفد بالكامل من تقنيات المعالجة المسبقة، أو لم تطبق ضبطا متقدما للمعلمات، كما أن قابلية تعميم النماذج على بيانات متنوعة كانت محدودة.

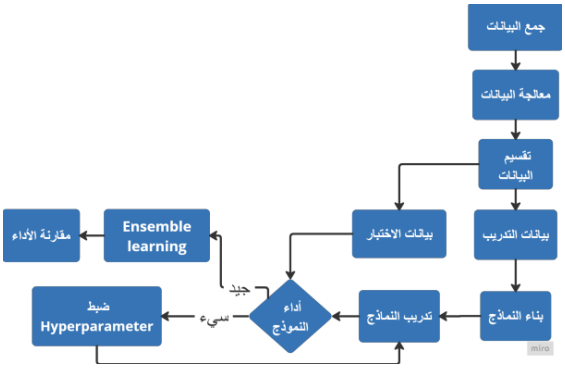
تسعى هذه الدراسة إلى معالجة هذه القيود من خلال دمج ثلاثة نماذج رئيسية للتعلم العميق (VGG19، DenseNet121، وCNN التقليدي)، مع استخدام الضبط التلقائي للمعلمات عبر خوارزمية (Hyperband)، والاستفادة من أوزان (CheXNet) المهيأة للصور الطبية، كما سيتم تطبيق منهجية الدمج الجماعي (Soft Voting Ensemble)؛ لتحسين دقة الكشف، مع التركيز على معالجة تحديات توازن البيانات، وتعزيز قدرة النماذج على التعميم على مجموعات بيانات متنوعة.

هذه الدراسة خطوة نحو تطوير أنظمة كشف آلي أكثر دقة وموثوقية لكسور العظام، مع قدرة أكبر على التكيف والتعميم في البيئات السريرية المستقبلية.

2. الجانب العملي والمنهجية

اتبعنا منهج شامل حيث بدأنا بالحصول على بيانات الاشعة السينية التي تشمل حالات الكسر وغير الكسر، ثم تم تطبيق تقنيات المعالجة المسبقة لضمان جودة البيانات؛ بالتالي ضمان جودة النماذج، ثم استخدمنا البيانات المعالجة لتدريب واختبار ثلاث هياكل تعلم عميق: الشبكة العصبية التلافيفية(CNN) ، و(VGG19)، و(DenseNet121) ثم استخدمنا التعلم الجماعي(Soft Voting) ؛ لدمج نقاط القوة في النماذج الثلاثة.

وتم تقييم أداء النماذج الفردية والمقارنة فيما بينها؛ بالتالي قد سمح هذا النهج بإجراء تقييم شامل لنقاط القوة لكل نموذج في سياق اكتشاف كسور العظام من صور الأشعة السينية.



شكل 1: منهجية الدراسة.

وستقدم الأقسام التالية شرحا لكل خطوة في إطارنا المنهجي، مما يضمن إمكانية إعادة الإنتاج وتسهيل الأبحاث المستقبلية في هذا المجال.

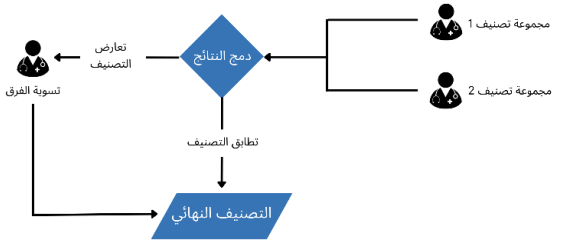
2.1 مجموعة البيانات

استخدمنا مجموعة بيانات (FracAtlas) التي تضم 4,083 صورة أشعة سينية للعظام، جمعت خلال عامي 2021م و2022م من ثلاثة مستشفيات ومراكز تشخيص في بنغلادش، صممت المجموعة لمهام تصنيف الكسور، وتحديد مواقعها، وتقسيمها، وتتميز بتعليقات توضيحية عالية الجودة تغطي أجزاء تشريحية متنوعة مثل: الكتف، والساق، والورك، والإصبع، واليد [14].

هذه المجموعة متاحة للاستخدام العام تحت رخصة Creative Commons Attribution 4.0 (CC-BY 4.0)، مما يتيح نسخها ومشاركتها وإعادة تشكيلها أو البناء عليها شريطة نسب المصدر.

تحتوي البيانات على 717صورة بها كسور (922 حالة كسر) مقابل 3,366 صورة دون كسور، مع تنوع ديمغرافي يشمل مرضى بأعمار بين 8أشهر و78عاما، وتوزيع جنسي نسبته 62% ذكور و38% إناث، تتضمن الصور أحيانا وجهات نظر متعددة للعضو، وأجهزة تثبيت العظام (Orthopedic Fixation Devices)، مما يعكس واقعية البيانات وقابليتها للتطبيق السريري [14].

وجرى تصنيف البيانات بواسطة اثنين من اختصاصي الأشعة، وفي حال الاختلاف احيلت الصور على جراح عظام خبير لضمان الدقة وموثوقية التصنيفات النهائية.



شكل 2: عملية تصنيف البيانات.

ووقع الاختيار لقاعدة بيانات (FracAtlas)؛ لاستخدمها في دراستنا نظرا لتغطيتها الشاملة، وموثوقية تعليقاتها، وتنوعها الذي يعكس ظروفا سريرية حقيقية، إضافة إلى التزامها بسرية المرضى؛ مما يسمح باستخدامها لأغراض البحث دون المساس بسرية المريض [14].

2.2 معالجة البيانات

خضعت البيانات لمرحلة معالجة؛ لضمان جودة وموثوقية النماذج، شملت هذه المرحلة أولا اكتشاف الصور التالفة وإزالتها، حيث تم فحص جميع صور الأشعة السينية باستخدام (TensorFlow)، وتم تحديد 59 صورة تالفة استبعدت من مجموعة البيانات؛ لمنع تأثيرها السلبي على التدريب، بعد ذلك تم تطبيع الصور (Normalization) عن طريق تقسيم قيم البكسل التي تتراوح عادة بين 0 و 255، [15]على 255 لتحويلها إلى نطاق 1–0؛ مما يحقق تجانس البيانات ويسهل تدريب الشبكات العصبية[16]، كما تم توحيد أبعاد جميع الصور لتصبح 224×224 بكسل؛ وهو الحجم المطلوب لنماذج (VGG19) و(DenseNet121)؛ الأمر الذي حافظ على التفاصيل الدقيقة للكسور وساهم في تحسين الكفاءة الحاسوبية وسرعة التدريب، بعد معالجة الصور تم تقسيم مجموعة البيانات إلى ثلاث مجموعات رئيسية: التدريب بنسبة 70%، والتحقق بنسبة 20% والاختبار بنسبة 10% مع خلط البيانات لضمان توزيع عشوائي وتقليل أي تحيز محتمل[17].

ونظرا لوجود اختلال واضح في توازن الفئات، حيث كانت 3,307 صورة غير مكسورة مقابل 717 صورة مكسورة، تم في البداية تجربة (Data Augmentation) بهدف زيادة تنوع العينات وتعويض النقص في الفئة الأقل تمثيلا، إلا أن النتائج لم تكن مرضية؛ لذلك تم حساب أوزان الفئات من مكتبة (Scikit-Learn) باستخدام أداة (compute_class_weight) وضبط الوضع على (Balanced)، مما أدى إلى أوزان تقريبية 0.608 للفئة غير المكسورة و2.806 للفئة المكسورة،

وتم دمج هذه الأوزان في عملية التدريب؛ لتحسين قدرة النموذج على تعلم الفئة الأقل تمثيلا، وتعزيز دقة اكتشاف الكسور [18].

2.3 التعلم العميق في تصنيف صور الأشعة السينية

في هذه الدراسة استخدمنا ثلاث أطر رئيسية للتعلم العميق لتصنيف صور الأشعة السينية واكتشاف كسور العظام: (CNN)، و (VGG19)، و (DenseNet121).

1.2.3 الشبكة العصبية التلافيفية (CNN)

تم بناء شبكة عصبية تلافيفية (CNN) باستخدام أربع كتل تلافيفية متتالية، بحيث تتدرج عدد المرشحات (Convolutional Blocks) من 32 إلى 256، مع الحفاظ على بنية موحدة لكل كتلة.

كل كتلة تحتوي على طبقتي (Conv2D)، يلي كل طبقة (Batch Normalization)؛ لزيادة سرعة التدريب، ثم تفعيل (LeakyReLU) بمعامل 0.1 لإدخال اللاخطية، بعد طبقتي (Conv2D)، يتم تقليل الأبعاد المكانية باستخدام (MaxPooling2D)، مع إدراج (Spatial Dropout2D) لتقليل الإفراط في التعلم (Overfitting)، حيث تتدرج نسب (Dropout) من 0.25 في الكتلة الأولى إلى 0.5 في الكتلة الرابعة.

بعد الكتل التلافيفية، تستخدم طبقة (Global Average Pooling) لجمع المعلومات المكانية وتقليل الأبعاد مع الحفاظ على السمات الأساسية المهمة؛ مما يقلل خطر الإفراط في التعلم (Overfitting) ويحافظ على تمثيلات الميزات الرئيسية.

ثم تأتي رأس التصنيف التي تتضمن طبقتين كثيفتين (Dense) بعدد وحدات 256 و128 على التوالي، مع تطبيق (Regularization L2)، و(Batch Normalization)، و (LeakyReLU) للتنشيط، بالإضافة إلى (Dropout) بمعدلات 0.5 و0.3 لضبط النموذج وتقليل الإفراط في التعلم (Overfitting)، أخيرا تتكون طبقة الإخراج من خلية واحدة مع دالة تنشيط (Sigmoid)، ما يسمح بالتصنيف الثنائي لصور الأشعة السينية إلى وجود كسر أو عدمه.

نقوم بعدها بتجميع النموذج (Model Compilation)، فتم استخدام محسن (AdamW) بمعدل تعلم ابتدائي 0.0001، ومعامل (Weight Decay) مقداره 0.0001، ودالة خسارة (Binary Cross Entropy) لملاءمتها مع التصنيف الثنائي، وكمعايير تقييم تم تضمين (AUC) لمراقبة الأداء أثناء التدريب.

ولضمان التوازن بين التعلم الجيد ومنع الإفراط في التعلم؛ يتم استخدام (Early Stopping) لمراقبة الأداء على مجموعة التحقق (Validation Set) وإيقاف التدريب عند تدهور الأداء، مع تحديد 200 كحد أقصى لعدد الدورات (Epochs).

2.2.3 نماذج التعلم بالنقل (VGG19) و (DenseNet121)

تستخدم هذه الدراسة التعلم بالنقل (Transfer Learning) تحديدا (VGG19) و (DenseNet121) للكشف عن كسور العظام، حيث (VGG19) و (DenseNet121) هما شبكتين عصبيتين ملتويتان عميقتان مدربتان مسبقا عادة على مجموعة بيانات (ImageNet)؛ وبالتالي الاستفادة من قدرتهما على التعرف على الميزات العامة المكتسبة من مجموعة البيانات الضخمة التي تتكون من صور طبيعية وتكيفها للعمل على الصور الطبية.

تم تهيئة أوزان النموذجين في دراستنا باستخدام نموذج (CheXNet) بدلا من الاعتماد على أوزان (ImageNet)؛ والسبب في ذلك أن التدريب باستخدام أوزان (ImageNet) لم يحقق نتائج مرتفعة، وهو أمر منطقي نظرا لاختلاف طبيعة الصور الطبيعية في (ImageNet) عن الصور الطبية، وعلى العكس من

ذلك فإن اوزان (ChexNet) صممت خصيصا لتحليل صور الأشعة السينية؛ مما يجعله نقطة انطلاق مناسبة لاستخراج الميزات الطبية الدقيقة وتحسين أداء النموذج.

في المرحلة الأولى، تم تجميد الطبقات الأساسية للحفاظ على الأوزان المدربة مسبقا، بينما جرى إلغاء التجميد للطبقات العليا من النموذجين الإجراء عملية الضبط الدقيق (Fine-Tuning).

- بدأ إلغاء التجميد في VGG 19 من الطبقة الثامنة فما فوق.
- بدأ إلغاء التجميد في DenseNet121 من الطبقة التاسعة والخمسين فما فوق.

يعكس هذا النهج حقيقة أن الطبقات الدنيا للشبكات العصبية التلافيفية تلتقط ميزات عامة منخفضة المستوى مثل الحواف والأشكال البسيطة وهي قابلة لإعادة الاستخدام في مهام متعددة، بينما الطبقات العليا مسؤولة عن الميزات الأكثر تخصصا، لذلك أعيد تدريبها لتتكيف مع طبيعة صور كسور العظام وتتعلم أنماطا جديدة مرتبطة بها، وبهذا يحقق الضبط الدقيق توازنا بين الاستفادة من المعرفة السابقة المخزنة في الأوزان المدربة مسبقا وبين تخصيص النموذج المهمة التصنيف الطبي.

شكل 3: عملية (Transformer fine-tuning).

فوق القاعدة (VGG19 و DenseNet121) أضيف رأس تصنيف مخصص لمهمة التصنيف الثنائي وجود كسر أو عدمه، يتكون من:

1 طبقة (Flatten) لتحويل خرائط الميزات ثنائية الأبعاد إلى متجه أحادي الأبعاد.

2 طبقة (Dense) بعدد قابل للضبط من الوحدات (Units) يتراوح من 128 إلى 512 وتنشيط (ReLU).

3 طبقة (Dropout) للتنظيم بمعدل بين 0.2 و0.5.

4 طبقة (Dense) اخرى للإخراج مع خلية واحدة وتنشيط (Sigmoid).

وبعد تحديد البنية المعمارية للنموذج وإضافة رأس التصنيف المخصص، تأتي الخطوة التالية المتمثلة في ضبط المعلمات الفائقة (Hyperparameter tuning)، حيث يعتمد أداء نماذج التعلم العميق بشكل كبير على اختيار المعلمات الفائقة (Hyperparameters) المناسبة التي تضمن الوصول إلى أفضل النتائج الممكنة؛ ولتحقيق ذلك تم الاستعانة بخوارزمية (Hyperband) المدمجة في مكتبة (Keras Tuner)، حيث تعمل هذه الخوارزمية على اختيار الأنسب منها بصورة آلية.

تشمل عملية الضبط اختيار مجموعة من المعلمات الأساسية، وهي:

- عدد الوحدات في الطبقة الكثيفة (Dense Layer) والتي تتراوح من 128 إلى 512.
- معدل الإسقاط (Dropout Rate) والتي بين 0.2 و0.5.
- اختيار المحسن (Adam)، أو (RMSprop)، أو (SGD).
- معدل التعلم للمحسن التي تم اختياره.

وفي النهاية يتم تجميع النموذجين باستخدام (Binary Cross Entropy) كدالة خسارة، والتي تتناسب مع طبيعة التصنيف الثنائي لمهمة اكتشاف كسر العظام، ويتم تحديد المحسن تلقائيا خلال عملية الضبط مع (AUC) كمقياس تقييم.

3.2.3 التعلم الجماعي

في هذه الدراسة طبقنا نهج التعلم الجماعي (Ensemble Learning) لتعزيز دقة اكتشاف كسور العظام، فيجمع التعلم الجماعي بين التنبؤات من نماذج

متعددة لاتخاذ قرار نهائي؛ وبالتالي الاستفادة من نقاط قوتها الجماعية، لذلك استخدمنا طريقة التصويت الناعم (Soft Voting) التي تجمع بين تنبؤات نماذج التعلم العميق الثلاثة: (CNN) و (VGG19) و (DenseNet121) ويتم تنفيذ ذلك على النحو التالي:

1. يقوم كل من النماذج الثلاثة (CNN)، و (VGG19) و (DenseNet121) بتوليد احتمالية إذا كانت صورة الأشعة تحتوي على كسر أو لا.
2. يتم دمج هذه الاحتمالات وأخذ المتوسط المرجح لها.
3. يعتمد التصنيف النهائي على أعلى قيمة احتمال (كسر أو غير كسر)؛ مما يضمن الاستفادة من نقاط قوة كل نموذج على حدة.

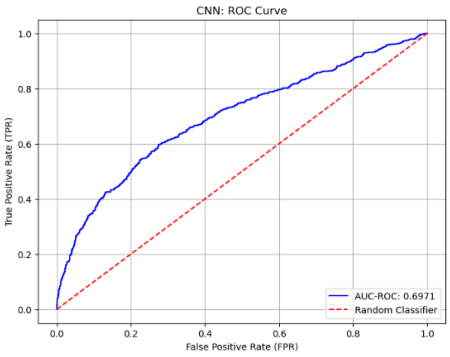
3. النتائج

لتقييم ومقارنة الاداء بين النماذج، تم تقييم النماذج باستخدام (Accuracy)، و (Precision)، و(Recall)، و(F1 Score)، و(AUC-ROC) لإعطاء صورة شاملة عن قدرتها على تمييز صور الكسور من غيرها؛ وبذلك يبرز هذا التحليل نقاط القوة والضعف لكل نموذج ويساعد في اختيار الأنسب للتطبيق السريري.

3.1 أداء النماذج الفردية

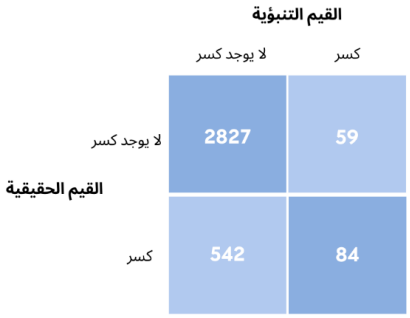
1.3.1 الشبكة العصبية التلافيفية (CNN)

حقق النموذج (Accuracy) بنسبة 82.89%، ومع ذلك هذه (Accuracy) منحازة؛ بسبب عدم توازن البيانات حيث بالنسبة للفئة التي لا يوجد بها كسر أظهر النموذج (Precision) بنسبة 84%، و (Recall) بنسبة 98%؛ مما يشير إلى قدرته العالية على تجنب الإيجابيات الكاذبة، أما بالنسبة لفئة الكسور، فقد بلغت (Precision) 59% و (Recall) 13% فقط، وهو ما يشير إلى صعوبة في التقاط حالات الكسر، وأدى ذلك إلى (F1 Score) منخفضة 22% وبلغت (AUC-ROC) للنموذج 0.6971؛ مما يدل على قدرة متوسطة على التمييز بين الفئتين.



شكل 4: منحنى (ROC) لنموذج (CNN).

أوضحت مصفوفة الارتباك دقة النموذج بشكل أكبر، حيث تم تصنيف 2,827 حالة غير كسر بشكل صحيح، و59حالة منها فقط تم تصنيفها بشكل خاطئ ككسور (إيجابيات كاذبة)، في حين تم تصنيف 84 حالة كسر بشكل صحيح، بينما تم تصنيف 542 حالة كسر على أنها غير كسور (سلبيات كاذبة)؛ هذا التوزيع يبرز التحدي الذي يواجه النموذج في تحديد جميع حالات الكسر بدقة.

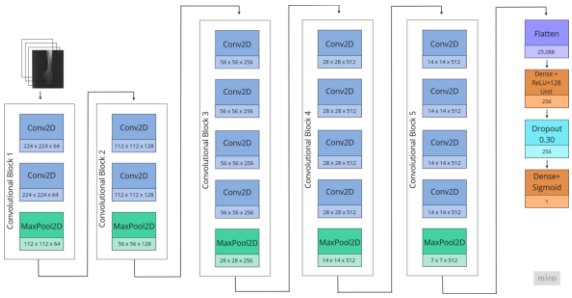


شكل 5: مصفوفة الارتباك لنموذج (CNN).

2.3.1 بنية VGG19

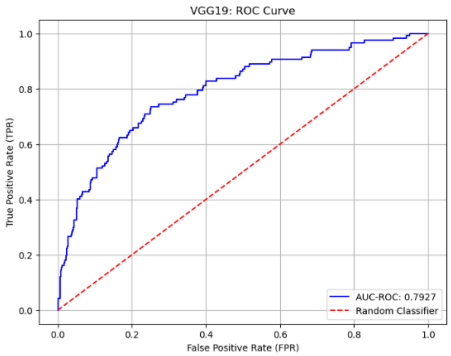
قمنا بضبط نموذج (VGG19) للكشف عن كسور العظام، فحدد ضبط المعلمات الفائقة (Hyperparameter tuning) تكوين لرأس التصنيف طبقة كثيفة بها 256 وحدة (Units)، ومعدل

0.30 (Dropout Rate)، ومحسن (RMSprop) بمعدل تعلم 0.0001.



شكل 6: بنية (VGG19).

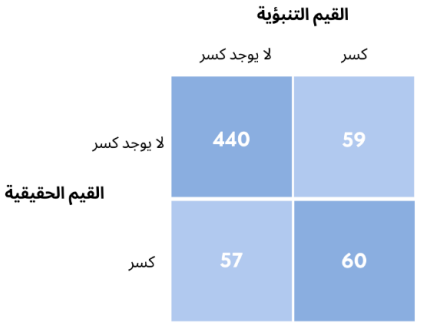
وأظهر نموذج (VGG19) أداء جيدا على مجموعة بيانات الاختبار، فحقق دقة (Accuracy) بنسبة 81.17%، أما بالنسبة لفئة غير الكسر كانت (Precision) ، و 89% (Recall) ،و 88% على التوالي؛ مما أدى إلى (F1-Score) بنسبة 88%، بالنسبة لفئة غير الكسر، حقق النموذج (Precision) بنسبة 50% وتذكر (Recall) بنسبة 51% و (F1-Score) بلغت 51% ،و أظهر منحنى (ROC) أن النموذج يمتلك قدرة تصنيفية جيدة، حيث بلغ معدل (AUC-ROC) حوالي 0.7927؛ وهو ما يعكس توازنا نسبيا في التمييز بين الفئتين، إلا أن هناك مجالا لتحسين التمييز خصوصا لفئة الكسر.



شكل 7: منحنى (ROC) لنموذج (VGG19).

تشير هذه النتائج إلى أن النموذج تمكن من تصنيف عدد كبير من الحالات بشكل صحيح، إلا أن الأداء ما زال أفضل مع فئة غير الكسر مقارنة بفئة الكسر، مما يعني أن بعض حالات الكسور قد لا تكتشف، وهو أمر بالغ الأهمية في التطبيقات الطبية ويتطلب تحسينا مستقبليا.

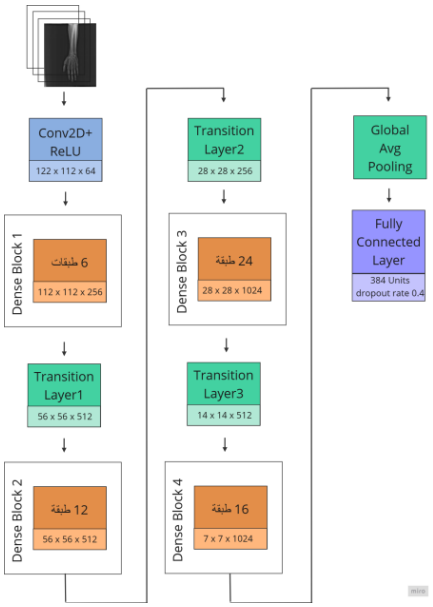
وأكدت مصفوفة الارتباك هذه النتائج، حيث حدد نموذج (VGG19) بشكل صحيح 440 حالة غير كسر و60 حالة كسر، وتم تصنيف 59 حالة غير كسر بشكل خاطئ على أنها كسور (إيجابيات كاذبة)، وتم تصنيف 57 حالة كسر بشكل خاطئ على أنها غير كسور (سلبيات كاذبة).



شكل 8: مصفوفة الارتباك لنموذج (VGG19)

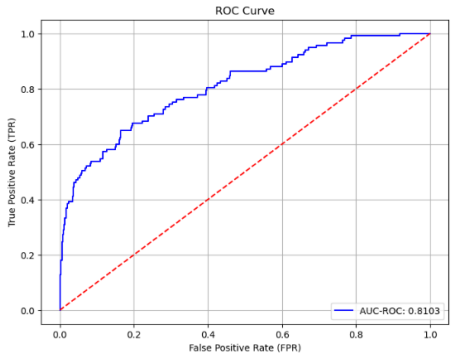
2.3.1 بنية DenseNet121

قمنا بضبط نموذج (DenseNet121) القائم على التعلم بالنقل لتصنيف صور الأشعة السينية للكشف عن كسور العظام، وحددت عملية ضبط المعلمات الفائقة (Hyperparameter tuning) التكوين المناسب لرأس تصنيف (DenseNet121)، فتضمن طبقة كثيفة تحتوي على 384 وحدة (Units)، ومعدل 0.4، ومحسن (RMSprop) بمعدل تعلم 0.001.



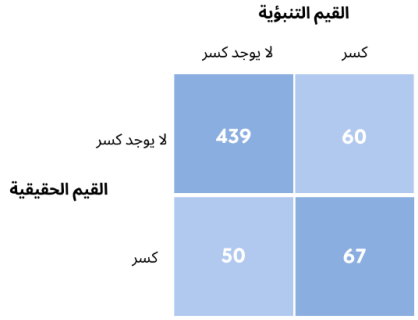
شكل 9: بنية (DenseNet121).

في مجموعة بيانات الاختبار، أظهرت (DenseNet121) أداء جيد، محقق دقة (Accuracy) بنسبة 82.14% أما بالنسبة غير المكسورة حقق دقة (Precision) بنسبة 90%، وتذكر (Recall) بنسبة 88%؛ مما أدى إلى (F1-Score) بنسبة 89%، وأما بالنسبة للكسور كانت الدقة 53% (Precision)، مع تذكر (Recall) بنسبة 57%مما أدى إلى (F1 Score) بنسبة 55%، كما حقق النموذج (AUC-ROC) بنسبة 81%؛ مما يعني قدرة النموذج على التمييز بين الصور التي تحتوي على كسور وتلك غير المكسورة.



شكل 10: منحنى (ROC) لنموذج (DenseNet121).

كشفت مصفوفة الارتباك عن تمكن (DenseNet121) من تحديد 439 حالة غير مكسورة 67 حالة كسر بشكل صحيح، ومع ذلك كان هناك60 نتيجة تصنيف صور غير مكسورة بشكل خاطئ على أنها كسور (إيجابية كاذبة)، و50 نتيجة تصنيف الكسور بشكل خاطئ على أنها غير كسور (سلبية كاذبة).

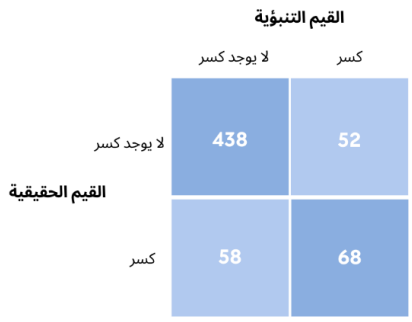


شكل 11: مصفوفة الارتباك لنموذج (DenseNet121).

3.2 تأثير التعلم الجماعي

أظهرت نتائج نموذج التعلم الجماعي تحسنا ليس بالكبير مقارنة بالنماذج الفردية، حيث بلغت (Accuracy)%82.14، أما (Precision) فكانت 88% لفئة غير الكسر و57% لفئة الكسر، مما يعكس قوة النموذج في تمييز الحالات السليمة أكثر من الكسور، وبلغ 89% (Recall) لغير الكسر و54% للكسر، ما يشير إلى قدرة عالية على اكتشاف غير الكسور مقابل ضعف نسبي في اكتشاف الكسور، كما حقق النموذج (F1 Score) بلغت 82%.

أظهرت مصفوفة الارتباك أن النموذج صنف 438 حالة غير كسر بشكل صحيح مقابل 52 حالة إيجابية كاذبة، بينما اكتشف 68 حالة كسر صحيحة مقابل 58 حالة سلبية كاذبة، توضح هذه النتائج أن النموذج يتمتع بدقة أعلى في التعرف على الحالات السليمة مقارنة بالحالات المصابة، مما يكشف عن تحدي عدم التوازن في الأداء، ورغم ذلك فإن دمج هياكل مثل: (CNN)، و(VGG19)، و(DenseNet121) يوفر أساسا مهما لتحسين الكشف عن الكسور وتعزيز إمكانية التطبيق السريري بعد إجراء تحسينات إضافية.



شكل 12: مصفوفة الارتباك للتعلم الجماعي (Ensemble Learning).

3.3 التحليل المقارن للنماذج

جدول رقم (1): مقارنة دقة النماذج.

المقياس	CNN	VGG19	DenseNet121	Ensemble Learning
Accuracy	%82.89	%81.17	%82.14	%82.14

جدول رقم(2): نتائج النماذج لفئة غير الكسور.

المقياس	CNN	VGG19	DenseNet121	Ensemble Learning
Precision	%84	%89	%90	%88
Recall	%98	%98	%88	%89
F1-Score	%90	%88	%89	%89

جدول رقم(3) :نتائج النماذج لفئة الكسور.

المقياس	CNN	VGG19	DenseNet121	Ensemble Learning
Precision	%59	%50	%53	%57
Recall	%13	%51	%57	%54
F1-Score	%22	%51	%55	%55

4. الاستنتاجات

أظهرت نتائج التقييم أن النماذج المدروسة تمتلك قدرة متفاوتة على تمييز صور كسور العظام من الصور السليمة، حيث برزت جميعها بأداء جيد على فئة الصور غير المكسورة مقابل ضعف نسبي في اكتشاف حالات الكسر، النموذج القائم على الشبكة العصبية التلافيفية (CNN) حقق أعلى دقة إجمالية، إلا أنه أظهر قصورا واضحا في استدعاء حالات الكسر ، وهو ما يعكس تأثير عدم توازن البيانات؛ في المقابل أظهرت النماذج المعتمدة على التعلم بالنقل مثل: (VGG19)، و (DenseNet121) ، أداء أكثر توازنا، إذ حسنت من قدرة النموذج على التمييز بين الفئتين مع تحقيق قيم (AUC-ROC) مرتفعة نسبيا، مما يعكس جدوى الاستفادة من الأوزان المدربة مسبقا على بيانات طبية.

كما أن نموذج التعلم الجماعي (Ensemble Learning) وفر أداء مقاربا للنماذج الفردية مع بعض التحسن في مقاييس التوازن بين الدقة والتذكر، إلا أن الفائدة كانت محدودة، وتظهر هذه النتائج أن التحدي الرئيسي يكمن في تحسين اكتشاف حالات الكسر، حيث تميل النماذج إلى تفضيل الحالات السليمة على حساب الحالات المرضية.

ويرجع السبب الأساسي على الأرجح إلى عدم توازن البيانات، رغم تطبيق بعض الدراسات لاستراتيجيات للتغلب على هذه المشكلة، إلا أن أثرها ظل واضحا في النتائج، كما أن تنوع البيانات ديموغرافيا، ووجود بعض الصور التي تحتوي على أجهزة تثبيت العظام (Orthopedic Fixation Devices)، قد ساهم في ضعف الأداء العام للنماذج، وعلى النقيض من ذلك، ركزت الدراسات الاخرى غالبا على جانب محدد أو عضو واحد، أو لم توفر التنوع الذي توفره قاعدة البيانات المستخدمة في هذه الدراسة، مما قد يفسر ظهور نتائج عالية في هذه الدراسات مع ضعف قابلية تطبيقها سريريا على بيانات أكثر تنوعا، توضح هذه النقاط التحديات العملية التي تواجه تطبيق نماذج الذكاء الاصطناعي سريريا، وهو ما يميز هذا البحث حيث يسعى إلى تقريب نتائج الذكاء الاصطناعي من التجربة السريرية الفعلية؛ مما يعزز دقة وكفاءة تشخيص كسور العظام في البيئات الطبية الواقعية.

تؤكد هذه الدراسة على أن تطوير تقنيات موازنة الفئات أثناء التدريب وتحسين استراتيجيات الضبط الدقيق، ودمج تقنيات متقدمة، قد يسهم في تعزيز قدرة النماذج على اكتشاف الكسور بدقة أعلى، وبذلك تمثل النتائج خطوة

مهمة نحو بناء أنظمة دعم قرار سريري موثوقة تساعد الأطباء في تشخيص كسور العظام من صور الأشعة السينية بشكل أكثر دقة وكفاءة.

5. التوصيات

استنادا إلى نتائج هذا الدراسة أن تقنيات التعلم العميق أظهرت نتائج واعدة في تصنيف صور الأشعة السينية واكتشاف الكسور، ومع ذلك هناك فرص للتطوير، ومن هنا تبرز التوصيات التالية:

- توسيع قاعدة البيانات: زيادة حجم البيانات وتطبيق المنهجية على أنواع مختلفة من طرق التصوير الطبي.
- تحقيق التوازن: استخدام بيانات أكثر توازنا؛ عبر توظيف أدوات وأساليب مختلفة لمعالجة عدم التوازن في العينات.
- تنويع النماذج: تجربة خوارزميات وأطر تعلم عميق أخرى تختلف عن النماذج المستخدمة في هذه الدراسة؛ وذلك لاختبار الأداء من زوايا متعددة وتحسين جودة النتائج.
- اختبارات سريرية: إجراء تجارب واسعة قبل اعتماد النماذج في البيئات الطبية للتأكد من موثوقيتها.
- اختبارات الدلالة الإحصائية: تطبيق اختبارات مثل: (McNemar) و(Bootstrapping)، للتحقق من الفروق الإحصائية بين النماذج.
- التحقق الخارجي ونقل التعلم التكيفي (Domain Adaptation): اختبار النماذج على بيانات من مؤسسات ومصادر مختلفة، مع تطبيق تقنيات (Domain Adaptation)؛ للتأكد من قدرة النموذج على التعميم في بيئات طبية متنوعة.
- توجيه الأطباء: التأكيد على دور هذه النماذج كأداة مساعدة مكملة للخبرة الطبية وليست بديلا عنها.

إن تنفيذ هذه التوصيات من شأنه أن يسهم في تطوير أنظمة الكشف الآلي عن كسور العظام في صور الأشعة السينية، وتعزيز فاعليتها في البيئات السريرية الواقعية.

6. المراجـــــع

[1] Medical News Today. Accessed 2024 Aug 28. Availablefrom: https://www.medicalnewstoday.com/articles/32044

[2] Bigham-Sadegh A, Oryan A. Basic concepts regarding fracture healing and the current options and future directions in managing bone fractures. Int Wound J. 2015;12(3):238–47. doi:10.1111/iwj.12231.

[3] منظمة الصحة العالمية. الكسور الناجمة عن هشاشة العظام. Accessed 2025 Jan 7. Available from: https://www.who.int/ar/news-room/fact-sheets/detail/fragility-fractures

[4] Kuo RYL, et al. Artificial Intelligence in Fracture Detection: A Systematic Review and Meta-Analysis. Radiology. 2022;304(1):50–62. doi:10.1148/radiol.211785.

[5] Tanzi L, Vezzetti E, Moreno R, Moos S. X-Ray Bone Fracture Classification Using Deep Learning: A Baseline for Designing a Reliable Approach. Appl Sci. 2020;10(4):1507. doi:10.3390/app10041507.

[6] Olczak J, et al. Artificial intelligence for analyzing orthopedic trauma radiographs: Deep learning algorithms—are they on par with humans for diagnosing fractures? Acta Orthop. 2017;88(6):581–6. doi:10.1080/17453674.2017.1344459.

[7] Yamada Y, et al. Automated classification of hip fractures using deep convolutional neural networks with orthopedic surgeon-level accuracy: ensemble decision-making with antero-posterior and lateral radiographs. Acta Orthop. 2020;91(6):699–704. doi:10.1080/17453674.2020.1803664.

[8] Üreten K. Use of deep learning methods for hand fractures detection from plain hand radiographs. Turk J Trauma Emerg Surg. 2020. doi:10.14744/tjtes.2020.06944.

[9] Naeem S, Naseer A, Rehman SU, Gruhn V, Akram S. Enhancing Finger Fracture Diagnosis: A Deep Learning Approach Using ResNet and VGG. 2023 Nov 30. doi:10.20944/preprints202311.1990.v1.

[10] Beyaz S. Femoral neck fracture detection in X-ray images using deep learning and genetic algorithm approaches. Jt Dis Relat Surg. 2020;31(2):175–83. doi:10.5606/ehc.2020.72163.

[11] Huang S-T, Liu L-R, Chiu H-W, Huang M-Y, Tsai M-F. Deep convolutional neural network for rib fracture recognition on chest radiographs. Front Med. 2023;10:1178798. doi:10.3389/fmed.2023.1178798.

[12] Uysal F, Hardalaç F, Peker O, Tolunay T, Tokgöz N. Classification of Shoulder X-ray Images with Deep Learning Ensemble Models. Appl Sci. 2021;11(6):2723. doi:10.3390/app11062723.

[13] Tahir A, Saadia A, Khan K, Gul A, Qahmash A, Akram RN. Enhancing Diagnosis: Ensemble deep learning model for fracture detection using X-ray images. Clin Radiol. 2024 Aug;S0009926024004197. doi:10.1016/j.crad.2024.08.006.

[14] Abedeen I, Rahman MA, Prottyasha FZ, Ahmed T, Chowdhury TM, Shatabda S. FracAtlas: A Dataset for Fracture Classification, Localization and Segmentation of Musculoskeletal Radiographs. Sci Data. 2023;10(1):521. doi:10.1038/s41597-023-02432-4.

[15] Albahadily HK, Tsviatkou VY, Kanapelka VK. Grayscale image compression using bit plane slicing and developed RLE algorithms. Int. J. Adv. Res. Comput. Commun. Eng.. 2017 Feb;6:309-14.

[16] Ahmed HA, Muhammad Ali PJ, Faeq AK, Abdullah SM. An Investigation on Disparity Responds of Machine Learning Algorithms to Data Normalization Method. ARO Sci J Koya Univ. 2022;10(2):29–37. doi:10.14500/aro.10970.

[17] Birchha V, Nigam B. Feature Selection Techniques And Hyper Parameter Tuning Impact On Classifier Performance For Breast Cancer Detection. Journal Of Pharmaceutical Negative Results. 2022 Oct 8;13.

[18] Krawczyk B. Learning from imbalanced data: open challenges and future directions. Prog Artif Intell. 2016;5(4):221–32. doi:10.1007/s13748-016-0094-0.