



Research Paper

Deep learning based suture training system

Mohammed Mansour^{a,*}, Eda Nur Cumak^a, Mustafa Kutlu^a, Shekhar Mahmud^b^a Department of Mechatronics Engineering, Sakarya University of Applied Sciences, Sakarya, Turkey^b Department of Systems Engineering, Military Technological College, Muscat, Oman

ARTICLE INFO

Keywords:

Deep learning
Suture training
Classification

ABSTRACT

Background and objectives: Surgical suturing is a fundamental skill that all medical and dental students learn during their education. Currently, the grading of students' suture skills in the medical faculty during general surgery training is relative, and students do not have the opportunity to learn specific techniques. Recent technological advances, however, have made it possible to classify and measure suture skills using artificial intelligence methods, such as Deep Learning (DL). This work aims to evaluate the success of surgical suture using DL techniques.

Methods: Six Convolutional Neural Network (CNN) models: VGG16, VGG19, Xception, Inception, MobileNet, and DensNet. We used a dataset of suture images containing two classes: successful and unsuccessful, and applied statistical metrics to compare the precision, recall, and F1 scores of the models.

Results: The results showed that Xception had the highest accuracy at 95 %, followed by MobileNet at 91 %, DensNet at 90 %, Inception at 84 %, VGG16 at 73 %, and VGG19 at 61 %. We also developed a graphical user interface that allows users to evaluate suture images by uploading them or using the camera. The images are then interpreted by the DL models, and the results are displayed on the screen.

Conclusions: The initial findings suggest that the use of DL techniques can minimize errors due to inexperience and allow physicians to use their time more efficiently by digitizing the process.

Introduction

Medical education aims to provide students with basic professional competencies, with patient safety as the top priority. While theoretical education is important, practical training is also necessary to equip students with the knowledge and skills they need. The acquisition of basic skills, such as surgical suturing, is crucial in the field of healthcare. A study conducted in 2013 at the Istanbul Faculty of Medicine found that only 71.6 % of 283 medical students surveyed were able to participate in surgical suture training, with 27.6 % feeling inadequate in this area [1].

Despite being a basic skill that all medical and dental students should learn (see Fig. 1), surgical suture training is not always emphasized in educational institutions for various reasons. Typically, suture placement by students is evaluated by experienced faculty members in surgical departments. However, this grading process can be burdensome for experienced surgeons, particularly under increasing workload and difficult conditions, and may also be a challenge for students who do not have enough time for training. The gold standard for suture

classification is expert evaluation during surgical training, but this can be difficult to coordinate, especially in undergraduate education. The COVID-19 pandemic has further disrupted the feasibility of these evaluations in medical education. One of the main purposes of medical education is to provide medical school students with basic professional competencies. Although it is not enough for students to receive only theoretical education, patient safety should be chosen as the most important goal, practical training should be given, and they should be equipped with knowledge and skills.

Recent medical studies have used computer vision [2] methods to classify and diagnose medical data, often achieving better results than medical doctors. It is well-known that these methods have been used to address deficiencies in the field. The development of basic medical skills has been a crucial focus in medical education in recent years. The training of medical students and hospital staff is evaluated by experts, and the feedback provided can be objectively applied without the need for subjective decision-making. This section reviews the use of technology-enhanced suture training based on system analysis and deep

* Corresponding author.

E-mail addresses: mmansour755@gmail.com (M. Mansour), edacumak@subu.edu.tr (E.N. Cumak), mkutlu@subu.edu.tr (M. Kutlu), shekhar.mahmud@mtc.edu.om (S. Mahmud).<https://doi.org/10.1016/j.sopen.2023.07.023>

Received 8 May 2023; Received in revised form 21 July 2023; Accepted 29 July 2023

Available online 6 August 2023

2589-8450/© 2023 The Authors. Published by Elsevier Inc. This is an open access article under the CC BY-NC-ND license (<http://creativecommons.org/licenses/by-nc-nd/4.0/>).

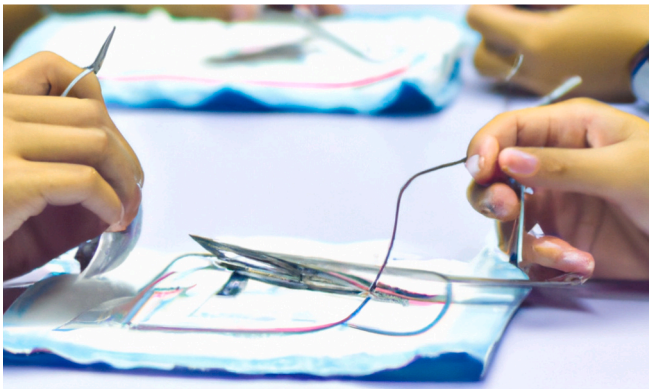


Fig. 1. Suture training example.

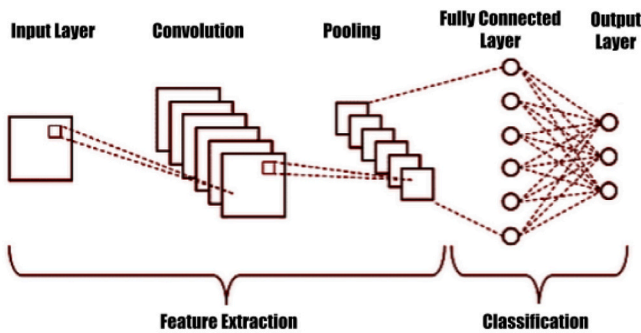


Fig. 2. CNN structure example.

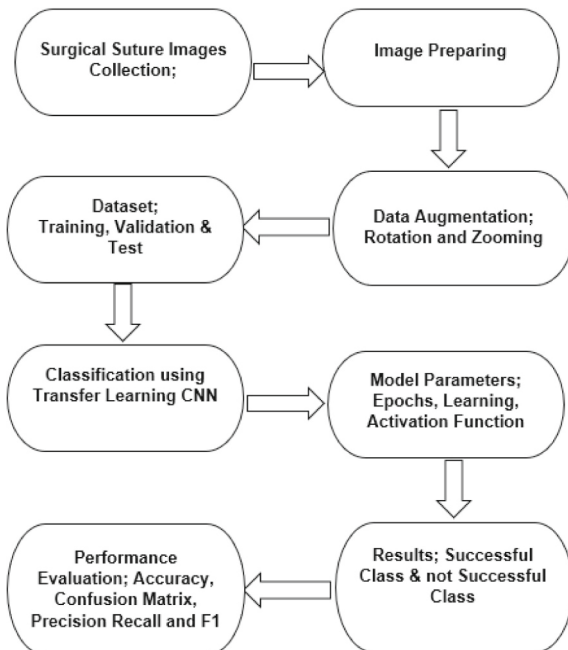


Fig. 3. Study flow chart.

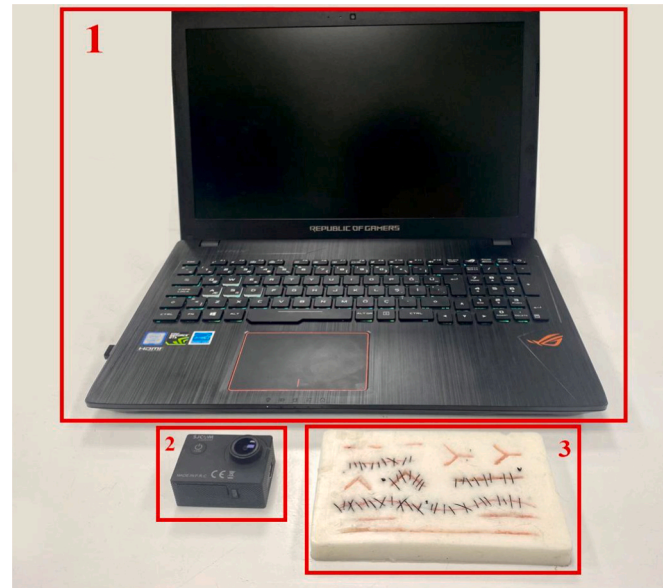


Fig. 4. System Setup: 1) Laptop, 2) Camera 3) Suture pad.

sewing operation on the platform, the algorithm takes into account various metrics such as the entry and exit times of the suture needle, the entry and exit points, and the length of the needle track. Woonjae et al. developed a flexible sensor for suture training to improve the training and working conditions of hospital personnel [4]. They believed that this sensor, embedded in an artificial skin simulator, would enhance the acquisition of surgical skills, particularly in terms of sensory training. Handermann et al. used a combination of computer vision-based software and fiber optic strain sensors to evaluate suture performance in the fields of general surgery and eye surgery [5]. The suture quality was evaluated using computer vision software, while the suture flow was assessed based on the voltage measurement of an optical fiber placed near the wound. Dubrovski et al. aimed to create an effective feedback system to facilitate motor learning skills by determining the forces and movements exerted on tissue during suturing [6]. They believed this system could be a viable tool for teaching basic technical skills. The main focus of the study was to examine, through detailed analysis, whether specific process measurements can be applied to hand movements and forces. Dose et al. have argued that it is possible to conduct an objective assessment of surgical skills using skill-based video analysis systems [7]. The aim of their study was to create a comprehensive surgical instrument. The device, developed by the Surgical Computing and Oncology and Imaging Research Group, is a dexterity-based motion analyzer. While these methods have been useful, there is still room for improvement by incorporating new data and simpler methods.

Deep learning (DL) and deep transfer learning are important techniques in data science and artificial intelligence that are used in statistical and predictive modeling [8–11]. Convolutional neural networks (CNN) are a type of DL architecture specifically designed for input formats such as images and are often used for image recognition and classification (see Fig. 2) [12–15]. These deep neural networks have been successful in a variety of real-world applications, including image classification, object detection, segmentation, and face detection. Transfer learning involves taking the classifier layer from a pretrained CNN and fine-tuning it on the target dataset [16]. This can reduce the training demands and is a common technique for using deep CNNs on small datasets.

The goal of this study is to predict the success of suture images using CNN transfer learning. This is the first study to use DL to classify suture images and assess the performance of different CNN models. Six well-known CNN models; VGG 16, VGG 19, Xception, Inception, Mobile

learning models. In their study, Ton et al. developed a computer vision algorithm to assess suturing ability [3]. This algorithm uses a synthetic deep-formed platform to evaluate suture skills. When performing the

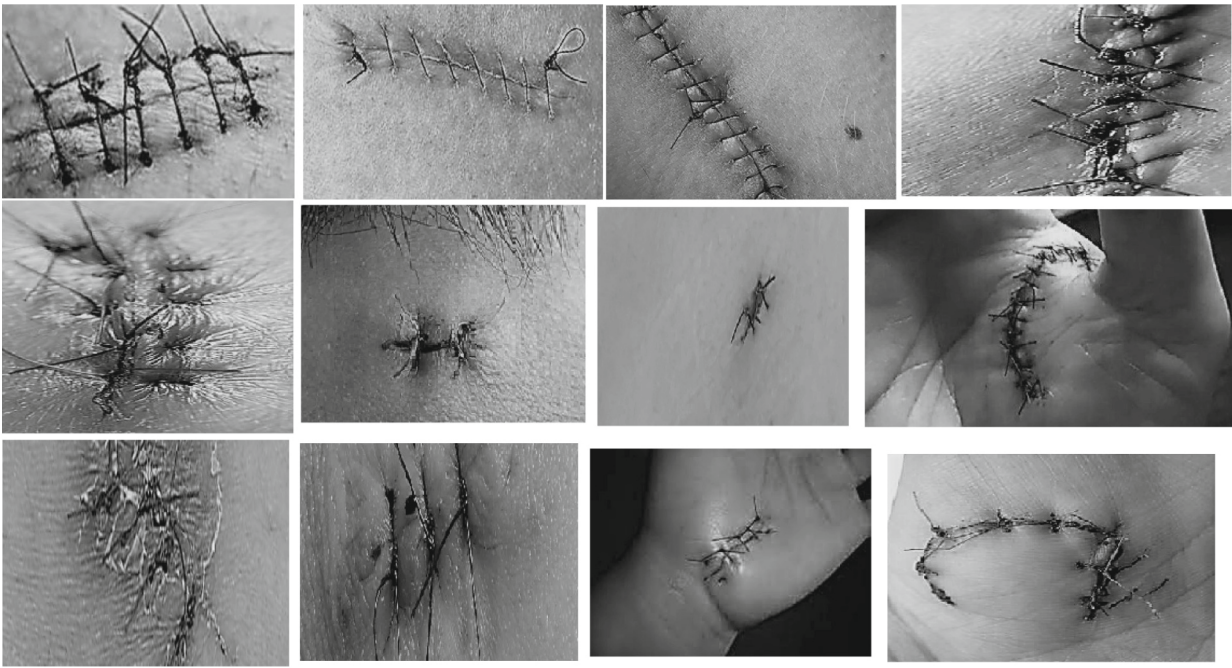


Fig. 5. Successful suture images example.

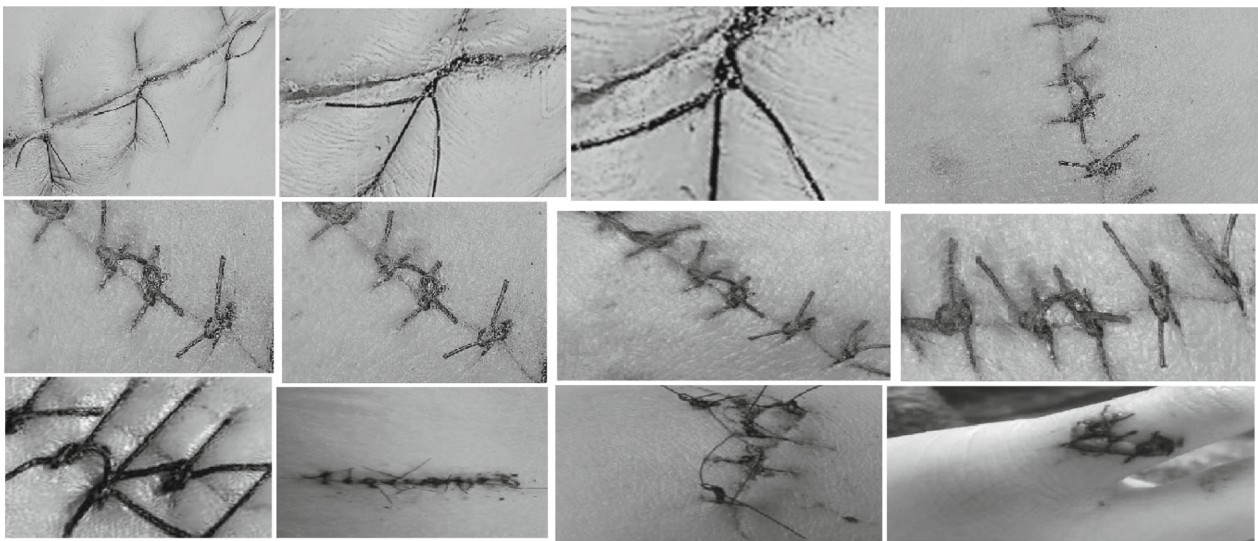


Fig. 6. Failed suture images example.

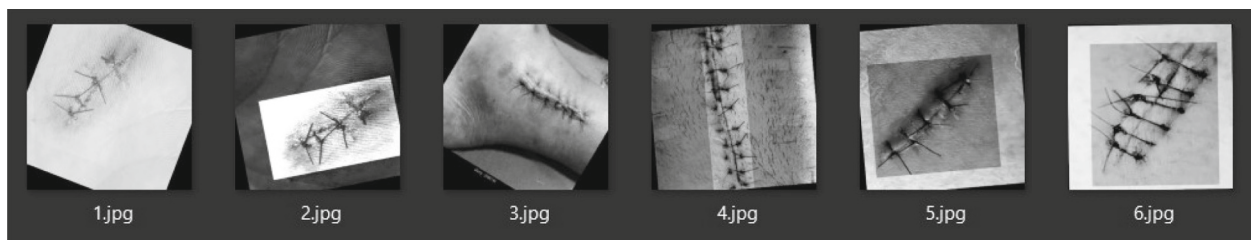


Fig. 7. Data augmentation.

Table 1
Data augmentation methods and parameters.

Method	Probability	Min. factor	Max. factor	Max. left rotation	Min. left rotation
Zoom	0.3	1.1	1.6	–	–
Rotation	0.7	–	–	10	10

Table 2
Parameter setting.

Parameter	Value
Image size	224*224*3
Convolutional and Max Pooling	CNN Transfer learning
Learning rate	0.0001, 0.0002
Epochs	100
Activation function	Sigmoid
Dropout rate	0.5, 0.7

Net, and Dens Net were used to classify a dataset containing two image classes: successful and unsuccessful [17]. The methods and system description are provided in [Methods](#) section. In [Results](#) section, the results are presented and explained. [Discussion](#) section gives the discussion. Finally, conclusions are drawn and the potential future applications are discussed in [Conclusion](#) section.

Methods

The study began by constructing the dataset, which consisted of two types of suture images: successful and unsuccessful. Data augmentation was then performed, and CNN transfer learning models were used and evaluated for classification to determine the best model. The flow chart of the study is shown in [Fig. 3](#).

Data collection

The data was collected from a polymer suture training pad (RTV 2) produced with a mold (see [Fig. 4](#)). In addition, to generate sufficient data for DL, search engine results were also collected and added to the database based on keywords such as “suture,” surgical suture material,” and” suture surface.” Examples of both the successful and unsuccessful

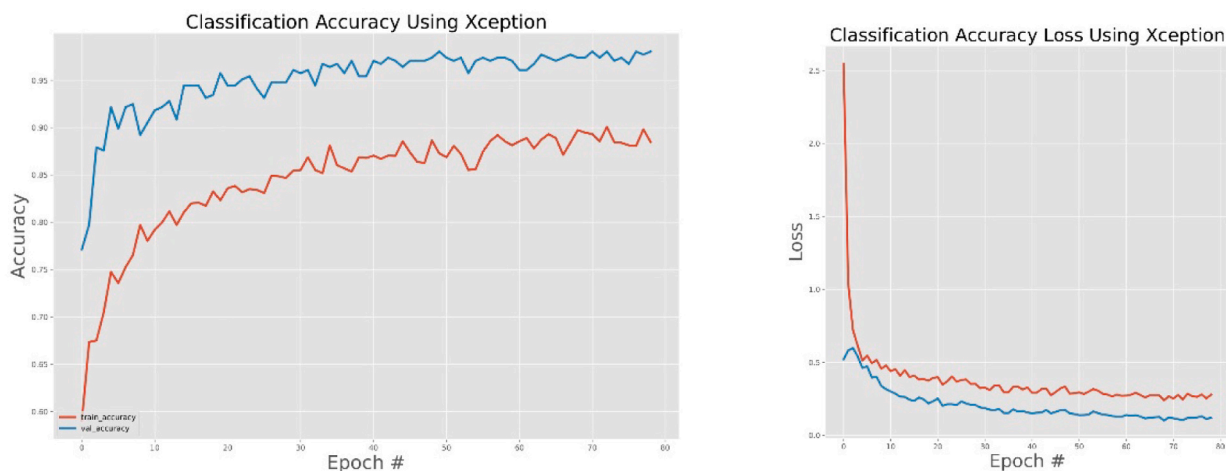


Fig. 8. Xception model training and validation accuracy and accuracy loss.

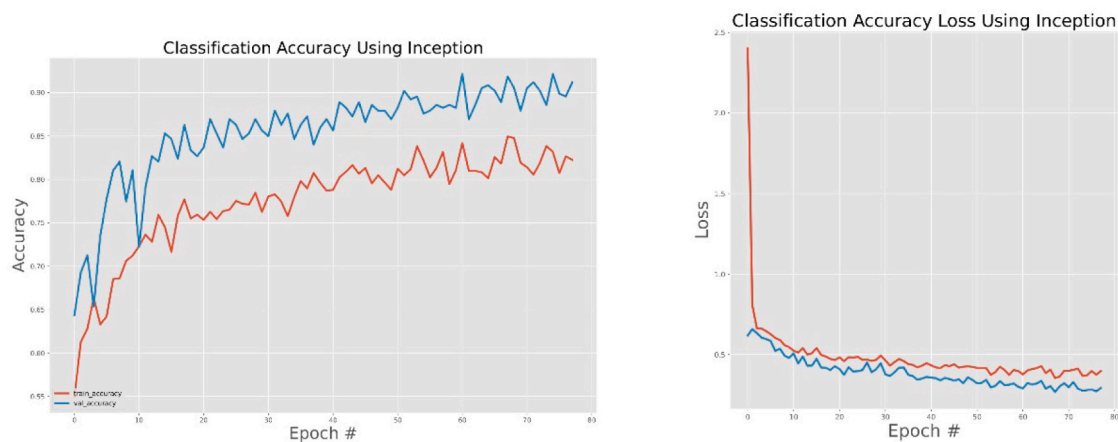


Fig. 9. Inception model training and validation accuracy and accuracy loss.

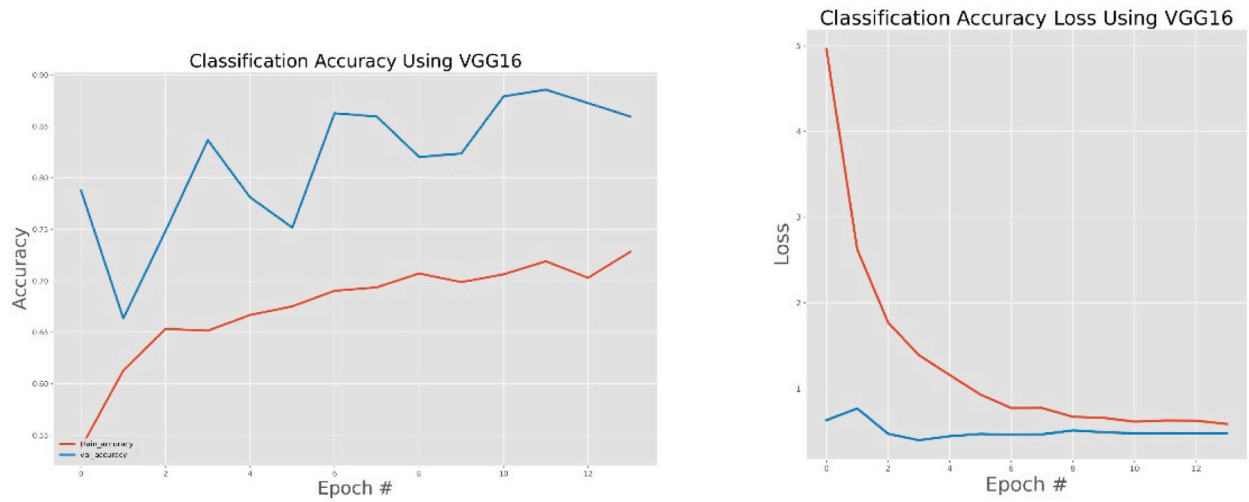


Fig. 10. VGG16 model training and validation accuracy and accuracy loss.

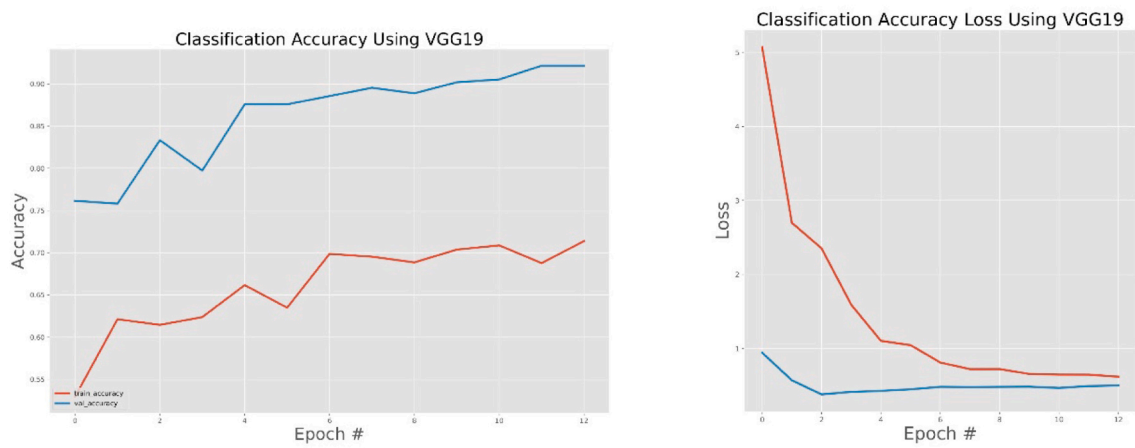


Fig. 11. VGG19 training and validation accuracy and accuracy loss.

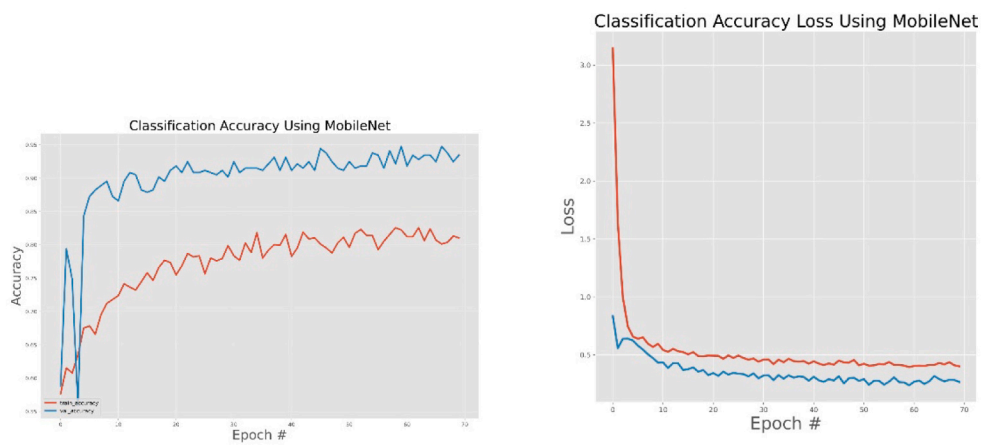


Fig. 12. MobileNet training and validation accuracy and accuracy loss.



Fig. 13. DensNet training and validation accuracy and accuracy.

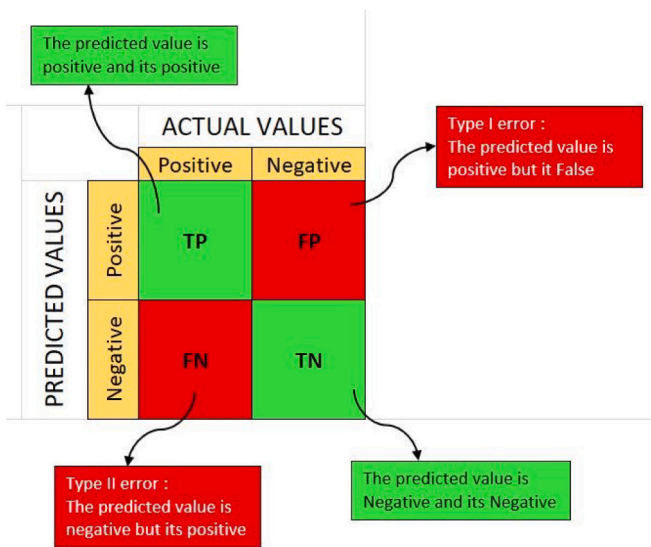


Fig. 14. Confusion matrix basic.

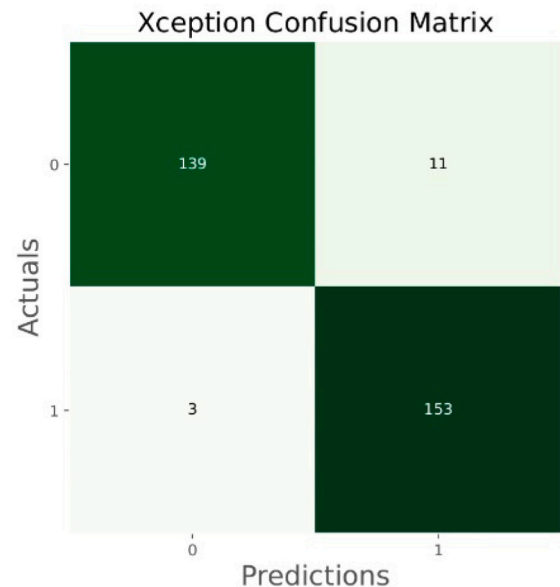


Fig. 15. Confusion matrix of Xception.

class are shown in Fig. 5 and Fig. 6, respectively. All of the collected results were examined by a medical specialist, and a dataset was created by sequentially adding those that were considered suitable for use in the study.

Image processing and data augmentation

To process and analyze the data, a custom Python application was created. The images selected for the dataset were also subjected to a series of data augmentation techniques, which involve generating new data from the existing dataset. The primary purpose of this process is to strengthen the deep neural network through diversification [18,19]. As shown in Fig. 7, data augmentation techniques such as rotation, zooming, cropping, and filtering can be used to increase the number of data (see Table 1) [18,20–23]. The table gives the changes of zoom and rotation few specifications done to the original images where Probability is the probability that a brightness chosen from the region is less than or equal to a given brightness value, the max and min filter effectively follows the smooth edges of the image, and max and min left rotation are counter clockwise rotation of the original images. These specifications of changes were applied to gather new images from the

original dataset and a total of 2005 images were used.

Transfer learning & CNN models

Transfer learning, or learning from related tasks, has gained increasing attention in recent years [24–26]. In this study, we used six popular CNN architectures with pre-trained weights for transfer learning. The suture images were classified as successful or unsuccessful using VGG 16, VGG 19, Xception, Inception, MobileNet, and DenseNet. CNNs are a type of DL algorithm used to process data with spatial or temporal relationships. While CNNs are similar to other neural networks, they incorporate several convolutional layers, which adds an additional level of complexity. A CNN network consists of four main components: the input layer, the convolutional layer, the pooling layer, and the fully connected layer [12,27]. Convolutional layer extracted feature H_i feature map and W_i is the weight, b_i is offset and ρ is the rectified Linear unit (Eq. (1)).

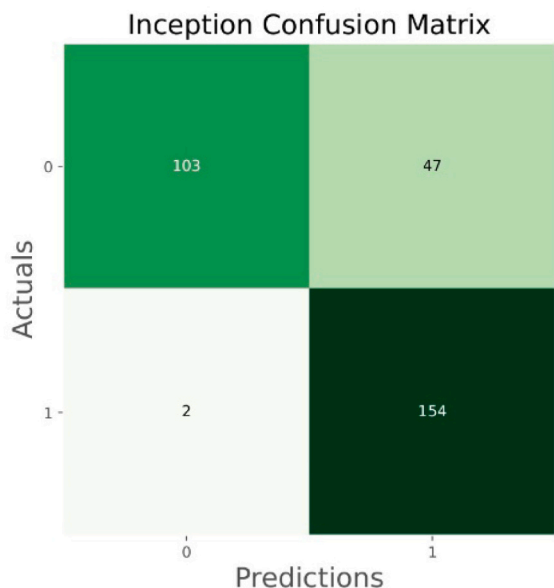


Fig. 16. Confusion matrix of Inception.

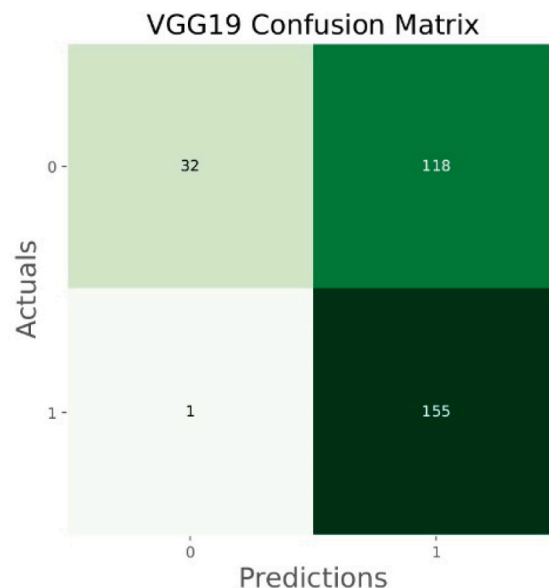


Fig. 18. Confusion matrix of VGG19.

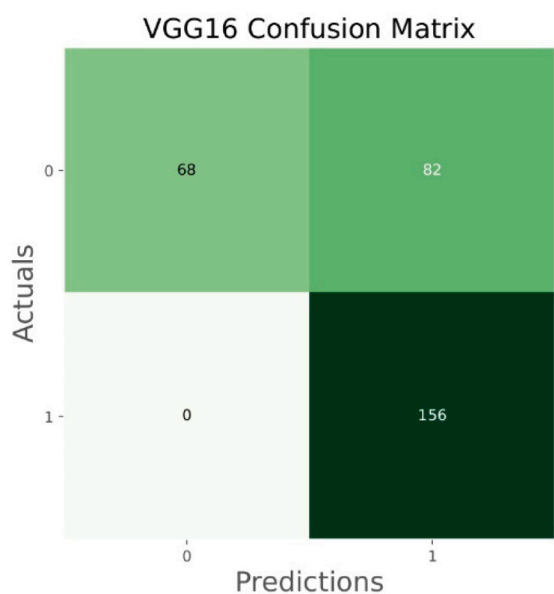


Fig. 17. Confusion matrix of VGG16.

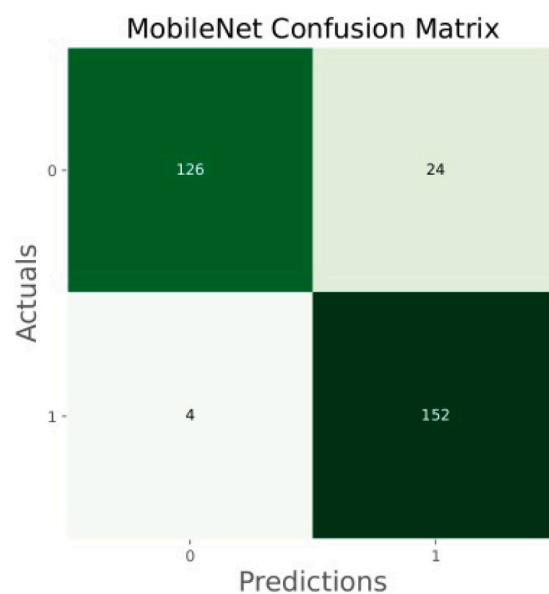


Fig. 19. Confusion matrix of MobileNet.

$$H_i = \rho(H_i W_i + b_i) \tag{1}$$

A crucial part of a CNN is the layer clustering, which reduces the dimensionality of the size of convolved features while also decreasing the computational power needed for image processing. Pooling can be divided into two categories: max pooling and average pooling. While average pooling returns the average value of the image section, max pooling returns the maximum value. Dropout layers can improve the performance of a trained model by reducing the correlation between neurons and preventing overfitting. All activation functions employ the dropout process, although it is scaled by a factor. It collapses the spatial dimensions of the mapped pooled features while preserving the channel

dimensions. When the layer is flattened and converted into a vector, more dimensions are added. The dense layer, also known as the fully connected layer, receives this vectorized feed. Fully connected layers are necessary due to the specific function of extracted image classification features. The sigmoid function makes predictions based on previously extracted image properties from earlier levels, and the sigmoid activation function is used in the output layers for classification into two classes. For the CNN models, the input image size was set to 224×224 . The images were augmented before being processed through the CNN convolutional layer. The CNN architectures were imported directly from the Keras and Tensorflow libraries [28–30] with their trained weights. The layers and filter sizes were used without any modifications. A global

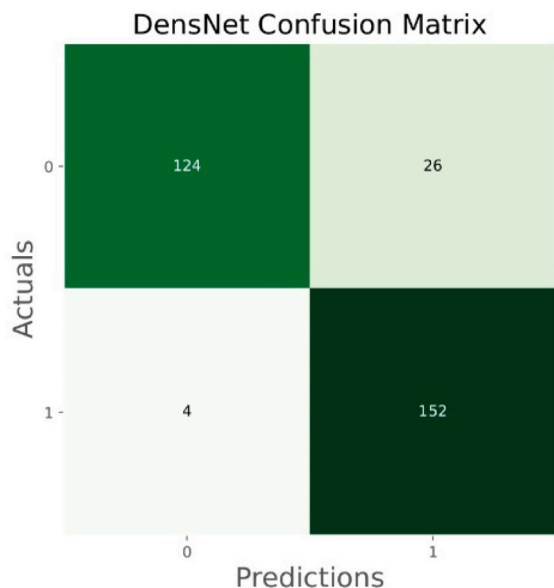


Fig. 20. Confusion matrix of DensNet.

Table 3
Accuracy table.

Algorithm	Accuracy (%)
VGG19	61
VGG16	73
Inception	84
DensNet	90
MobileNet	91
Xception	95

average pooling 2D layer was added to convert the features to a single element vector per image, and the sigmoid activation function was used in the final layer.

CNN models application

The CNN architecture was applied to classify the success of suture

images. The experiments were performed using the Python programming language and its libraries, including Keras, sci-kit, and Tensorflow. Before the classification process, the images were prepared and augmented to improve the accuracy of the model and reduce overfitting. Several hyperparameters, including the number of epochs, hidden layers, hidden nodes, activation functions, dropout, learning rates, and batch size, were used to fine-tune the model. The performance of the model can be influenced by hyperparameter tweaking. The hyperparameters used are shown in Table 2. These hyperparameters represent the specification of DL models applied.

The test-split Python function was used to split the dataset. The dataset included (1220) images for training, (479) images for validation, and (306) images for testing. Six pre-trained CNN algorithms were used. For all CNN models, the input layer used 224 by 224 images, 100 epochs, a batch size of 32, a learning rate of 0.0001, early stopping mode to prevent overfitting, and Adam optimizer. The models performed a max-pooling operation in each pooling layer with a modified pool size and the network's rectified linear unit (RELU) function. The output classes used a sigmoid activation function. The hyperparameters, including the learning rate and epoch size, were modified during the network's training phase. The learning rate was tested at various settings to maximize the targeted performance measurement. The validation procedure was based on the total number of images in the collection. The accuracy improved as different epochs and batch sizes were adjusted.

Results

The effectiveness of the models was assessed and validated using training, testing, and validation techniques. Fig. 8–13 show the training and validation accuracy, as well as the accuracy loss, for the models where accuracy score is the number of correct predictions obtained and loss values are the values indicating the difference from the desired

Table 4
CNN models weighted average assessments.

Algorithm	Precision	Recall	F1 score
VGG19	76	61	54
VGG16	82	73	71
Inception	87	84	84
DensNet	91	90	90
MobileNet	92	91	91
Xception	96	95	95

Accuracy Metrics

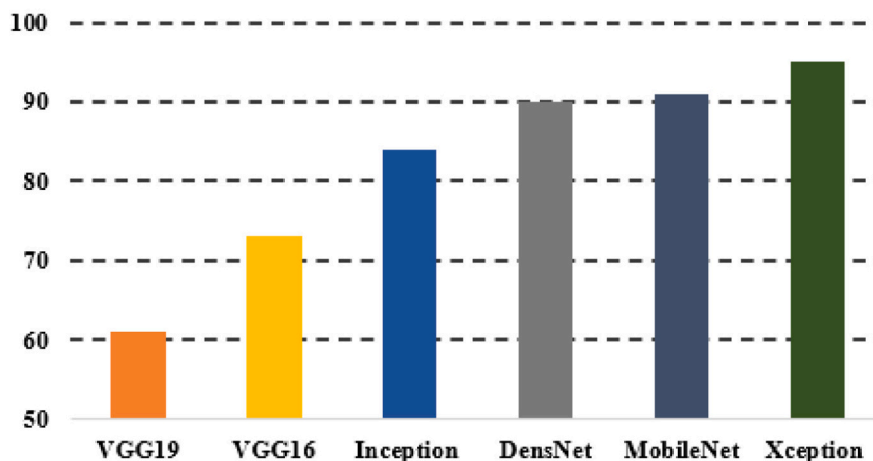


Fig. 21. Accuracy chart.

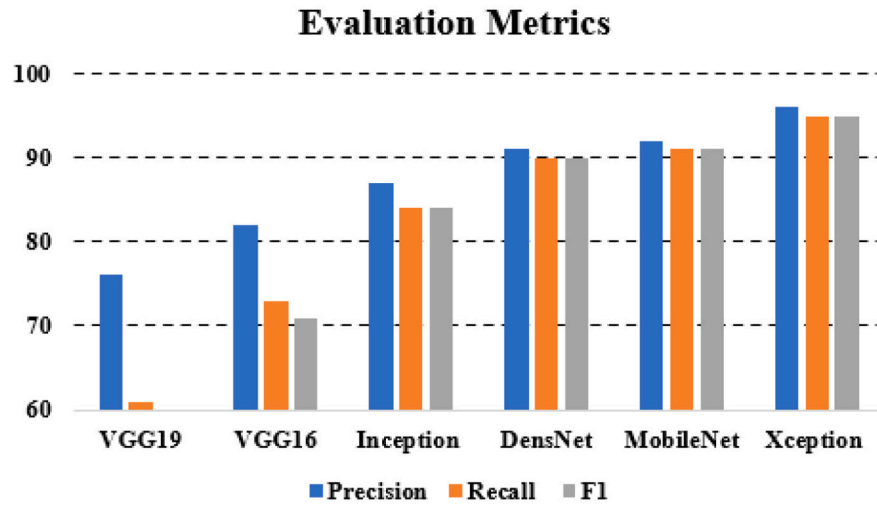


Fig. 22. Evaluation metrics for CNN architectures.



Fig. 23. The graphical user interface.

target states [31]. I.e. the accuracy is the correctness of the model and accuracy loss is the sum of errors made for each example in training and validation sets. Loss value implies how poorly or well a model behaves after each iteration of optimization. The epoch refers to the one entire passing of training data through the algorithm. It's a hyperparameter that determines the process of training the DL model. Each time a dataset passes through an algorithm, it is said to have completed an epoch.

The confusion matrix is a table that is used to define the performance of a classification algorithm (see Fig. 14) [32]. A confusion matrix visualizes and summarizes the performance of a classification algorithm using the test data. Figs. 15–20 show the confusion matrices, which illustrate the predicted classes and their disturbance. The accuracy which is calculated from confusion matrix is defined as follows (Eq. (2)), were TP (observation is predicted positive and is actually positive), TN (observation is predicted negative and is actually negative), FN (refers to the number of predictions where the classifier incorrectly predicts the positive class as negative), and FP (observation is predicted positive and

is actually negative). The accuracy is best when it is close to 100 [33,34,36].

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{2}$$

The general accuracies were 61 % with VGG19, 73 % with VGG16, 84 % with Inception, 90 % with DensNet, 91 % with MobileNet and 95 % with Xception. Table 3 and Fig. 21 present the accuracies of CNN architectures. The Xception algorithm is the most accurate algorithm among the other models. Table 4 and Fig. 22 show the assessment parameters of CNN models. In this study, these primary criteria were precision (Eq. (3)), recall (Eq. (4)), and F1 score (Eq. (5)) for the two predicted classes (success and not success) [37–40]. The results of these equation are evaluated based on confusion matrix as accuracy and they have more success when they are close to 100 [33,34,36].

$$Precision = \frac{TP}{TP + FP} \tag{3}$$

$$\text{Recall} = \frac{TP}{TN + FN} \quad (4)$$

$$F1 = 2 * \frac{\text{Precision} * \text{Recall}}{\text{Precision} + \text{Recall}} \quad (5)$$

QtDesigner were used to design the graphical user interface. PyQt is a Python adaptation of the Qt library written in C++ [41]. The interface includes six DL models using PyQt5 version. The suture images, either taken from the camera or uploaded by the user, are interpreted by DL, and the result information is displayed on the screen. The implemented graphical user interface is shown in Fig. 23.

Discussion

In this study, CNN models were applied to analyze and evaluate suture images. The performance of the CNN models was assessed using several standard metrics, including accuracy, specificity, precision, recall, and F1. Our results showed that the accuracy performance of the studied models was high and statistically significant. In particular, among the different models tested, the Xception model demonstrated the highest accuracy (95 %), precision (96 %), recall (95 %) and F1 measure (95 %). This suggests that the Xception model was particularly effective at correctly identifying and classifying suture images. There have been very less methods proposed in the literature [3–5] for evaluating suture images and suture performance. These methods used computer vision-based software to evaluate the suture performance. In this new study, a novel approach for detecting the success of suture images was presented using CNN models, which achieved an accuracy of 96 %. This new method is simple and can be easily accessed through an application interface. The research indicates that the use of CNN techniques to predict the success of suture photos could potentially be beneficial for medical personnel. The findings of this study have the potential to significantly advance surgical education by reducing errors due to insufficient practice and providing physicians with efficient tools for digitizing the process. By using CNN models to evaluate suture images, the effectiveness of medical training and practice were aimed to be improved. In summary, the goal of this study was to investigate the use of CNN models as a means of evaluating suture images in order to enhance medical education and practice.

The developed graphical user interface also allows for the evaluation of suture images in a user-friendly way, which has not been previously reported in the literature. The findings of this study have contributed new insights into the potential use of DL techniques to improve the efficiency and accuracy of suture training. By leveraging the power of CNN models, we were able to demonstrate high levels of accuracy in detecting the success of suture images. These results suggest that the use of DL approaches could potentially be beneficial for enhancing the effectiveness of suture training programs. Additionally, the approach is simple and easily accessible through an application interface, making it a potentially useful tool for medical professionals looking to improve their suture skills. Overall, the results of this study provide evidence for the potential utility of deep learning in suture training and highlight the need for further research in this area.

Conclusion

A dataset of suture images was collected and compiled for this study, which involved applying CNN models through data augmentation, dataset preprocessing, training, and testing. Compared to other techniques, CNN models have more extensive evaluation metrics parameters. With an accuracy of 95 %, the proposed research is suitable for deployment. Given these considerations, the results of this study, which aims to provide suture training with the help of an application under the guidance of experienced surgeons in medical schools, have the potential to make significant contributions towards reducing the negative impacts

of increased workload.

In future studies, we plan to make the developed system available on mobile platforms to ensure wider accessibility. To further evaluate the effectiveness and user-friendliness of the system, user testing will be conducted on a sample of medical students. The results of the user testing will be used to identify any deficiencies in the system and to gather feedback on potential improvements. These findings will be taken into consideration as we work to address any identified deficiencies and enhance the overall performance of the system.

Funding sources

Not applicable.

Ethical approval

Not applicable.

CRedit authorship contribution statement

Mohammed Mansour (M.Sc.): Software preparing and manuscript writing.

Eda Nur Cumak (M.Sc.): Collecting surgical suture image.

Mustafa Kutlu (PhD): Manuscript writing.

Shekhar Mahmud (PhD): Manuscript writing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

- [1] Zeynep Solakoglu. Evaluating the educational gains of the 6th year medical students on injection and surgical suture practices. *J Ist Faculty Med* 2014;77:1–7.
- [2] George Stockman, Shapiro Linda G. *Computer vision*. Prentice Hall; 2001. PTR 285.
- [3] Irfan Kil, Anand Jagannathan, Singapogu Ravikiran B, Groff Richard E. Development of computer vision algorithm towards assessment of suturing skill. In: 2017 IEEE EMBS International Conference on Biomedical & Health Informatics (BHI): 29–32. IEEE; 2017.
- [4] Woonjae Choi, Bummo Ahn. A flexible sensor for suture training. *IEEE Robot Autom Lett* 2019;4:4539–46.
- [5] Amir Handelman, Yariv Keshet, Eitan Livny, Refael Barkan, Yoav Nahum, Ronnie Tepper. Evaluation of suturing performance in general surgery and ocular microsurgery by combining computer vision-based software and distributed fiber optic strain sensors: a proof-of-concept. *Int J Comput Assist Radiol Surg* 2020;15: 1359–67.
- [6] Adam Dubrowski, Ravi Sidhu, Jason Park, Heather Carnahan. Quantification of motion characteristics and forces applied to tissues during suturing. *Am J Surg* 2005;190:131–6.
- [7] Aristotelis Dosis, Rajesh Aggarwal, Fernando Bello, et al. Synchronized video and motion analysis for the assessment of procedures in the operating theater. *Arch Surg* 2005;140:293–9.
- [8] Aish Mohammed A, Abu-Naser Samy S, Abu-Jamie Tanseem N. Classification of pepper using deep learning. 2022.
- [9] Koosha Sharifani, Mahyar Amini. Machine learning and deep learning: a review of methods and applications. *World Inf Technol Eng J* 2023;10:3897–904.
- [10] Mohammadreza Iman, Reza Arabnia Hamid, Khaled Rasheed. A review of deep transfer learning and recent advancements. *Technologies* 2023;11:40.
- [11] Kumar N, Gupta M, Gupta D, Tiwari S. Novel deep transfer learning model for COVID-19 patient detection using X-ray chest images. *J Ambient Intell Humaniz Comput* 2023;14:469–78.
- [12] Yandong Li, Zongbo Hao, Hang Lei. Survey of convolutional neural network. *J Comput Appl* 2016;36:2508.
- [13] Neena Aloysius, Geetha M. A review on deep convolutional neural networks. In: 2017 international conference on communication and signal processing (ICCSP): 0588–0592. IEEE; 2017.
- [14] Mohammed Mansour, Mert Suleyman Demirsoy, Mustafa Kutlu. Kidney segmentations using CNN models. *J Smart Syst Res* 2023;4:1–13.
- [15] Bharadiya J. Convolutional neural networks for image classification. *Int J Innov Sci Res Technol* 2023;8:673–7.
- [16] Nabil Sayed Aya, Yassine Himeur, Faycal Bensaali. Deep and transfer learning for building occupancy detection: a review and comparative analysis. *Eng Appl Artif Intell* 2022;115:105254.

- [17] Sanskruti Patel. A comprehensive analysis of Convolutional Neural Network Models. *Int J Adv Sci Eng Technol* 2020;29:771–7.
- [18] Connor Shorten, Khoshgoftaar Taghi M. A survey on image data augmentation for deep learning. *J Big Data* 2019;6:1–48.
- [19] Buket Kaya, Muhammed Onal. COVID-19 Tespiti için Akciger BT Goruntulerinin Bolutlenmesi *Avrupa Bilim ve Teknoloji Dergisi*. 2021:1296–1303.
- [20] Luis Perez, Jason Wang. The effectiveness of data augmentation in image classification using deep learning. *arXiv preprint, arXiv:1712.04621*; 2017.
- [21] Jason Wang, Luis Perez, et al. The effectiveness of data augmentation in image classification using deep learning. *Convolutional Neural Networks Vis Recognit* 2017;11:1–8.
- [22] Agnieszka Mikolajczyk, Michal Grochowski. Data augmentation for improving deep learning in image classification problem. In: 2018 international interdisciplinary PhD workshop (IIPhDW). IEEE; 2018. p. 117–22.
- [23] Xuejie Hao, Liu Lu, Rongjin Yang, Lizeyan Yin, Le Zhang, Xiuhong Li. A review of data augmentation methods of remote sensing image target recognition. *Remote Sens* 2023;15:827.
- [24] Hoo-Chang Shin, Roth Holger R, Mingchen Gao, et al. Deep convolutional neural networks for computer-aided detection: CNN architectures, dataset characteristics and transfer learning. *IEEE Trans Med Imaging* 2016;35:1285–98.
- [25] Mingsheng Long, Zhu Han, Wang Jianmin, Jordan Michael I. Deep transfer learning with joint adaptation networks. In: *International conference on machine learning*: 2208–2217PMLR; 2017.
- [26] Chuan Li, Shaohui Zhang, Yi Qin, Edgar Estupinan. A systematic review of deep transfer learning for machinery fault diagnosis. *Neurocomputing* 2020;407: 121–35.
- [27] Ivars Namatevs. Deep convolutional neural networks: structure, feature extraction and training. *Inf Technol Manag Sci* 2017;20:40–7.
- [28] Ketkar Nikhil, Ketkar Nikhil. Introduction to keras Deep learning with python: a hands-on introduction 2017:97–111.
- [29] Jojo Moolayil, Jojo Moolayil, Suresh John. *Learn Keras for deep neural networks*. Springer; 2019.
- [30] Bo Pang, Erik Nijkamp, Nian Wu Ying. Deep learning with tensorflow: a review. *J Educ Behav Stat* 2020;45:227–48.
- [31] Nathalie Japkowicz, Mohak Shah. *Evaluating learning algorithms: a classification perspective*. Cambridge University Press; 2011.
- [32] Marina Sokolova, Guy Lapalme. A systematic analysis of performance measures for classification tasks. *Inf Process Manag* 2009;45:427–37.
- [33] Yogesh Singh, Kumar Bhatia Pradeep, Omprakash Sangwan. A review of studies on machine learning techniques. *Int J Comput Sci Secur* 2007;1:70–84.
- [34] Salvador Garcia, Alberto Fern'andez, Juli'an Luengo, Francisco Herrera. A study of statistical techniques and performance measures for genetics-based machine learning: accuracy and interpretability. *Soft Comput* 2009;13:959–77.
- [35] Ming Yin, Jennifer Wortman Vaughan, Hanna Wallach. Understanding the effect of accuracy on trust in machine learning models. In: *Proceedings of the 2019 chi conference on human factors in computing systems*; 2019. p. 1–12.
- [36] Jesse Davis, Mark Goadrich. The relationship between Precision-Recall and ROC curves. In: *Proceedings of the 23rd international conference on Machine learning*; 2006. p. 233–40.
- [37] Brendan Juba, Le Hai S. Precision-recall versus accuracy and the role of large data sets. In: *Proceedings of the AAAI conference on artificial intelligence*. 33; 2019. p. 4039–48.
- [38] Wang Ruibo, Jihong Li. Bayes test of precision, recall, and F1 measure for comparison of two natural language processing models. In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*; 2019. p. 4135–45.
- [39] Reda Yacoub, Dustin Axman. Probabilistic extension of precision, recall, and f1 score for more thorough evaluation of classification models. In: *Proceedings of the first workshop on evaluation and comparison of NLP systems*; 2020. p. 79–91.
- [40] Daniel Molkenin. *The book of Qt 4: The art of building Qt applications*. No Starch Press; 2007.