Research Article

Emergency Diagnosis and Triage with Artificial Intelligence in Intracranial Hemorrhage

Kafa içi Kanamalarda Yapay Zeka ile Acil Tanı ve Triaj

Dr. Şiyar Bahadır

Hacettepe University, School of Medicine, Department of Neurosurgery, Ankara, Turkey

Abstract: Acute intracranial hemorrhages, regardless of their type, are pathologies with high mortality and require rapid diagnosis and treatment, however the patient group who will benefit most from early operation is operated later than the patient group with less favorable outcome, because they do not admit with a severe clinical presentation. In this study, we aimed to evaluate a deep learning model that can distinguish the presence of intracranial hemorrhage in a small data set. Material Method: 3 healthy patients and 5 patients with intracranial hemorrhages were randomly seleceted for the study from the qure.ai Cranial CT database. The data set was created with a total of 200 CT cross-section images, 100 of which were hemorrhagic and 100 were healthy, and it was divided into three groups as training, validation and test set. The artificial neural network was trained in the training set and its accuracy was tested in the validation set, the accuracy did not improve after reaching around 80% and the training of the artificial neural network was stopped. Later, this artificial neural network was evaluated in the test set. Results: The deep learning model was run on the test set. Results were as follows; Sensitivity 90.0%, Specificity: 70.0%, Positive Predictive Value: 75.0%, Negative Predictive Value: 87.5% Total Accuracy: 80.0%. The deep learning model made only one false-negative assessment in 20 crosss-sections that it had never seen before. As a result, we think that a deep learning model can produce highly accurate results even if they are trained in a small data set and potentially be used for rapid triage in emergency departments.
Keywords: intracranial hemorrhage, neurosurgery, artificial intelligence, deep learning

Özet: Akut kafa içi kanamalar, hangi türden olursa olsun, mortalitesi yüksek, hızlı tanısı ve tedavisi yüksek önem arzeden patolojilerdir, ancak erken operasyondan en fazla fayda görecek hasta grubu, gürültülü bir tabloyla gelmediği için, fayda görmeyecek olan hasta grubuna göre daha geç opere edilmektedir. Bu çalışmada, küçük bir veri setinde , kafa içi kanamanın varlığını ayırt edebilen bir derin öğrenme modelini değerlendirmeyi amaçladık. Materyal Metod: Çalışmaya qure.ai beyin BT veritabanından, 5 intrakraniyal kanamalı 3 sağlıklı hasta rastgele olarak dahil edildi. 100 adet kanamalı 100 adet de sağlıklı olmak üzere toplamda 200 adet BT kesit görüntüsü ile veri seti oluşturuldu, eğitim, doğrulama ve test seti olarak üçe bölündü. Yapay sinir ağı eğitim setinde eğitilerek doğrulama setinde hassasiyeti test edildi, hassasiyet %80 dolaylarına çıktıktan sonra sabitlendi ve yapay sinir ağının eğitimi durduruldu. Daha sonra bu yapay sinir ağı, test setinde değerlendirildi. Sonuçlar: Derin öğrenme modeli test seti üzerinde çalıştırıldı. Sensitivite %90.0 , Spesifite: %70.0 ,Pozitif Prediktif Değer: %75.0 , Negatif Prediktif Değer: %87.5 Toplam Doğruluk: %80.0 olarak geldi. Derin öğrenme modeli, daha önce hiç görmediği 20 kesitte, yalnızca 1 defa yanlış negatif değerlendirme yaptı. Neticede, bir derin öğrenme modelinin küçük bir veri setinde bile oldukça yüksek doğrulukta sonuçlar çıkarabileceği ve potansiyel olarak acil servislerde hızlı triaj amacıyla kullanılabileceğini düşünmekteyiz.

Anahtar Kelimeler: kafa içi kanama, beyin cerrahisi, yapay zeka, derin öğrenme

Correspondence Address : Şiyar Bahadır

ORCID ID of the author: Ş.B. 0000-0003-2329-9669

Hacettepe University, School of Medicine, Department of Neurosurgery,
Ankara, Turkey

siyarbahadir@gmail.com

Please cite this article in press at: Bahadır S. , Emergency Diagnosis and Triage with Artificial Intelligence in Intracranial Hemorrhage, Journal of Medical Innovation and Technology, 2020; 2 (2):115-120

1.Introduction

Acute intra-cranial hemorrhages, regardless of their type, are pathologies with high mortality that require rapid diagnosis and treatment [12]. In intracranial hemorrhages, the period from admission to the emergency room until the skin incision has a significant effect on life expectancy, such that some authors recommend intervention within the first four hours in acute traumatic bleeding [3]. In a study conducted by Tuntanathin et al., this information was confirmed, moreover, the "unfavorable outcome" group, who would benefit the least from early operation, was put into surgery earlier than the patients who would benefit more; probably due to the severity of their clinical presentation, and the operation time of "favorable outcome" group was prolonged as their initial presentation was not as catastrophic as the "unfavorable outcome" group [4].

Machine learning, known as "Artificial Intelligence", is based on the principle of the computer's self-learning from data without explicit coding, contrary to the usual computer algorithms [5]. In recent years, a sub-branch called artificial neural networks and deep learning has become widespread with the emergence of computers with higher processing power. Artificial neural networks refer to a mathematical approximation function organized similarly to the human nervous system. Machine learning methods have been used in many fields of medicine until now. In diagnosis of tuberculosis6 compression fractures [7] differentiation of hepatic and pulmonary nodules8 pancreas cancer [9] and coronary diseases [10] It has been shown that the algorithms can make estimation with a high accuracy, close to an expert clinician.

Of course, machine learning is neither an alternative nor a replacement for the clinician, but will assist the clinician in clinical triage and decision-making processes. For example, Topol et al., Who established a system to classify brain CTs as "critical", "high importance" and "routine", reported that they made progress in the speed of clinical diagnosis and "early diagnosis of the right patient" [11].

Turkey, in terms of Tomography Scans per 1,000 people, occupies fourth place after the United States, Iceland and Korea according to 2018 OECD data. [12] (Table 1). While obtaining this many CT scans alone is a huge burden on public finances, when the evaluation and reporting process is considered, the dimensions of the high cost will be understood more clearly. Automating the imaging decision and triage evaluation, referring the correct requests to the radiologist's interpretation can prevent an important expense in terms of public finance and increase the quality of the service provided.

United States	271,5
Korea	228,1
Iceland	227,3
Turkey	225.1
Luxembourg	218.5
Greece	213.9
Belgium	201.9
France	195.7
Denmark	184.6
Austria	183.6
Latvia	180.8

Table 1: OECD 2018 - Total number of Brain CTs obtained according to countries.

The aim of this study is to design a simple artificial intelligence algorithm that works with the principle of deep learning without a need for high computational processing power and expensive technology investmens, that can be established locally in small-sized hospitals in remote districts, where there is a tomography machine but not always a Radiologist to interpret, and emergency services where rapid interpretation of brain CT examinations is required. Another aim of our study is to evaluate accuracy of this simple artificial intelligence algorithm trained on a small database.. The training database of the artificial neural network was kept very small in order to meet the low processing power precondition and to ensure local reproducibility. In total 200 brain CT slices from 8 patients were used. We have tested whether artificial intelligence algorithms with minimized false negative results can be developed even with a very small sample size and a low-level computer.

2.Materials and Methods

This study is a retrospective study. The study dataset was created with [13] randomly selected patients from the open source "qure.ai" database, which contains anonymised brain CTs, available online . Since the study was conducted using an anonymised, online and open source data set, ethics committee approval was not sought. As for the inclusion criteria, 500 brain CT's evaluated and patients were randomly assigned to pathological class, and when the pathological class with a predetermined quota for each condition was filled, the remaining CT's of the same pathology were excluded and next pathology was searched for (one patient each for acute subdural, chronic subdural, epidural, parenchymal, subarachnoid classes, and three

patients for the healthy group).

100 hemorrhagic and 100 healthy slices were selected, the criteria for selecting 200 slices was reproducibility of the results with a middle grade desktop computer. After the slices were saved in DICOM format, the necessary DICOM pre-processing operations were performed with the pydicom library of the Python programming language. First, attenuation values in CT (the amount of reduction after the x-ray beam passes tissue) were converted to housefield units. Later, due to the large range of the housefield unit scale, windowing was applied in order to enhance images. During this process, 40 to 80 housefield units were chosen as imaging window; values below were accepted as equivalent to water and values above equivalent to bone. Thus, parenchyma was displayed with the most accurate contrast values.

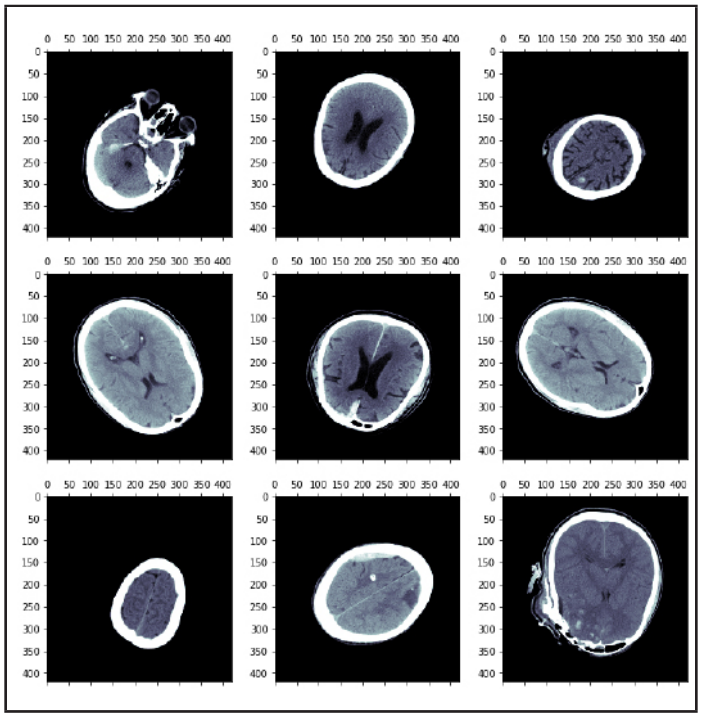

Figure 1: CT sections were added to the data set by changing their orientation to enrich the data set.

Later, in order to improve the quality of the images, sections of image unrelated to the training of the neural network, namely the air outside the skull, and the image of CT table's head rest were deleted to reduce noise in the data. For this, a mask was created to surround the cranium and brain regions in the image, and the area outside this mask was removed from the image. Then the part inside the mask was centered on a black background with a size of 420x420 pixels. Thus, standardization was achieved in training, validation and test images. Finally, an array of size 200, 420, 420 was obtained that stored the cleaned and pre-processed data. Later, due to the large data requirements of artificial neural network traininall slices were included again in the data set by rotating them at different angles, a process calleddata augmentation. (Figure 1)

Then, three separate groups were determined as training, validation and test clusters, randomly. The artificial neural network was trained with the training set, its accuracy was determined during the training with the verification set, and after the training was finished and model was ready, the accuracy of the artificial neural network was tested with the test set. A convolutional neural network architecture was used as an artificial neural network, the neural network was modeled with the help of Tensorflow and Keras libraries.

```
Model: "sequential"
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
conv2d (Conv2D)              (None, 210, 210, 32)      896
_________________________________________________________________
max_pooling2d (MaxPooling2D) (None, 105, 105, 32)      0
_________________________________________________________________
conv2d_1 (Conv2D)            (None, 53, 53, 32)        9248
_________________________________________________________________
max_pooling2d_1 (MaxPooling2 (None, 26, 26, 32)        0
_________________________________________________________________
conv2d_2 (Conv2D)            (None, 13, 13, 64)        18496
_________________________________________________________________
global_average_pooling2d (Gl (None, 64)                0
_________________________________________________________________
dropout (Dropout)            (None, 64)                0
_________________________________________________________________
dense (Dense)                (None, 32)                2080
_________________________________________________________________
dropout_1 (Dropout)          (None, 32)                0
_________________________________________________________________
dense_1 (Dense)              (None, 1)                 33
=================================================================
Total params: 30,753
Trainable params: 30,753
Non-trainable params: 0
```

Figure 2: Deep Learning Model and Structure of Layers

In addition, an accuracy assessment (check_accuracy) function was defined for the model to be able to self-evaluate during its training and to record the structure in which the successful prediction was made, when it made more successful predictions than the previous round. False negative and false positive values were calculated, and with the help of these values, Sensitivity, Specificity, positive predictive value (PPV) and Negative predictive value (NPV) calculations were also embedded in the same function. Binary cross-entorpy method was chosen for model loss calculation, "rmsprop" as optimizer, and accuracy as evaluation metric during training. The 2D convuliton algorithm was used in the intermediate layers of the artificial neural network and "relu (rectified linear unit)" was used as the activation function. Since two results are expected as "bleeding" and "no bleeding" in the output layer, the sigmoid

activation function was used (Figure 2).

A confusion matrix was drawn for diagnostic accuracy assessment. Positive predictive value was calculated from true positive, true negative, false positive and false negative values.

A total of 8 patients were included in the study, 5 of them had signs of bleeding on Brain CT (subdural, subarachnoid, intraparenchymal and epidural), while 3 of them had no evidence of bleeding on Brain CT.

Slices with positive bleeding findings were selected from the brain CTs of patients with bleeding, so that all slices in the dataset labeled as "bleeding present" had sign of bleeding. For the dataset labeled as no bleeding, brain CTs of healthy patients were transferred as they were, and vertex and cervical images were not included in the database in order to meet the 100-slice requirement due to the design of the study. As a result, 100 sections were labeled as "bleeding present" and 100 sections were labeled as "no bleeding".

Since all of the patients were anonymized, no information about their demographic characteristics and clinical conditions could be provided.

3.Results

After the first training of the Artificial Neural Network, it was tested in training, verification and test sets. When the artificial neural network is tested in the training set, values were as follows; True Positive: 71, True Negative: 62, False Positive: 21, False Negative: 8 Sensitivity: 89.8, Specificity: 74.6, Positive Predictive Value: 77.1, Negative Predictive Value: 88.5 Total Accuracy: 82.0%.

However, the evaluations made in the training sets in artificial neural networks are generally not accepted, if the training is carried out in a certain closed set and is tested in that set, because of a phenomenon called "overfitting", which means artificial neural network recognize the samples in the training set at a near perfect rate, and do not recognize the samples outside the training set.

When tested later in the validation set we had similar results; True Positive: 11, True Negative: 6, False Positive: 0, False Negative: 1 Sensitivity: 91.6, Specificity: 100.0, Positive Predictive Value: 100.0, Negative Predictive Value: 85.7 Total Accuracy: 94.4%. However, since the validation set was also used during the training, success in this set was not considered an absolute success of the model. (Figure 3)

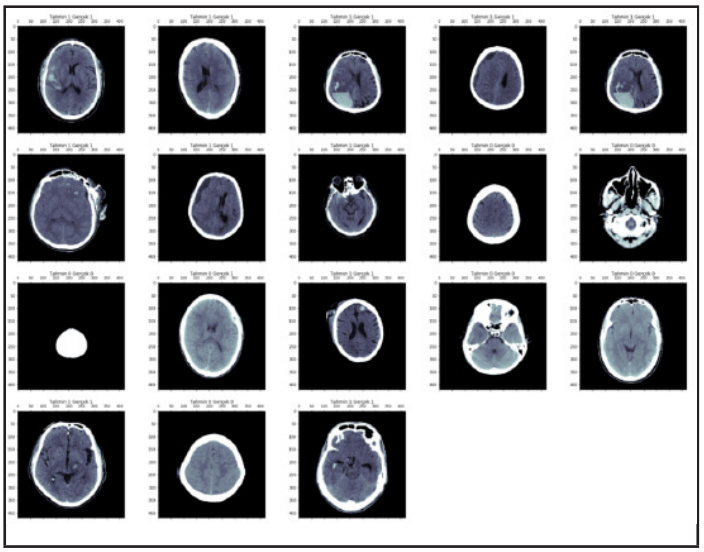

Figure 3: Results obtained in the Validation Set

In the next step, considering that false negative evaluation will cause death, the data set was imbalanced to increase the number of positive (bleeding) sections, and the model was retrained through this new data set, in order to reduce the rate of false negative evaluation.

When tested in the training set again after this training, True Positive: 71, True Negative: 65, False Positive: 18, False Negative: 8 Sensitivity: 89.8, Specificity: 78.3, Positive Predictive Value: 79.7, Negative Predictive Value: 89.0 Total Accuracy: 83.9%. That is, the number of false negatives did not decrease, but the number of false positives was decreased and the number of true negatives increased.

The results did not change when the model trained with the imbalanced set was tested in the validation set.

Finally, in the last stage, when the model is evaluated in the test set, which is the set that the model has never encountered before; True Positive: 9, True Negative: 7, False Positive: 3, False Negative: 1 Sensitivity: 90.0, Specificity: 70.0, Positive Predictive Value: 75.0, Negative Predictive Value: 87.5 Total Accuracy: 80.0%. The model we trained made only one false-negative evaluation in 20 slices that it had never encountered before. When the slice of the wrong evaluation was examined, it was observed that the slice was an epidural hemorrhage with artifacts. (Figure 4)

Since Artificial Neural Networks are nonparametric classification functions, the confidence interval could not be calculated 14 .

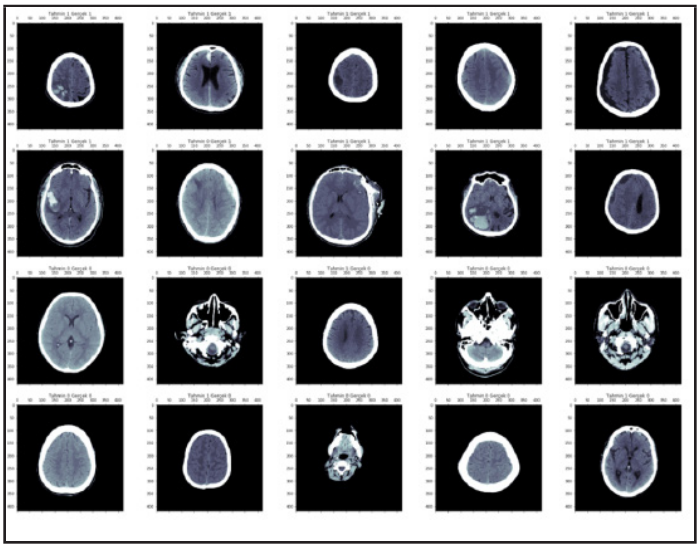

Figure 4: Results Obtained in the Test Set

4.Discussion

The most important limitation of the study is that the database was arranged with slices taken from an open database, without knowing the patients' clinical presentation. In addition, the number of patients added to the database was kept as low as possible, which can be considered as a limitation of the study.

The use of Machine Learning algorithms to distinguish bleeding in Brain CT images has become more frequent with the widespread use of artificial neural networks and deep learning thanks to the increasing computer power. Chilamkurthy et al. [13], achieved more than 90% accuracy in the classification and detection of intracranial hemorrhages, and suggested that artificial intelligence can be used in the triage process. While 21.095 Brain CT scans were used in their study, in our study, only 8 Brain CT scans were selected from the same data set and 200 cross-sections from these 8 scans were included. Although our study is poor in terms of accuracy and depth, it is easily reproducible locally and requires very little computational resources as it is trained in a very small data set.

In 2008, Yuh et al. [15] evaluated traumatic brain injuries with machine learning algorithms in terms of hemorrhage and mass effect, and motion artifacts were identified as the main cause of false positive results. In our study, when the false positive results were examined, it was seen that movement artefacts mainly caused it too.

Hoon Ko et al. In a study by Hoon et al. [16], 4,516,842 brain CT images were evaluated using a previously developed image recognition algorithm called Xception, in 3 different imaging windows (parenchyma, bone and subdural), and classification was made for 5 main bleeding types. A resulting accuracy of 92% was achieved. In our study, the accuracy rate is similar, but the smallness of the data set makes the paradigm used in our study more advantageous in terms of local reproducibility. In our study, the brain CT images are divided into two classes as "bleeding present" and "no bleeding". Although compared to other studies in the literature, [17,18] our study is apparently weaker in terms of detail and depth, it should be considered that our aim to prevent an emergency from being overlooked and help the tirage process.

In conclusion, our study stands out as a rather small and modest study when compared to other studies in the literature. While studies in the literature also have features such as classifiying bleeding subtypes [13,17-19], calculating the amount of bleeding [13,15] and are run in fairly large data sets, our study focused only on distinguishing the presence of bleeding. Separating bleeding subtypes is not possible in a small dataset like the one we used. However, considering the purpose we have defined, it should also be taken into account that it may be sufficient to ensure earlier evaluation by Radiologist and shortening the ER door to operating table interval for patients who may require urgent surgical intervention.

Our source codes [20] can be re-used, and if the artificial neural network used in the study is trained from the beginning it will be observed that similar results can be obtained. By changing the dataset and tag file, the effect of different datasets can be evaluated.

This study might lead to the development of a system that can be used in emergency departments in Turkey. An ideal artificial intelligence-based emergency diagnosis system should be trained in computers with powerful processors on very large databases. After being trained, it should receive information directly from the imaging unit in DICOM format without losing the attenuation data, then it should be able to makes a multi-class classification instead of two classes and finally, it is expected to deliver the result after the evaluation to the doctor quickly (with the help of automatic pager, SMS or computer alerts). If all these conditions are met, we believe that the resulting system will both increase survival and quality of service while reducing costs.

References

1. Alagoz F, Yildirim AE, Sahinoglu M, et al. Traumatic acute subdural hematomas: Analysis of outcomes and predictive factors at a single center. Turk Neurosurg. 2017;27:187-91.

2. Solaroglu I, Kaptanoglu E, Okutan Ö, Beşkonakli E, Taşkin Y. Prognostic value of initial computed tomography findings in patients with traumatic acute subdural hematoma. Turk Neurosurg. 2002.

3. Haselsberger K, Pucher R, Auer LM. Prognosis after acute subdural or epidural haemorrhage. Acta Neurochir (Wien). 1988.

4. Karnjanasavitree W, Phuenpathom N, Tunthanathip T. The optimal operative timing of traumatic intracranial acute subdural hematoma correlated with outcome. Asian J Neurosurg. 2018;13:1158.

5. Koza JR, Bennett FH, Andre D, Keane MA. Automated Design of Both the Topology and Sizing of Analog Electrical Circuits Using Genetic Programming. In: Artificial Intelligence in Design '96. Springer Netherlands; 1996:151-70.

6. Lakhani P, Sundaram B. Deep learning at chest radiography: Automated classification of pulmonary tuberculosis by using convolutional neural networks. Radiology. 2017.

7. Bar A, Wolf L, Amitai OB, Toledano E, Elnekave E. Compression Fractures Detection on CT. June 2017. http://arxiv.org/abs/1706.01671. Accessed February 6, 2020.

8. Yasaka K, Akai H, Abe O, Kiryu S. Deep learning with convolutional neural network for differentiation of liver masses at dynamic contrast-enhanced CT: A preliminary study. Radiology. 2018;286:887-96.

9. Liu F, Xie L, Xia Y, Fishman EK, Yuille AL. Joint Shape Representation and Classification for Detecting PDAC. April 2018. http://arxiv.org/abs/1804.10684. Accessed February 6, 2020.

10. Shadmi R, Mazo V, Bregman-Amitai O, Elnekave E. Fully-convolutional deep-learning based system for coronary calcium score prediction from non-contrast chest CT. In: Proceedings - International Symposium on Biomedical Imaging. Vol 2018-April. IEEE Computer Society; 2018:24-8.

11. Topol EJ. High-performance medicine: the convergence of human and artificial intelligence. Nat Med. 2019;25:44-56.

12. Organisation for Economic Co-operation and Development O. OECD - Computed Tomography (CT) Exams (indicator). Computed tomography (CT) exams (indicator).

13. Chilamkurthy S, Ghosh R, Tanamala S, et al. Deep learning algorithms for detection of critical findings in head CT scans: a retrospective study. Lancet. 2018;392:2388-96.

14. Shafer G, Vovk V. A Tutorial on Conformal Prediction. Vol 9.; 2008.