Soft Computing (2022) 26:665–676
https://doi.org/10.1007/s00500-021-06510-w

DATA ANALYTICS AND MACHINE LEARNING

Deep learning assisted cognitive diagnosis for the D-Riska application

Cristian Cuerda[1] · Alejandro Zornoza[1] · Jose A. Gallud[1] · Ricardo Tesoriero[1] · Dulce Romero Ayuso[2]

Accepted: 19 October 2021 / Published online: 22 November 2021

Abstract

In this article, we expose a system developed that extends the Acquired Brain Injury (ABI) diagnostic application known as D-Riska with an artificial intelligence module that supports the diagnosis of ABI enabling therapists to evaluate patients in an assisted way. The application is in charge of collecting the data of the diagnostic tests of the patients, and due to a multi-class Convolutional Neural Network classifier (CNN), it is capable of making predictions that facilitate the diagnosis and the final score obtained in the test by the patient. To find out the best solution to this problem, different classifiers are used to compare the performance of the proposed model based on various classification metrics. The proposed CNN classifier makes predictions with 93 % of Accuracy, 94 % of Precision, 91 %, of Recall and 92% of F1-Score.

Keywords Acquired brain injury · Cognitive test · Deep learning · Convolutional neural networks · D-Riska

1 Introduction

According to public well-known statistics, Acquired Brain Injury (ABI) represents a serious health problem, mainly due to the large number of people affected (more than 400,000 in Spain), the duration of this type of injury, which is usually chronic, the severity and variety of sequelae. Furthermore, it should be noted that ABI represents the leading cause of disability in adults in developed countries (Cerebral Injury 2014). The causes of ABI are diverse (Cerebral Injury 2014), these include from traumatic brain injury, stroke, anoxia or hypoxia, brain tumors, to encephalitis of various etiologies, among many others. In addition, ABI can affect all areas of human functioning.

The affected area and the deficits presented by affected people depend on the type of injury, the initial location and severity of the injury, as well as the characteristics of the affected people, such as age, personality or skills prior to the injury (Red Menni 2015). This type of injury represents a serious public health problem, both due to the number of affected people and the severity of their injuries, which led us to introduce technological advances into this discipline.

Nowadays, existing technology enables developers to generate several user interface (UI) configurations, where UI components can be offered to users through different devices. This set of interrelated devices, also known as multi-device ecosystems, allows users to interact with the system through different interaction mechanisms that are distributed in physical environments(W3C Community 2018).

This paper presents an application that employs the Distributed User Interface (DUI) (Penalver et al. 2011) paradigm in the D-Riska application (Cuerda et al. 2018) for the evaluation and diagnosis of Acquired Brain Injury. This application is based on the traditional Riska test, which is part of the Loewenstein Occupational Therapy Cognitive Assessment Battery (LOTCA)(Katz et al. 1989).

The Riska test is performed together with other tests defined in the LOTCA, and consists of a card game (Fig. 1), in which patients should group a total of 18 cards with different shapes and colors making as many groupings as possible. The more grouping patients are able to form, the higher is the score obtained in the test. The higher the score, the lower the degree of patients' possible brain injuries.

The solution we have developed focuses on the first stage of the rehabilitation of brain injury assisting the diagnosis, evaluation and treatment of patients, facilitating the work of therapists and improving the work of patients, that is, on the evaluation of the sequelae produced and the difficulties or deficits they cause. To achieve this goal, we have relied on a traditional evaluation methodology in this area called

✉ Jose A. Gallud
jose.gallud@uclm.es

[1] Universidad de Castilla-La Mancha, Campus universitario s/n, 02071 Albacete, Spain

[2] Universidad de Granada, Facultad de Ciencias de la Salud, Avenida de la Ilustración 60, 18071 Granada, Spain

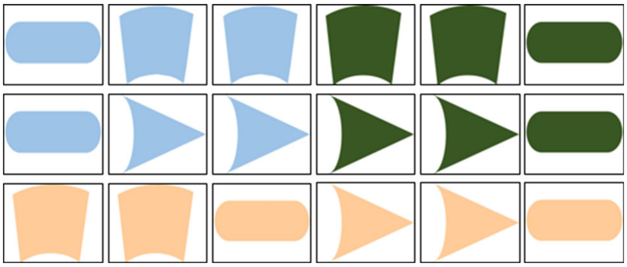


Fig. 1 View of the set of cards that the patient has to classify in D-Riska test

LOTCA, and we have created a digital platform that facilitates the diagnosis and evaluation of the therapist.

Thanks to the use of distributed user interfaces, our application allows the patient and therapist to work simultaneously, each one on a different screen, saving us from using the rest of the devices necessary for the test. The application is responsible for storing all the information related to the session, and allows the therapist to focus on observing how the patient performs the test to facilitate and improve their diagnosis. In addition, it allows this observation to be less invasive,since it is not necessary even for them to be in the same physical space (Cuerda et al. 2020).

The decision regarding patients' final diagnosis by therapists only depends on therapists knowledge and interpretation. However, it is considered essential to endow this type of applications with the ability to assist or support therapists' final decision. this paper presents an expert module based on Artificial Intelligence (AI) using techniques stand out in the current state of the art to reach this goal.

Various techniques are considered to tackle this problem in order to achieve a great capacity for success in the model interpretation.

The system consists of two separated physical UIs which are synchronized. While therapists operate the *Therapist UI*; patients perform assessments on the *Patient UI*.

1.1 Therapist UI

It enables therapists to conduct and analyze the assessment process. Therefore, it enables therapists to introduce assessment session information (i.e. patient personal information and condition, therapist personal information, etc. as well as to have full control of the assessment process in real-time, since both UI are connected and synchronized. For instance, therapist can control patients' UI during the assessment process).

1.2 Patient UI

It enables therapist to evaluate patients' condition. Therefore, this UI enables patients to grouping cards (instead of physical plastic objects) using a touch screen. In fact, card movements performed by patients are transferred to the *Therapist UI* in real-time in order to observe unexpected patterns or behaviors in patients as soon as possible to intervene in the assessment process as required.

1.3 Paper contribution

This paper proposes the development of an expert module to extend the tool previously developed and evaluated in Cuerda et al. (2018) in order to support patients' final diagnosis automatically. The data sources employed to build the expert module were gathered from the use and testing of this application with real patients at the Centro de día Los Tulipanes (day center) and Residencia Hermanitas de los Pobres (residence) in Granada, Spain, which are detailed in the Sect. 3.2. This data consists of sets of images classified by therapists (experts in the field) which enabled the definition of a supervised image classification problem using models based on artificial vision and specific image classification to build a diagnostic module based on AI. This module assists therapists to evaluate patients distinguishing among different types of groupings patients have formed during the assessment process. Consequently, it assigns a candidate score which reduces the probability of errors in the diagnosis process.

This module is not intended to replace the work done by therapists, it only assists and support their decisions on the evaluation. However, as this tool supports session recording to review the assessment process a posteriori, it enables therapists to move their attention to other aspects, such as putting down notes while the patient is performing the test, instead of focusing their full attention only to every movement performed by patients. Some of these that the therapist can focus on are the procedure that follows to carry out the test, recheck what it has happened when the session is recorded, and in the case of novel therapists, it provides guidelines for evaluation that facilitate clinical reasoning. In addition, it also facilitates a first self-diagnosis performed by the patients autonomously, and in case of detecting any behavior or abnormal response, schedule an appointment with a professional therapist to carry out a more exhaustive test.

Since the proposed application collects and manages tests results using images, they can be reused to train multi-class classification models in Convolutional Neural Network classifier (CNN), (LeCun et al. 2015) to predict the probability of the grouping class associated to an existing grouping defined in the test itself.

The proposed expert module provides:

- A greater probability of success at diagnosis time.
- A reduction in the average diagnosis time.
- A self-performing and open-source diagnostic tool.

- Enables the therapists to focus their attention on other fundamental aspects of the evaluation.

To reduce the development time, a set of existing convolutional neural networks presented as part of the state of the art in Donahue et al. (2014) is used as a stating point to implement the expert module instead of defining them from the scratch. Due to the high demand of resources required to train these networks, the architecture infrastructure is supported by Cloud Computing (CC) services which provide high levels of computing and storage capacity.

In general, the contribution of this work can be summarized as follows:

- Presents a DUI system which employs an AI module that assists therapists in the diagnosis of patients with ABI; which is migrated to a CC service provider.
- Collects and stores information obtained from tests carried out on real patients to generate sets of valid data to train and validate the system model, while providing a reliable persistence and management of this information.
- Presents a classifier that infers the class of a grouping and its probability of membership.
- Validates and compares the performance of the a set of models based on the most representative and descriptive metrics in the state of the art.

1.4 Organization of the paper

This article is structured as follows. Section 2 describes the state of the art related to this research area. In addition, while section 3 details the proposed solution, the evaluation and validation process carried out is presented in section 4. Finally, section 5 summarizes the conclusions and future works of this work.

2 Related work

We start providing an overview of the traditional Loewenstein Occupational Therapy Cognitive Assessment Battery (LOTCA)(Katz et al. 1989) evaluation process, which inspired the design and implementation of the D-Riska application detailed in Cuerda et al. (2018). The LOTCA was developed as a technique to assess basic cognitive skills and visual perceptions in adults with neurological disabilities. Provides an in-depth assessment of basic cognitive skills that can be used for treatment planning as well as for treatment progress reviews (McDermott 2011).

The LOTCA battery assess the basic cognitive skills required for daily functions including: orientation, visual perception and psychomotor skills, problem solving skills, and thought operations. The development of this battery is based on information obtained from clinical and neuropsychological experience and development theories. Moreover, LOTCA is generally used at the initial stage of patient evaluation; however, it can be used to set therapeutic goals and review the cognitive status of patients over time (AnnesG and KatzN 1996; Zwecker et al. 2002).

Regarding people's fitness, LOTCA can be used with patients who suffered a stroke, elderly people with dementia, patients with aphasia, people with traumatic brain injuries, or people with intellectual disabilities (JangY and Chern 2009) and mental illness (JosmanN 2006).

The D-Riska application is based on one of the tests proposed by the LOTCA battery, known as Riska. In the traditional Riska test, 18 plastic objects of 3 different colors and 3 different shapes are delivered to patients. The classification of Riska objects consists of two sub-tests. While in the *Unstructured sub-test* therapists ask patients to form groupings of objects spontaneously; in the *structured sub-test* therapists ask patient to form groupings of objects according to a class following a given pattern which is presented as an example.

Thus, the D-Riska application enables therapist to carry out patients' assessment process in a similar way to the traditional one while providing the advantages that digital technologies introduce in the process(Cuerda et al. 2018).

The development of the expert module integrated into the D-Riska application supported by a CC architecture is described in detail in Sect. 3.1.

Internet-based CC is the most powerful type of architecture in the area of computing. It consists of a compilation of hardware, software and infrastructure integrated and available on the network. This type of architecture has several advantages over grid computing and other types of computing configurations and infrastructures (Srivastava and Khan 2018). The current literature collects numerous reviews on CC (Sulich et al. 2020), (Singh and Yadav 2019) (Verma and Sharma 2019).

Recently, advances in AI techniques have encouraged the development of intelligent solutions for CC applications. AI methods, such as Artificial Neural Networks (ANN), Deep Learning, fuzzy logic, and evolutionary algorithms have allowed CC paradigms to be improved through their capabilities to extract knowledge from large amounts of real-world data, optimizing even more so its design, performance and safety compared to traditional techniques (Labati et al. 2020).

Cloud Computing service providers such as Google Cloud Platform, Amazon Web Services, Microsoft Azure and IBM Cloud have incorporated the necessary resources for information management and cloud computing. These also include a large battery of cognitive services, most of them based on AI in areas such as computer vision, voice recognition, text analysis, intelligent indexing, among others (Kumar 2016).

As for references to Machine Learning (ML) applications in the field of mental health and, more specifically in the field of Occupational Therapy (OT), there are already great advances and novel approaches.

Firstly, (Khondoker et al. 2016) presents a comparison of ML techniques for classification based on mental medicine studies on different data sources.

Similarly to the approach presented in our proposal, ML techniques are used to guide patients in the evaluation process through a set of tests employing digital screens in Tsoi et al. (2018).

In addition, the relationship of the behaviors between participants, with or without symptoms of dementia, is studied to measure the brain's response time when drawing a simple figure that can be digitized for early detection of the disease.

A review and a compilation of the latest advances in the area of expert systems and artificial intelligence applications in decision-making and consultation is presented in Bennett and Doub (2016).

Moreover, a review of automated techniques for the early diagnosis and classification of some classes of the Alzheimer disease is presented in Tanveer et al. (2020). This review of 165 papers includes the application of various techniques such as Support Vector Machines, Artificial Neural Networks and Deep Learning.

In Benrimoh et al. (2018) Deep Learning is used to solve problems in treatment selection and diagnosis prediction of mental health, starting with depression. It also stands out along the same lines the work presented in Leis et al. (2019), where an attempt to identify the characteristics of this type of diseases in Spain is based on written language patterns; which are decisive in the detection of the most frequent signs of this pathology.

As may be seen, all reviews refer to the application and success of the use of ML techniques in supporting the diagnosis of mental illnesses, either through specific and generic means or obtained from different sources, such as the social media, IoT mechanisms or public data sources available online.

Finally, the work presented in Majumder et al. (2017) describes a method based on Deep Learning techniques determine the paper author personality type through textual analysis (i.e. given a text relating a set of characteristics with the psychological profile of the paper author).

Based on the presented experience, this paper extends the D-Riska application to support decision-making in the RISKA assessment process taking advantage of the use and efficiency of CC technologies and ML techniques.

3 The proposed solution

3.1 System architecture

Figure 2 shows an overview of the application architecture hosted by Amazon Web Services (AWS) CC provider.

The former design of the D-Riska application was based on a client-server architecture where therapists create a personalized sessions for patients. These sessions collect personal information from patients and therapists to related them to the assessment session.

The server part of the application synchronizes the *Patient UI* with the *Therapist UI* that act as the clients in this architecture. In addition, the server is responsible for managing the session information and guiding the assessment session process. The system was implemented using the Node.js environment and operates through a Representational State Transfer (ReST)(Fielding 2000) Application Programming Interface (API) that is connected to a non-relational database (NoSQL) where the information related to the session and

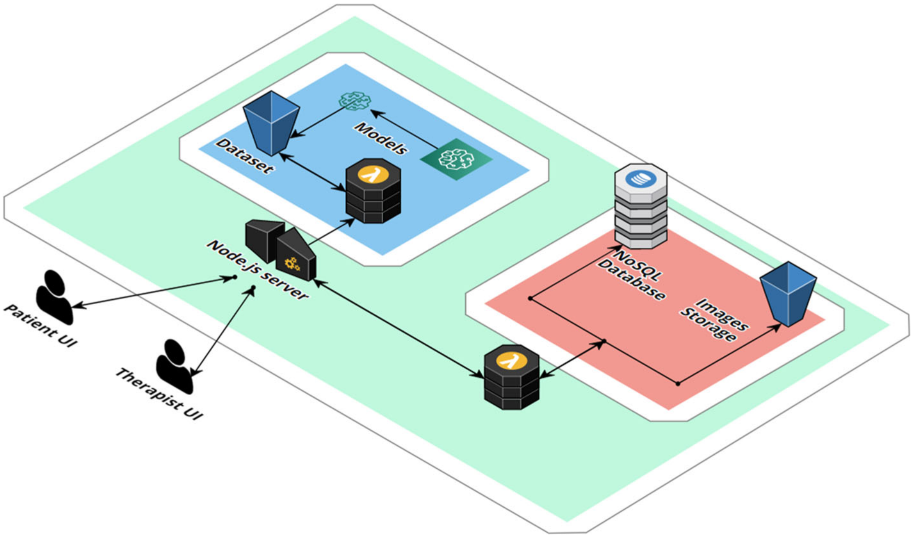


Fig. 2 Overview of the architecture of our application deployed using Amazon Web Services

its participants is stored. The server is not only responsible for collecting the session information entered by therapists, it also collects all the patients' actions through Web Sockets to replicate the patients' actions in the *Therapist UI* in real-time.

In addition, the server is also responsible for storing all assessment session related information in the database, in order to access it later on. The server is also in charge of storing the images gathered during session processes in the image database jointly with the corresponding diagnosis (class).

On the other hand, we have the part of the architecture that corresponds to the module developed in this work that connects to the server in order to use the information generated by the existing D-Riska application.

When therapists evaluate a patient session, they determine in which of the 5 possible categories the card grouping fits, storing the image and its class relationship in the system dataset.

A scheduled retraining of the model is performed every week in case new images were included. Thus, the model is constantly learning. Every time a therapist evaluates a new patient, and adds a new categorized image to the dataset; it is included in the model training set.

At the same time, when therapists perform patients diagnosis, they are able to can visualize the response of the model including the type of category, or class, the image resulting of the test belongs to. It is this moment when the model analyzes the image related to the test, and assigns it to a class based on previous learning. If the prediction is wrong, when the therapist corrects it, the system stores the image with the correct class and a new training launch to match the characteristics of the image to its corresponding class.

3.2 Data collection

The D-Riska application we have used as a starting point manages different types of information. On the one hand, it stores a log of sessions carried out by pairs of patients-therapists where patient and therapist personal information. On the other hand, it stores other session information such as the results associated to each assessment session including the following:

- Duration of the test.
- Observations of the therapist.
- Results of the evaluation by the therapist.
- One image for each test scenario, or grouping, associated to the final layout of the cards.

These images are the ones taken to build the data set that the classification model will use, in its training, validation and evaluation phases (see Sect. 4.1).

3.3 Convolutional neural network

This section presents the Convolutional Neural Networks that were used in the diagnosis and classification of the images belonging to the tests carried out, determining for each of these the class or type of grouping to which they belong to.

The proposed model assigns classes to card groupings in images using a percentage, or value, of matching for each of them. The class associated to an image results from the highest probability of prediction matching. The advantage of this model is that it allows the extraction of significant characteristics from the images and the possibility of making predictions in reasonable periods of time.

The techniques based on Deep Learning are based on Neural Networks which are ML techniques which emphasis is placed on learning through data representation that a series of layers perform aiming to learning more significant representations from gathered data. This technique is based on neurology concepts based on in the brain synapse process. However, this representation drawn from biology does not imply that convolutional neural networks are an exact model representations of the human brain.

The term Deep does not refer to any kind of deep understanding achieved by the model; it is based on the idea of successive layers of representations, which greatly increases the complexity of the model (currently a Deep Learning model can include around tens or hundreds of successive layers and all of them learn through exposure of the training set). Thanks to this complexity, it is possible to learn much more complex and invisible patterns by simpler models, so that Deep Learning must be understood as a mathematical framework to learn data representations.

The basic scheme of a Deep Learning algorithm specifically applied to a classification problem is depicted in Fig. 3 where the network transforms an image into a representation that is incrementally different from the original image and incrementally informative about the final result.

Building a Convolutional Learning network from scratch is a very expensive process that can take a long time to develop, including long periods of trial and error in order to choose the right configuration of layers and hyperparameters that best fits data. In addition, it requires long training periods for large images that can pose a problem in terms of computational performance and training period. In order to solve this problem, there are two main escape routes. The first one is using the computing provided by CC services which provide all the infrastructure, frameworks and environments required to carry out these processes. However, it should be borne in mind that the accommodation, and use of these services, may entail a high cost for the project.

On the other hand, a common, and highly effective solution used in Deep Learning, and specifically in small data sets, as it is the case, is to use pretrained networks, a con-

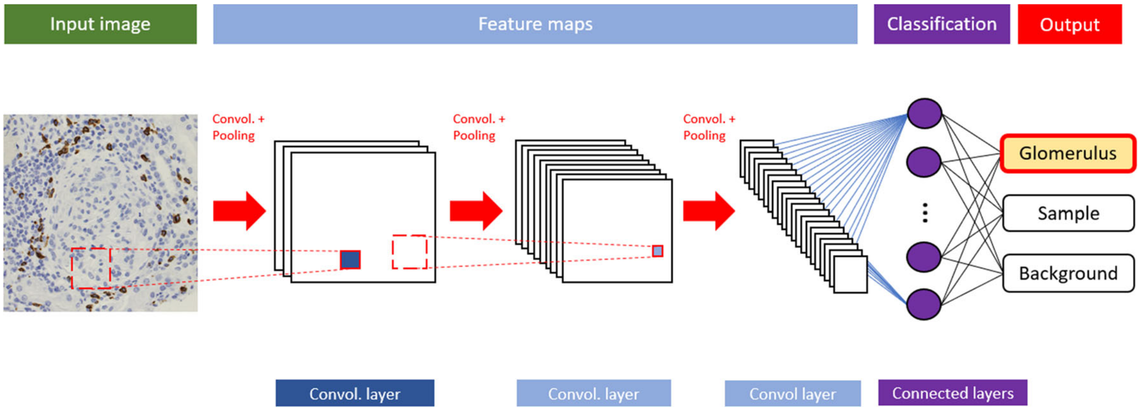


Fig. 3 Example of Deep Learning network working on transforms an image into a representation

cept known as Transfer Learning (Donahue et al. 2014). A pretrained network is a stored network that was previously trained to solve a problem on a high-dimensional data set, most often finding pretrained networks in image classification problems on a large data set. If the data set is large and general enough, the learned characteristics space can be effectively used as a generic world view model, and its variables can provide more than enough knowledge to solve different computer vision problems, keeping in mind that these new problems, like the one provided in this paper, may include totally different classes from the original problem. For example, you can train an ImageNet network (where classes are mostly animals and objects of daily life) and retrain part of that network in order to identify other objects in images. This portability of the features learned about different problems is a key advantage in our project, since it allow us to apply Deep Learning on our small set of images.

There are two different approach to use pretrained networks, feature extraction and fine-tuning. As the second one is not taken into account in this work, only the first one is detailed in next section.

3.3.1 Feature extraction

Feature extraction involves using representations learned from a pretrained network with the goal of extracting interesting features from new examples. These characteristics are taken and processed on a new classifier, which is trained from the start. It is necessary to clarify that this new classifier is not an independent model, such as an Support Vector Machines (SVM) or a Random Forest, this classifier is defined in terms of series of dense layers (fully connected layers) that are incorporated into the pretrained network. While the first part of the network (already trained) extracts the most relevant characteristics of the model; the second part becomes the classifier of the model (pending training with our images). It is not required to detail the meaning of all the layers used because they do not increase complexity and focus on the basis of these models.

All these elements are easy to deploy using the framework that stands out in the state of the art Tensorflow (Abadi et al. 2016), currently owned by Google and supported in the Python programming language.

Initially, Tensorflow version 1 was complex framework to use compared to the latest version that includes the Keras library update (*tensorflow.keras*). This library enables the creation of sequential models and the edition its hyperparameters easily. Thus, *tensorflow.keras* defines a module that enables a quick inclusion and deployment of pretrained networks. In addition, to easy pretrained network modifications, this module enables freezing or unfreezing layers (i.e. sets of networks already trained, or trainable) as well as adding or deleting model layers. The following set of networks were used to pre-train the proposed model employing the ImageNet dataset (Deng et al. 2009).

- InceptionResNetV2.
- MobileNet.
- VGG16.
- VGG19.

As mentioned, this framework also provides an easy way to edit layer parameters, such as setting the number of hidden layers, the activation/deactivation of functions, the number of learning steps, and, for each layer, the number of neurons to be considered.

The configuration process is simple, each of the trained models incorporates two dense layers (fully connected layers) which are the only layers that are trained with the images of the training set, since the rest of the layers (dedicated to the extraction of characteristics) are frozen. Finally, a dense layer is incorporated. This layer defines 5 neurons where the output of each of these layers is the output corresponding to each problem class. The function Softmax (see Eq. 1) is used to return the probability for each class and it is used as the activation function of the last layer. The rest of the layers employ the Rectified Linear Unit (ReLU) function (see Eq. 2) to generate their outputs. The following steps summarize the procedure followed the defined networks:

- Load data sets, train, validate and test.
- Reduce of the size of all the sets (resizing of images).
- Perform data augmentation of the training set images (rescaled).
- Add classification layers to feature extraction models.
- Model training using the training set for learning, and the validation set for error measurement.
- Evaluate the predictive capacity on the validation and test, using Accuracy, Precision, Recall, F1-Score, Mathews Correlation Coefficient (MCC), Cohen's Kappa and confusion matrix.

$$Softmax(y)_i = \frac{exp(y_i)}{\sum_j exp(y_j))} \quad (1)$$

$$ReLU(x) = max(0, x) \quad (2)$$

4 Evaluation

4.1 Dataset

After the decision support model is trained, validated and evaluated with data from the available data sources (see Sect. 3.2) using the final images of the finished tests. A total of 860 images, or instances, were processed. The resolution of these images is 1320 x 410 pixels in RGB format. They were arranged in 5 different proportion classes, taking into account that the minority class is the class that corresponds to a random grouping (grouping more susceptible to suffering a pathology).

Since the number of images is too small to address the problem using Deep Learning, it is required to generate extra images to increase the set of these initial images using the data augmentation technique presented in Perez and Wang (2017).

To carry out this task, the initial images resulting from therapists' evaluations were compiled to apply various noise introduction procedures. These procedures perform negligible movements on the cards captured in the image, generating a totally new, and classy instance, similar to the real one. It was exhaustively verified that these movements do not modify the type of grouping to keep the class given by the therapist.

Since the number of images is still small to include in neural networks, the re-scaling technique was employed by Data Augmentation image generators to increase the number of images.

Consequently, the final set of images consists of a total of 9000 images divided into 3 independent sets based on a logical proportion criteria. In 4 we can see a sample with four images includes in our set of images, for four different possible groupings of cards (by form, by color, complete series or by form and color)

The training, validation and test sets are presented in Table 1.

Table 1 Dataset Description

Dataset name	ISE-Driska	
Resolution	1320 x 410	
Num. Instances	9000	
Num. Classes	5	
Proportion classes	Unbalanced	
Set	Dimension	Proportion
Training	8000	89%
Validation	700	8%
Test	300	3%

4.2 Metrics

To evaluate and understand the results obtained by each network, and to be able to easily select the classifier that best suits our problem, we have used the indexes in the domain of multi-class classification presented in Tharwat (2018).

The problem to tackle presents an unbalanced case (i.e. there are different proportions of groups) because, in reality, the proportion of people who is ill or suffer related problems (i.e. those who carry out a"random"grouping) is much lower than the rest. Therefore, it is necessary to take into account more complete indices to provide a more specific vision of the discriminatory power of the model, and how it classifies on all cases according to each type of grouping in particular (Hossin 2015). In multi-class classification problems like this, the commonly selected indices are: Accuracy, Precision, Recall, MCC, Kappa and F1-Score (Sokolova and Lapalme 2009).

To evaluate the performance of the classifier at the individual level of each grouping it is essential to visualize the confusion matrix (Fig. 5) since in multi-class problems it collects, as in binaries, true positives or TP on its diagonal (TP_A for class A, TP_B for class B and TP_C for class C in Fig. 5. An example of a correctly classified image suppose a unit hit sum on the corresponding diagonal.

The rest of the elements that do not correspond to the diagonal suppose cases where the predicted class differs from the real class meaning that there are errors in our classifier ([E_{BA}, E_{CA}] for class A, [E_{AB}, E_{CB}] for class B and [E_{AC},

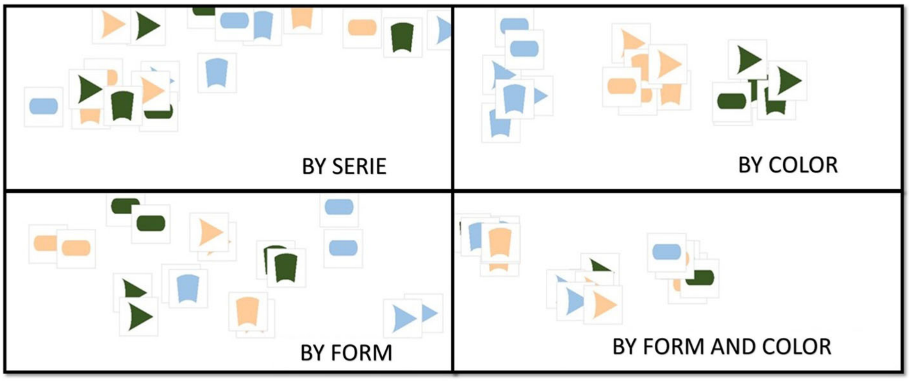
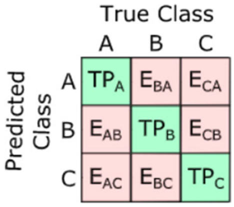


Fig. 4 Example of some images within our dataset corresponding to different groupings of possible cards



	True Class		
Predicted Class	A	B	C
A	TP_A	E_{BA}	E_{CA}
B	E_{AB}	TP_B	E_{CB}
C	E_{AC}	E_{BC}	TP_C

Fig. 5 Example of the confusion matrix for a multi-class classification test

E_{BC}] for class C (see Fig. 5). The ideal assumption is that all values are concentrated on the diagonal, so that when representing this matrix in a heat map format it is easy to observe the success or not of the model, or in which classes it has more discriminative error, using the intensity of color and looking for its location on the corresponding diagonal.

As for model comparison, for each of the models, and each of the indices, their calculations are made from the predictions. Once the values of these indices were obtained, independently for each of the classes, the mean of the corresponding values is selected. This enables models, and their corresponding indices, to be compared at the macro average level. The reason for using these indices is the maximization of the success score for each of the categories or classes.

A simplified description for each of these indices is exposed as follows:

4.2.1 Accuracy (AC)

The Accuracy (AC) is the most used metric to measure classifier performance and represents the percentage of instances classified correctly and is measured in percentage (%). This is a sensitive index in the case of unbalanced problems, since a high success rate can hide flaws as far as minority classes are concerned, so that it must always be as far as possible, accompanied by other metrics. The general equation to calculate this metric is shown in Eq. 3.

$$Accuracy = \frac{TP + TN}{P + N} \tag{3}$$

Where TP are the True Positive values, TN are True Negatives values, FP are the False Positives and FN are the False Negatives values, all of them calculated over all classes.

4.2.2 Precision

The Precision, also known as Positive Prediction Value (PPV) determines the ratio of cases that, within those classified as positive, which really are. This index is beneficial when working with minority classes, which allows us to summarize to some extent the general discriminatory power of our classifier (see general Eq. 4).

$$Precision = \frac{TP}{TP + FP} \tag{4}$$

4.2.3 Recall

The Recall, also known as the True Positive Rate (TPR), is the proportion of positive cases that were correctly identified by the classifier. The equation to calculate this metric is presented in general Eq. 5.

$$Recall = \frac{TP}{TP + FN} \tag{5}$$

4.2.4 F1-Score

The F1-Score determines the relationship between Precision and Recall. Combine those measurements to return a more general quality measurement of the model. The value of the measurement F is a harmonic value, located in the range 0 to 1, being 1 when both Precision and Recall values are around 1, since it calculates the mean between these, so that an imbalance in one of them, it significantly reduces the quality of the model. The general Eq. 6 shows how to calculate this metric.

$$F - Score = 2 * \frac{Precision * Recall}{Precision + Recall} \tag{6}$$

4.2.5 Cohen's Kappa

As for the Cohen's Kappa, it is a statistician who is in charge of measuring the agreement between two classifiers. This score type measure expresses the level of agreement defined in Eq. 7 where p_o is the empirical probability of agreement on the label assigned to any example (by way of agreement ratio), and p_e is the expected agreement when both models randomly assign labels. p_e is estimated using empirical criteria on the different classes. The output is therefore a value between -1 and 1. The maximum value implies complete agreement; a value of 0 or less implies a random or meaningless agreement.

$$K = \frac{p_o - p_e}{1 - p_e} \tag{7}$$

4.2.6 Matthews correlation coefficient (MCC)

The Matthews Correlation Coefficient (MCC) is a measure of the quality of the rankings. This takes true and false positives and negatives into account and is generally interpreted as a balanced measure which can be used even when classes have very different sizes. A coefficient of 1 implies a perfect prediction, a 0 implies a random prediction and a -1 implies a totally inverse prediction. The statistic is also known as the phi coefficient (see Eq. 8.)

$$MCC = \frac{(TP * TN - FP * FN)}{\sqrt{(TP + FP)(TP + FN)(TP + FP)(TN + FN)}} \tag{8}$$

4.3 Results and discussion

Table 2 shows the results obtained for the different CNN classifiers proposed, based on the metrics selected at the macro average level: Precision, Recall, F1-Score, Accuracy and MCC. A classifier that clearly stands out above the rest in all metrics is the VGG16. The results obtained by this classifier exceed in all the metrics analyzed the score of 90 out of 100, which makes it the best solution we have found to solve the problem of multi-classification.

As shown in the table, the success rate obtained is 93 out of 100, that is to say, that out of every 100 times that the therapist selects the help of our tool to find out which category the test result belongs to, employing the VGG16 classifier, we would guess 93 out of the 100 times. The MobileNet classifier obtains results on the threshold of uncertainty, having an approximate hit rate when tossing a coin. However, both the InceptionResNetV2 and the VGG19 fail to adjust to the problem posed, obtaining very pessimistic results that make it unfeasible to use them to solve the problem.

As for the Precision-Recall balance (F1-Score), the VGG16 classifier is again the one with the best results. It implies that despite of the problem being unbalanced, the model correctly predicts minority classes, and especially the class random, which, as recalled, corresponds to the most susceptible cases of disease.

In the case of MCC, it can be seen how the VGG16 approaches remarkably perfectly in the predictions; however, the rest of the models are closer to the random prediction (which can be seen perfectly in the Accuracy of these models).

The matrix Fig. 6 presents the case of the Cohen’s Kappa classifier, which is the one that has the greatest agreement with the real classes is the VGG16 (0.91), followed by the MobileNet network (0.32). It is easy to see that there is also a certain agreement between them since their value is 0.32; however, it is not a high agreement value.

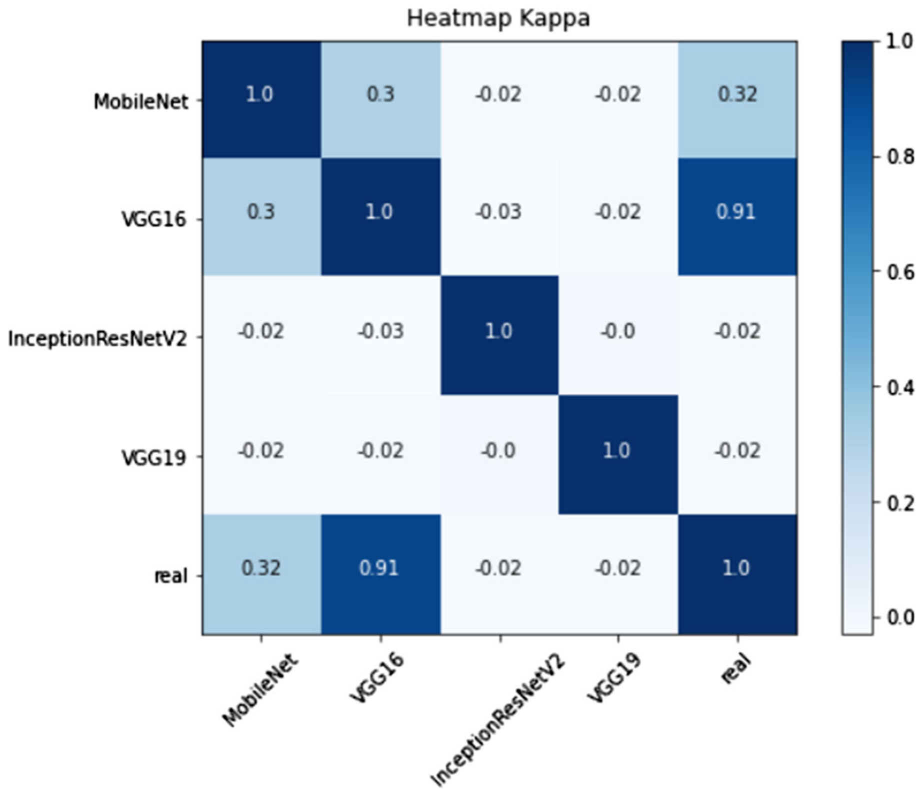
Figure 7 compares these results graphically for the all the metrics. The closer the score is to 100, the better the result of the classifier for that metric. At a glance, the reader can see how the VGG16 classifier is the one that obtains the best results in all metrics with a notable difference compared to the others.

Figure 8 shows the confusion matrix for the classifier that has obtained the best results on the test set, the VGG16. If we analyze this matrix, we obtain several conclusions about the behavior of the VGG16 classifier for this problem, and where it finds the greatest difficulties in prediction. As seen in the first row of the matrix, for the“random”case, the proposed system is always correct, having correctly predicted the 45“random” cases that have been provided. In the second row, for the“color” case, we see that he has been correct 72 times out of the 75 cases that have been provided to him, erroneously classifying two as “shape”and one as“color shape”. In the third row, for the“form” case, we see that it obtains similar results to the“color”case, hitting 72 of the 75 times, erroneously classifying 3 as“color” this time. For the“form color”case, in the fourth row, we see that once again, as in the“random”case, it has a full number of correct answers, correctly classifying the 60 proposed cases. In the last row, for the“complete series”case, it is where we find the most of

Table 2 Performance evaluation using CNN with various classifiers

Classifier	Precision	Recall	F1-Score	Accuracy	MCC
MobileNet	0,47	0,45	0,37	0,46	0.36
VGG16	0,94	0,91	0,92	0,93	0.91
InceptionResNetV2	0,25	0,22	0,2	0,24	-0.02
VGG19	0,18	0,18	0,18	0,19	-0.02

Fig. 6 Head Map for the Cohen's Kappa



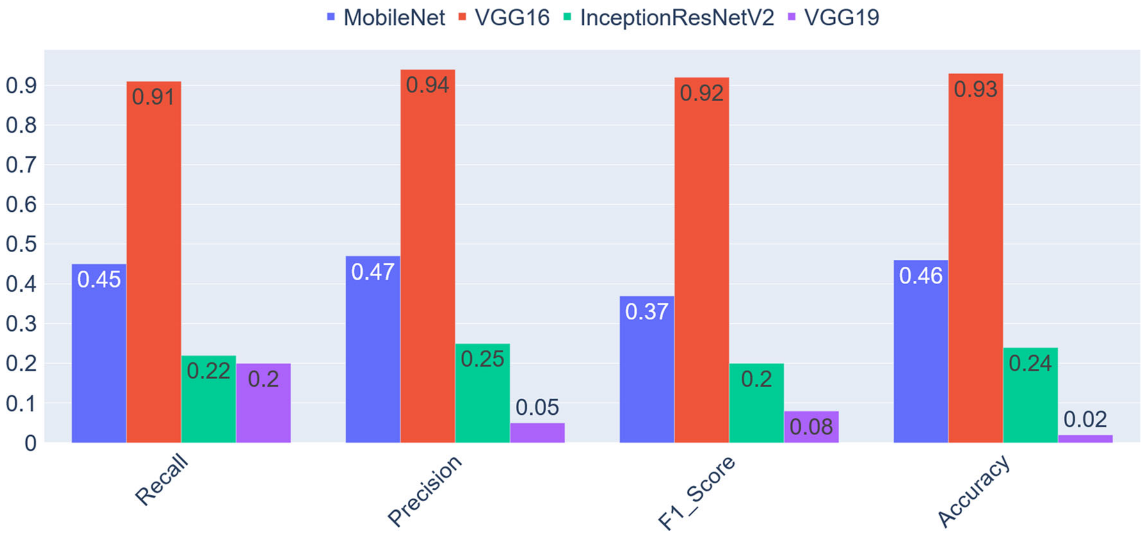


Fig. 7 Comparison of classifiers results in in terms of various measures.

errors in the proposed model. The Recall metric for this case drops to 64 out of 100 indicating that the proposed model is only capable of classifying 64% of"complete series"cases correctly, while the rest, in its vast majority, classifies them as"color form".

This implies that our classifier is sensitive to distinguish between the cases of"color form"and"complete series", that is, there is great uncertainty in the classifier's discrimination power between these two classes; therefore, it may be considered a misleading from the therapist part. However, the results obtained despite the reduced number of images and the possibility that our model improves with the increase in the training set, certifies that it is possible to improve the results. And it is also important to highlight that the tool responds to the therapist correctly in most of cases. It should also be noted that cases of complete series and color form are usually very rare cases in tests performed by patients.

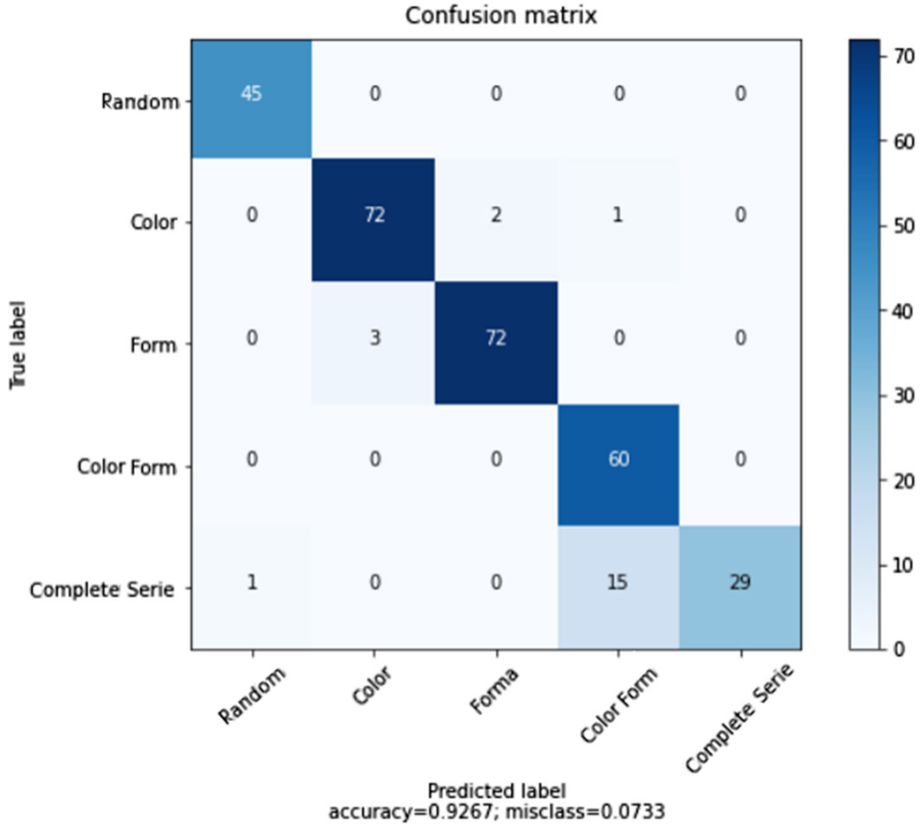


Fig. 8 Confusion matrix for the VGG16 classifier

5 Conclusions

In this article, we present a module that provides an application for the evaluation of Acquired Brain Injury (ABI) in patients with a model based on Artificial Intelligence (AI) that suggests diagnosis decision to therapists. The article briefly presents the D-Riska application, based on the Riska test of the LOTCA battery, which is used as the data provider of the developed AI module.

To achieve this goal, Cloud Computing services were employed to support the client-server architecture of the application. This architecture includes the model that implements the incremental learning process based on the CNN that is in charge of collecting the information resulting from the evaluations carried out by the therapists to learn and improve their predictions, while helping therapists to make decisions about patients' diagnosis providing a prediction based on the training performed.

The selection of the CNN classifier to be use in the system, we have carried out an experimental evaluation where we have compared several trained classifiers with test cases of the problem to solve. Consequently, we have decided that the classifier that best suits this problem is the VGG16, which is capable of making predictions with a 93% hit rate and a 0.92% F1-Score.

Observing the experimental results for this classifier, we assume that they validate the performance of the proposed model in terms of different ranking metrics over other ranking models.

As future works, this approach could be improved by implementing an expert system or a rule-based system that is capable of increasing the Precision the VGG16 classifier makes its predictions for the "complete series"case, which is the one that generates the most errors in order to increase the Precision to a value close to 100%.

Funding Open Access funding provided thanks to the CRUE-CSIC agreement with Springer Nature.

Declarations

Conflict of interest All authors certify that they have no affiliations with or involvement in any organization or entity with any financial interest or non-financial interest in the subject matter or materials discussed in this manuscript.

References

Abadi M, Barham P, Chen J, Chen Z, Davis A, Dean J, Devin M, Ghemawat S, Irving G, Isard M, et al (2016) Tensorflow: a system for large-scale machine learning. In: 12th {USENIX} symposium on operating systems design and implementation ({OSDI} 16), pp. 265–283

AnnesG CS, Katz N (1996) Comparison of younger and older healthy American adults on the Loewenstein occupational therapy cognitive assessment. Occup Ther Int 3:157–173

Bennett CC, Doub TW (2016) Chapter 2-expert systems in mental health care: AI applications in decision-making and consultation, In: DD Luxton (Ed.), Artificial intelligence in behavioral and mental health care, Academic Press, San Diego, pp. 27–51. https://doi.org/10.1016/B978-0-12-420248-1.00002-7. http://www.sciencedirect.com/science/article/pii/B9780124202481000027

Benrimoh D, Fratila R, Israel S, Perlman K, Mirchi N, Desai S, Rosenfeld A, Knappe S, Behrmann J, Rollins C, You RP (2018) T. Aifred Health Team, Aifred health, a deep learning powered clinical decision support system for mental health. In: Escalera S, Weimer M (eds) The NIPS' 17 competition: building intelligent systems. Springer International Publishing, Cham, pp 251–287

Cerebral Injury Association of Navarra (2014). https://www.adacen.org/blog/index.php/que-es-el-dano-cerebral-y-cuales-son-sus-secuelas/

Cuerda C, Romero-Ayuso D, Gallud JA, Tesoriero R, Fardoun HM (2020) Using information and communication technologies to enhance patient rehabilitation research techniques. Methods Inf Med. https://doi.org/10.1055/s-0040-1713086

Cuerda C, Gallud J.A, Tesoriero R, Fardoun H.M (2018) Digital LOTCA-how to evaluate acquired brain injury using technology, pp. 807–813. https://doi.org/10.5220/0006940808070813

Deng J, Dong W, Socher R, Li LJ (2009) Kai Li, Li Fei-Fei ImageNet: a large-scale hierarchical image database. In: Proceedings of the 2009 IEEE conference on computer vision and pattern recognition, pp. 20–21

Donahue J, Jia Y, Vinyals O, Hoffman J, Zhang N, Tzeng E,Darrell T (2014) Decaf: a deep convolutional activation feature for generic visual recognition. In: International conference on machine learning, pp. 647–655

Fielding RT (2000) Architectural styles and the design of network-based software architectures, Ph.D. thesis, University of California, Irvine

Hossin M, Sulaiman MN (2015) A review on evaluation metrics for data classification evaluations. Int J Data Min Knowl Manag Process 5:1–11. https://doi.org/10.5121/ijdkp.2015.5201

Jang Y, Chern JS, Lin KC (2009) Validity of the Loewenstein occupational therapy cognitive assessment in people with intellectual disabilities. Am J Occup Ther 63:414–422

JosmanN KN (2006) Relationships of categorization on tests and daily tasks in patients with schizophrenia, post-stroke patients and healthy controls. Psychiatry Res 141:15–28

Katz N, Itzkovich M, Averbuch S, Elazar B (1989) Loewenstein occupational therapy cognitive assessment (LOTCA) battery for brain-injured patients: reliability and validity. Am J Occup Ther 43:184–192

Khondoker M, Dobson R, Skirrow C, Simmons A, Stahl D (2016) A comparison of machine learning methods for classification using simulation with multiple real data examples from mental health studies. Statistical Methods in Medical Research, 25:1804–1823. https://doi.org/10.1177/0962280213502437

Kumar M (2016) An incorporation of artificial intelligence capabilities in cloud computing. Int J Eng Comput Sci. https://doi.org/10.18535/ijecs/v5i11.63

Labati R, Genovese A, Piuri V, Scotti F, Vishwakarma S (2020) Computational Intelligence in Cloud Computing, pp. 111–127. https://doi.org/10.1007/978-3-030-14350-36

LeCun Y, Bengio Y, Hinton G (2015) Deep learning. Nature 521:436–444

Leis A, Ronzano F, Mayer MA, Furlong LI, Sanz F (2019) Detecting signs of depression in tweets in Spanish: behavioral and linguistic analysis. J Med Internet Res 21:e14199 http://www.jmir.org/2019/6/e14199/. https://doi.org/10.2196/14199

Majumder N, Poria S, Gelbukh A, Cambria E (2017) Deep learning-based document modeling for personality detection from text. IEEE Intell Syst 32:74–79

McDermott A (2011) Loewenstein Occupational Theraphy Cognitive Assessment, Nicol Korner-Bitensky,

Penalver A, López-Espín JJ, Gallud JA, Lazcorreta E, Botella F (2011) Distributed user interfaces: specification of essential properties. Springer, London, pp 13–21. https://doi.org/10.1007/978-1-4471-2271-52

Perez L, Wang J (2017) The effectiveness of data augmentation in image classification using deep learning, arXiv:1712.04621

Red Menni of Cerebral Injury (2015). https://xn--daocerebral-2db.es/papel-de-la-terapia-ocupacional-en-la-rehabilitacion-del-dano-cerebral/

Singh S, Yadav S (2019) Application of cloud computing in healthcare a review. Int J Comput Sci Eng 7:909–914. https://doi.org/10.26438/ijcse/v7i3.909914

Sokolova M, Lapalme G (2009) A systematic analysis of performance measures for classification tasks, Information Processing and Management 45:427–437. https://doi.org/10.1016/j.ipm.2009.03.002. http://www.sciencedirect.com/science/article/pii/S0306457309000259

Srivastava P, Khan RA (2018) Review paper on cloud computing. Int J Adv Res Comput Sci Softw Eng 8:17. 10.23956/ijarcsse.v8i6.711

Sulich A, Zema T, Zema P (2020) Schemes for verification of resources in the cloud: comparison of the cloud technology providers, pp. 145–162. https://doi.org/10.1007/978-3-030-40417-89

Tanveer M, Richhariya B, Khan R, Rashid A, Prasad M, Khanna P, Lin C-T (2020) Machine learning techniques for the diagnosis of Alzheimer's disease: a review. ACM Trans Multimed Comput Commun Appl 16:35. https://doi.org/10.1145/3344998

Tharwat A (2018) Classification assessment methods. Applied Computing and Informatics. https://doi.org/10.1016/j.aci.2018.08.003. http://www.sciencedirect.com/science/article/pii/S2210832718301546

Tsoi KK, Lam MW, Chu CT, Wong MP, Meng HM (2018) Machine learning on drawing behavior for dementia screening, pp. 131–132. https://doi.org/10.1145/3194658.3194659

Verma D, Sharma T (2019) Issues and challenges in cloud computing. IJARCCE 8 188–195:188–195. https://doi.org/10.17148/IJARCCE.2019.8431

W3C Community and Business Groups (2018). https://www.w3.org/community/dui/

Zwecker M, Levenkrohn S, Fleisig Y, Zeilig G, Ohry A, Adunsky A (2002) Mini-mental state examination, cognitive fim instrument, and the Loewenstein occupational therapy cognitive assessment: Relation to functional outcome of stroke patients. Arch Phys Med Rehabil 83:342–5

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.