JOURNAL OF
NEUROIMAGING
THE OFFICIAL JOURNAL OF THE AMERICAN SOCIETY OF NEUROIMAGING

WILEY

A deep learning model for prognosis prediction after intracranial hemorrhage

Amaia Pérez del Barrio[1] | Anna Salut Esteve Domínguez[2] |
Pablo Menéndez Fernández-Miranda[1] | Pablo Sanz Bellón[1] |
David Rodríguez González[2] | Lara Lloret Iglesias[2] | Enrique Marqués Fraguela[3] |
Andrés A. González Mandly[1] | José A. Vega[4,5]

[1]Servicio de Radiodiagnóstico, Hospital Universitario "Marqués de Valdecilla", Santander, Spain

[2]Advanced Computation and e-Science, Instituto de Física de Cantabria (IFCA), Consejo Superior de Investigaciones Científicas (CSIC), Santander, Spain

[3]Servicio de Radiofísica, Hospital Universitario "Marqués de Valdecilla", Santander, Spain

[4]Departamento de Morfología y Biología Celular, Universidad de Oviedo, Oviedo, Spain

[5]Facultad de Ciencias de la Salud, Universidad Autónoma de Chile, Santiago de Chile, Chile

Correspondence
David Rodríguez González, Advanced Computation and e-Science, Instituto de Física de Cantabria (IFCA), Consejo Superior de Investigaciones Científicas (CSIC), Edificio Juán Jordá. Avda. de los Castros s/n, 39005 Santander, Spain.
Email: drodrig@ifca.unican.es

Present address
Amaia Pérez del Barrio, Servicio de Radiodiagnóstico, Hospital Universitario de Navarra, Pamplona, Spain.
Pablo Menéndez Fernández-Miranda, Servicio de Radiodiagnóstico, Clínica Universitaria de Navarra, Pamplona, Spain.

Abstract

Background and Purpose: Intracranial hemorrhage (ICH) is a common life-threatening condition that must be rapidly diagnosed and treated. However, there is still a lack of consensus regarding treatment, driven to some extent by prognostic uncertainty. While several prediction models for ICH detection have already been published, here we present a deep learning predictive model for ICH prognosis.

Methods: We included patients with ICH ($n = 262$), and we trained a custom model for the classification of patients into poor prognosis and good prognosis, using a hybrid input consisting of brain CT images and other clinical variables. We compared it with two other models, one trained with images only (I-model) and the other with tabular data only (D-model).

Results: Our hybrid model achieved an area under the receiver operating characteristic curve (AUC) of .924 (95% confidence interval [CI]: .831-.986), and an accuracy of .861 (95% CI: .760-.960). The I- and D-models achieved an AUC of .763 (95% CI: .622-.902) and .746 (95% CI: .598-.876), respectively.

Conclusions: The proposed hybrid model was able to accurately classify patients into good and poor prognosis. To the best of our knowledge, this is the first ICH prognosis prediction deep learning model. We concluded that deep learning can be applied for prognosis prediction in ICH that could have a great impact on clinical decision-making. Further, hybrid inputs could be a promising technique for deep learning in medical imaging.

KEYWORDS
deep learning, head CT, hybrid, intracranial hemorrhage, medical image, prediction, prognosis

INTRODUCTION

Intracranial hemorrhage (ICH) is a common life-threatening condition affecting over 2 million people worldwide every year.[1] It is defined as the presence of intracranial blood outside the brain vessels and may be spontaneous or traumatic. The bleeding may cause increased intracra-

nial pressure, which can rapidly lead to a fatal outcome. That is why ICH must be rapidly diagnosed and treated.[2,3] The diagnosis is usually performed using head CT, an imaging technique widely available in hospitals and with a high sensitivity for ICH detection.[1] However, regarding treatment options, which includes medical, interventionist, and surgical options, there are still some unsolved problems, such as

the management of external ventricular drains or the indications for early surgical treatment. One of the reasons for this lack of a clear therapeutic protocol is the uncertainty regarding the prognosis, which leads to difficult clinical decision-making, with the responsibility often falling to the individual neurosurgeon.[1,4-7]

In this context, an Artificial Intelligence (AI) model that provides information on the patient's prognosis could be useful. In other diseases, such as cancer[8,9] or coronavirus disease 2019 (COVID-19),[10] deep neural networks have already demonstrated good results in predicting prognosis, opening the door to the use of these systems in daily clinical practice. Regarding ICH, AI research has focused mainly on its detection in CT images,[11-15] not on prognostic prediction.

In this work, we aim to design a deep learning model able to classify patients with ICH according to their prognosis using CT images combined with clinical data as input.

METHODS

Study design

We developed a deep learning model for the classification of patients with ICH into those with poor prognosis and those with good prognosis, using 3-dimensional (3D) brain CT images (scans) combined with clinical data. We adopted this hybrid input given that both images and clinical variables could provide important information regarding prognosis, and that both kinds of data are usually available: on the one hand, all patients with ICH undergo a head CT scan at diagnosis, and on the other hand, the clinical variables included are ones that are usually collected on the patient's admission due to suspected ICH. Data were collected retrospectively.

The methodology was built according to the checklist for artificial intelligence in medical imaging requirements for AI in Medical Imaging.[16]

Ethical approval

This study was approved by the Ethical Committee of our institution, and due to the retrospective nature of the study and the low risk of data leakage, patient consent was waived.

Data: Sources and description

Initially, we included 277 patients diagnosed with ICH from 2010 to 2015, from a database provided by our institution's coding department, which conducted a search with the following criteria: all patients, regardless of their age and sex, admitted to the hospital from 2010 to 2015.

Scans of the included patients were obtained from the Picture Archiving and Communication System, and the clinical data for each patient was manually gathered from the Electronic Health Records (EHR) by three radiologists from our institution.

TABLE 1 Collected categorical and numerical variables

Categorical variables	Numerical variables
Sex	Age (years)
Smoker	Systolic AP
Alcohol	Diastolic AP
Head trauma	Oxygen saturation
Hypertension	Temperature
Diabetes Mellitus	Heart rate
Dyslipidemia	Respiratory frequency
Medical history of intracranial hemorrhage	Glasgow Coma Scale
	Glucose
Medical history of cardiovascular diseases	Creatinine
	Sodium
Medical history of neurologic diseases	Potassium
	White blood cells
Medical history of dementia	Hemoglobin
Medical history of cancer	Platelets
Medical history of hematologic diseases	Mean corpuscular volume
	Red blood cell distribution width
Medical history of other major diseases	Mean corpuscular hemoglobin concentration
Anticoagulant drugs	Mean platelet volume
Antiaggregant drugs	International normalized ratio
Antihypertensive drugs	Fibrinogen
Calcium antagonist drugs	
Alpha-blockers drugs	
Physical exploration with neurological signs and symptoms	

Abbreviation: AP, arterial pressure.

Images

All the scans included in the study were unenhanced sequential head CTs, acquired with a slice thickness ranging from 2.5 to 5.0 mm. None of these images had been previously used for research or published. We have now made them available at DIGITAL.CSIC (https://doi.org/10.20350/digitalCSIC/14706) for research purposes.[17]

Tabular data

Collected clinical variables, both categorical and numerical ($n = 41$), are shown in Table 1.

Ground truth and labels

Two labels were set: poor prognosis and good prognosis. The poor prognosis label was assigned to those patients who did not survive the hospital stay after diagnosis of ICH, whereas the good prognosis label was reserved for those who survived and were discharged. This approach represented the consensus among the neurologists and radiologists of our institution, since it was an objective outcome that was recorded in the EHR, whereas other potential criteria such as the grade of disability were subjective and not always recorded. The mean hospital stay was 17 ($\pm$5) days. Labels were encoded as 0 for good prognosis
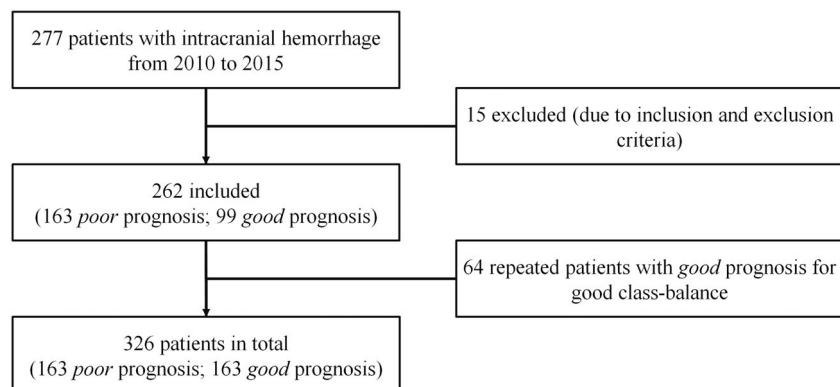
FIGURE 1 Flowchart showing patient selection method

TABLE 2 Demographic data of patients in both groups (poor prognosis and good prognosis groups)

	Poor prognosis	Good prognosis	All patients
Number of patients	163	163 (99 + 64)	326
Age (years)	73.48 ± 6.36	61.52 ± 12.73	67.50 ± 12.73
Lowest age (years) : Highest age (years)	32 : 101	17 : 88	17 : 101
Sex (male : female)	97 : 66	102 : 61	199 : 127

Note: All the data represent mean ± standard deviation unless otherwise indicated.

and 1 for poor prognosis. There was no missing information regarding patient outcome.

Patients' selection and image annotation

Scans were analyzed by two independent radiologists, with more than 3 years of experience and who report head CTs on a daily basis. Patients with significant motion artifacts and/or significant postsurgical changes were excluded ($n = 15$), with no discrepancies between radiologists. Thus, a total of 262 patients were finally included in the study (Figure 1).

Of the total number of patients included, 99 survived the hospital stay and were eventually discharged, while the remaining 163 patients did not survive. Annotation of patients (or images) into good and poor prognosis was performed manually by the three radiologists based on the information they found in the EHR.

Data partition

We partitioned the data at patient level and label balanced. First, we partitioned the data assigning approximately 85% of the patients to the training set, and the remaining 15% to the test set. None of the patients of the test set was included in the training set. Additionally, 20% of the training set was assigned to the validation set. Then, in order to achieve a good label balance, we oversampled patients within each dataset, repeating 64 patients in total (all with good prognosis). We performed this oversampling after data partition so that no patient was repeated within the different sets. After the oversampling, the total number of

TABLE 3 Data details of train and test sets

Details	Train set	Test set
Number of patients	276	50
Ratio of poor prognosis : good prognosis	138 : 138	25 : 25
Age (years)	68.01 ± 4.95	64.68 ± 11.31
Sex (male : female)	162 : 114	37 : 13

Note: All the data represent mean ± standard deviation unless otherwise indicated.

patients increased to 326 (demographic data are shown in Table 2), giving a total of 276 patients assigned to the training set and 50 patients to the test (Table 3).

Data preprocessing

Preprocessing of images

For image preprocessing, we chose a 3D approach, despite the associated high computational cost of working with 3D images, to introduce all the information contained in the scans, such as the correlation between layers, information that is lost in the 2-dimensional approach. We believe that information about hematoma volume and morphology could be of significant interest in predicting prognosis.[5–7]

In the present work, 262 unique DICOM volumes were used, all with an image size of 512 × 512 pixels and a variable number of slices. First, we de-identified them using DICOM Confidential.[18] Then, we rescaled
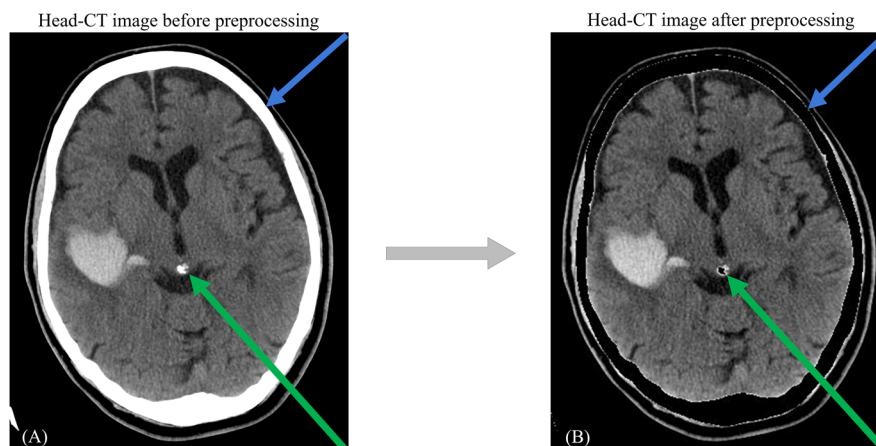
FIGURE 2   An example of a slice of a head CT scan showing an intraparenchymal hematoma located in the right temporoparietal lobe: before (A) and after (B) the preprocessing. Blue arrows point to skull before and after and green arrows point to the calcified pineal gland before and after
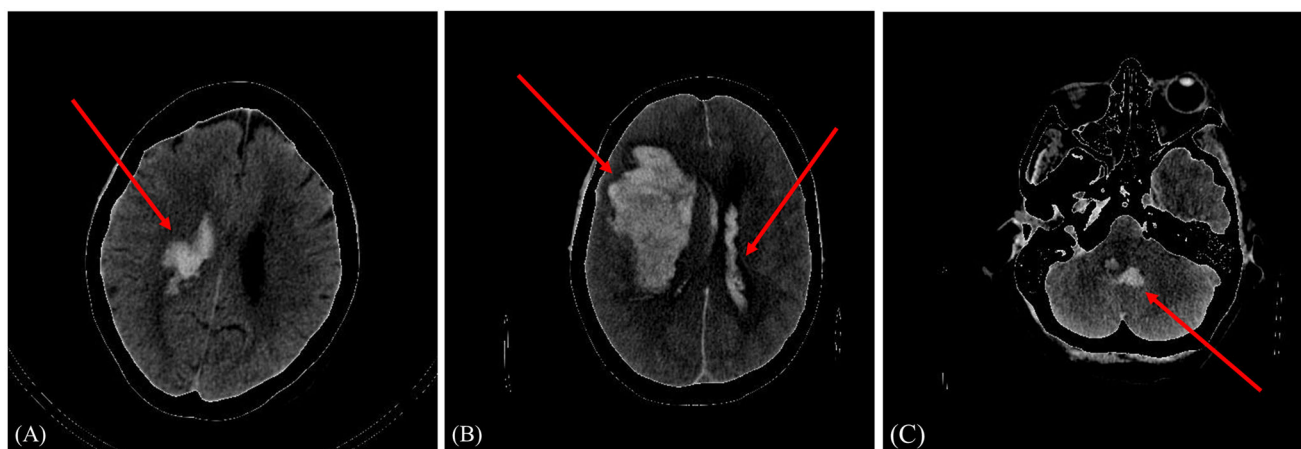

FIGURE 3   Sample of images of the dataset after the preprocessing (red arrows point to the intracranial hemorrhages): a right intraventricular hematoma (A), an intraparenchymal and intraventricular hematoma (B), and an infratentorial hematoma (C)

them to Hounsfield Units (HU), resampled, and applied an HU range between 15 and 100, setting to 0 those pixels with values out of range, such as calcified structures or bone (Figures 2 and 3). Other HU ranges were tried in a prior assessment obtaining the best performance with the 15–100 UH range, which is consistent with prior studies suggesting that the skull can complicate the detection of some ICH.[19] In the last preprocessing step, slices were downsampled to 128×128 and the slice number of scans was matched. To do the latter, since the maximum number of slices was shown to be 45, we filled in the volumes with fewer slices by adding empty slices to the top. We considered this to be the method that least disturbed the image.

Preprocessing of clinical data

Before introducing variables to the model, numerical variables were normalized using the min–max normalization (equation 1) to get num-

bers between 0 and 1, and categorical variables were normalized to 0 (absence) or 1 (presence).[20]

$$x' = \frac{x - \min(x)}{\max(x) - \min(x)} \tag{1}$$

Custom neural network: A model designed from scratch

We designed a custom model based on a 3D convolutional neural network (CNN) for image input and a feed-forward network for clinical data input. CNNs have achieved the best results in artificial vision and medical imaging in recent years,[21] while neural networks have been used with tabular data,[22] showing promising results in areas such as electrocardiogram reading automatization.[23] In the last part of our custom model, the output of the convolutional block and the output of the feed-forward block are concatenated, and this hybrid input is passed to the last dense layer with a sigmoid activation function. Hence, we
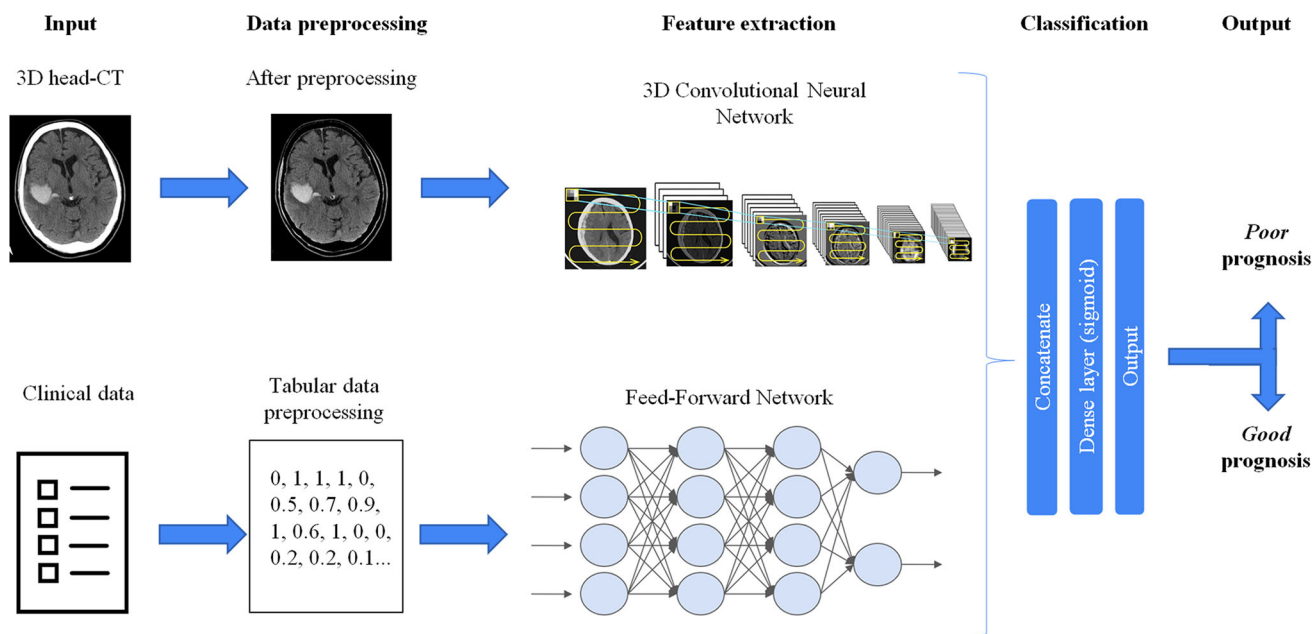
FIGURE 4 Schematic overview of the full architecture of the network
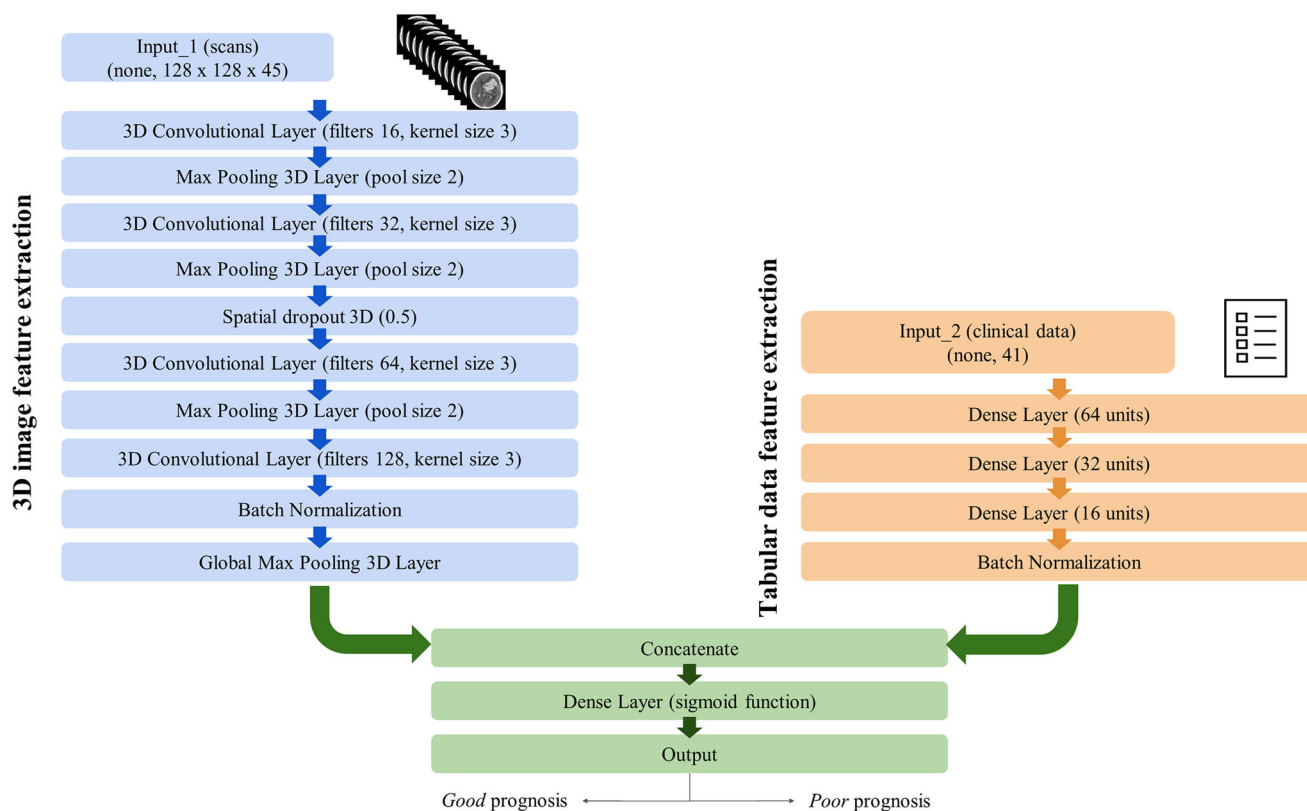
FIGURE 5 Detailed architecture of the hybrid-model

refer to this model as the hybrid model. Figures 4 and 5 show the architecture, and the code is available at: https://github.com/anesdo/ICH_DETECTION_PROGNOSIS_2D3D.git.

Two reference models were also trained, one on image data only (I-model), using the CNN, and the other on tabular data only (D-model), using only the feed-forward network.

TABLE 4 Hyperparameters used in all models

	Hybrid model	I-model	D-model
Parameters initialization	Random (seed = 2)	Random (seed = 2)	Random (seed = 2)
Optimizer	Adagrad	Adagrad	Adagrad
Learning rate	0.008	0.008	0.008
Batch size	6	6	6
Top model	GMP	GMP	GMP
Number of epochs	25	9	13

Abbreviation: GMP, Global Max Pooling.

TABLE 5 Metrics of the different models

	Hybrid model	I-model	D-model
AUC (95% CI)	.924 (.831-.986)	.763 (.622-.902)	.746 (.598-.876)
Accuracy (95% CI)	.861 (.760-.960)	.680 (.540-.80)	.697 (.56-.82)
Specificity (95% CI)	.960 (.862-1.0)	.643 (.44-.838)	.839 (.69-.962)
NPV (95% CI)	.800 (.76-.96)	.694 (.5-.875)	.652 (.468-.813)
F1-score (95% CI)	.843 (.703-.949)	.688 (.522-.826)	.642 (.471-.8)

Abbreviations: AUC, area under the curve; CI, confidence interval; NPV, negative predictive value.

Training and evaluation

Models were trained on the training set (using the validation set to optimize the model hyperparameters) and evaluated on the test set. For model performance assessment, the area under the receiver operating characteristic curve (AUC), accuracy, specificity, negative predictive value, and F1-score metrics were obtained. F1-score combines precision and recall in a single measure.

Software tools and statistical analysis

The training of deep learning models was implemented using Python (v.3.6.3, Python Software Foundation, Delaware, USA, https://www.python.org), and the Nvidia Graphics Processing Unit Titan X, with Tensorflow (v.2.6.2, Google Brain Team, California, USA, https://www.tensorflow.org/) and Keras (v.2.6, Google, California, USA, https://keras.io/) libraries as the backend engine.

Confidence intervals of 95% for all the metrics of the evaluation were calculated by bootstrapping (subsampling the test set) with a bootstrapped sample size of 1000.

RESULTS

All models were trained using the binary cross-entropy loss function and the early-stopping technique to avoid overfitting. The best results were obtained with the hyperparameters shown in Table 4.

The hybrid model, I-model, and D-model achieved an AUC of .924, .763, and .746, respectively. We note that the I-model has more false positives (FP) than false negatives (FN), whereas the D-model shows more FN than FP. The correct balance between FP and FN is achieved with the hybrid model. Further, a statistically significant improvement in specificity is seen in the hybrid model compared to the I-model (from .643 to .960). These metrics are shown in Table 5, while the learning (accuracy) curves and the confusion matrix are shown in Figure 6.

DISCUSSION

In this work, we trained a model that predicts patient survival with an AUC of .924, while being unable to achieve these results using only image features or tabular data features (Figure 7). Comparing the results of the hybrid model and the reference models, a better performance is observed in the model that integrates both the image information (providing the morphology, size, and density of the ICH among others) and the clinical information that is usually gathered by the admission service (first blood test, neurological symptoms at onset, etc.).

To the best our knowledge, this is the first deep learning network to predict ICH prognosis, and could be a step forward in investigating ICH-related poor prognostic factors. Further, we believe that this model could also be used to complement an ICH-detection system, as a second step after the diagnosis, by providing more information about the hemorrhage once it is detected, and therefore helping clinicians' decision-making.

Regarding prognosis prediction in other diseases, most of the research has been focused on cancer. An example of this is DeepProg, a model for cancer prognosis prediction using multi-omics data presented in 2021 by Poirion et al.[8] And, as an example of mixing images with clinical data, Chieregato et al.[10] presented in 2022 a hybrid model trained with images and clinical data from patients with COVID-19 that predicted the severity of the disease with an AUC of .949.
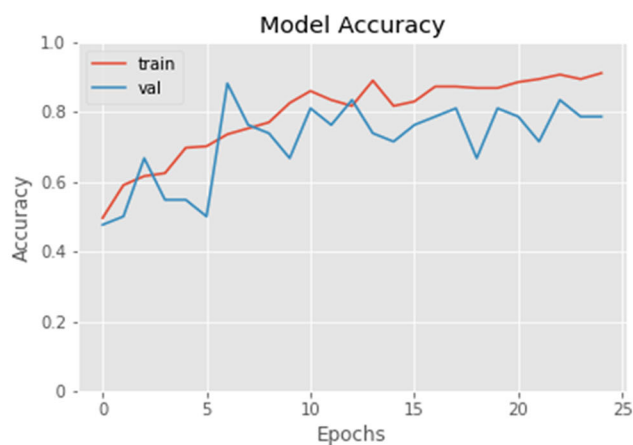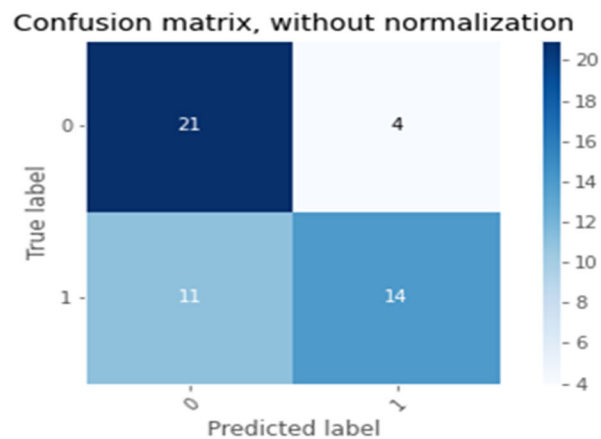
(A) Hybrid-model
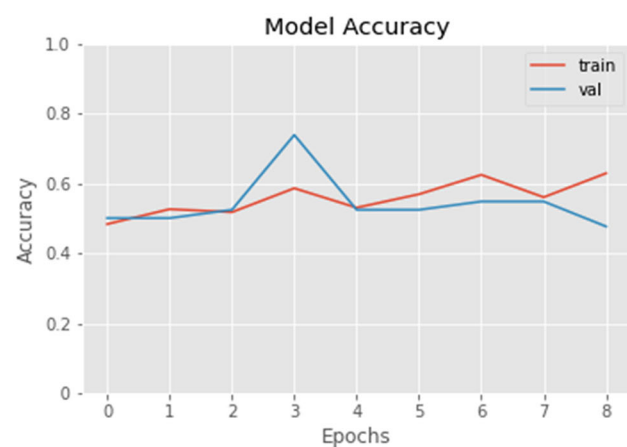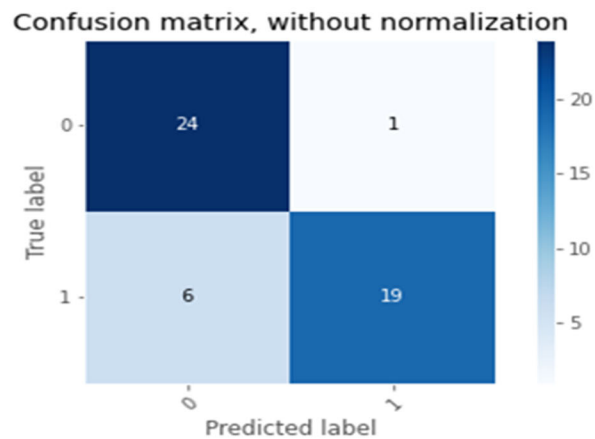
(B) I-model
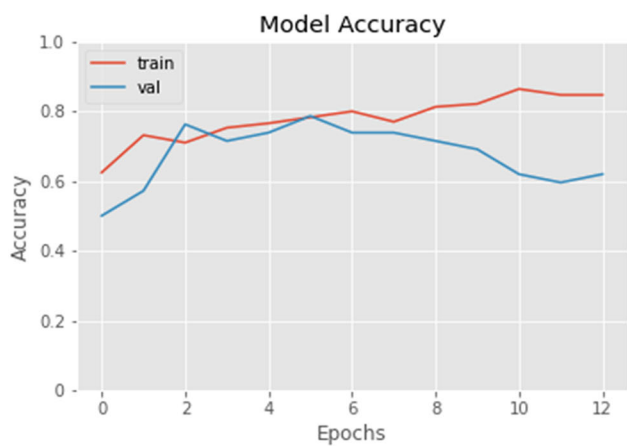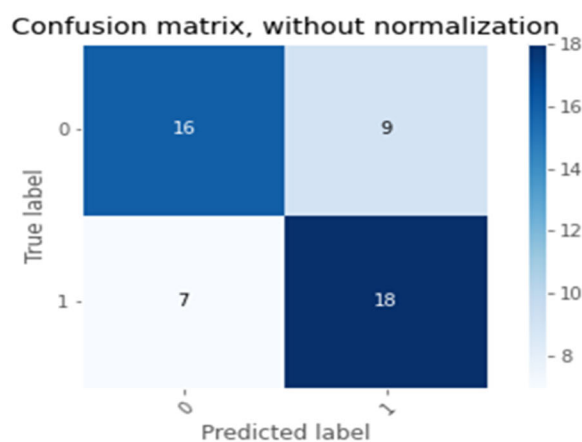
(C) D-model

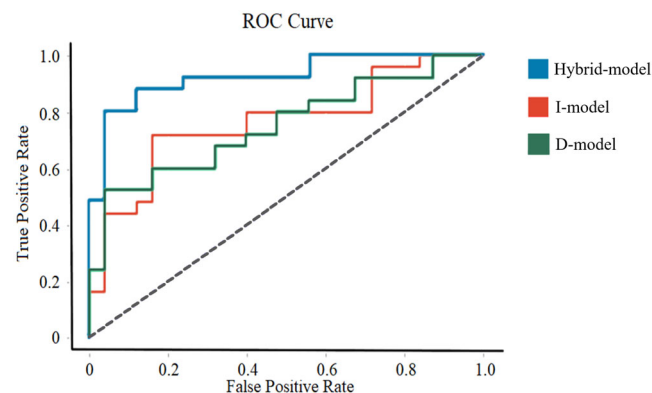**FIGURE 6** Learning curves and confusion matrix of the three models

FIGURE 7 A combination of the areas under the curve of the three models trained. ROC, receiver operating characteristic

Concerning the hybrid input approach, we found some other similar studies analyzing other diseases, and our conclusions are consistent with those previous studies, which also found similar or better performance when including tabular data. For instance, Heo et al.[24] found better results in identifying tuberculosis signs on X-rays mixing images with demographic data compared to using images alone. Accordingly, we believe that hybrid inputs could be a promising technique in medical challenges since most diseases or outcomes have a multifactorial cause.

Deep neural networks, which are data-driven systems, allow us to include large amounts of data as input to the network, without knowing in advance if they are important regarding the problem to be solved. The network itself will decide this during the training (feature selection), and the redundant or spurious data will not hinder the network. This is one of the main advantages they provide since they save us from having to manually process the data in advance (end-to-end approach), allowing us to feed models with large amounts of data directly from the patient's EHR.[22] Nonetheless, the development of these models is still limited by some obstacles, notably, the shortage of annotated medical images, the lack of data exploitation tools in hospitals, and the complicated ethical and legal framework regarding medical data transfer, three obstacles that are interconnected.

One of the reasons for the scarcity of large, annotated image datasets is that manual annotation is a very time-consuming task for radiologists. In recent years, some techniques have been developed to automate image annotation such as natural language processing techniques and structured radiological reports, but they are not yet widely available in hospitals.[25] And the same is true of data exploitation tools, such as the one published in 2021 by Yi et al.[26] We believe that with these tools it will soon be possible to feed models with large annotated image datasets and with large amounts of tabular data directly from the patient's EHR. To simulate this, in this study we introduced all the patients' blood test parameters, without manually choosing those important parameters beforehand, to avoid a manual feature selection. However, in preparing this work, we did not have any data exploitation tools in our hospital, so we were obliged to gather all the clinical data manually from the EHR, which was very time consuming. Conse-

quently, the clinical data we could get from each patient were limited. Hence, this study also demonstrates the importance of bringing these tools to hospitals, since the lack of them is an important barrier to the development of AI systems in clinical settings.

Regarding the last limitation, the complicated ethical and legal framework of medical data transfer is partially due to the requirement for meticulous data de-identification, and patient consent at least in prospective studies.[19,27] Despite this handicap, there are some publicly available databases that can be used for research purposes,[28] to which we are adding the dataset used in this work.[17] However, using a database that does not adequately represent the real population can lead to a drop in model performance when tested in a real setting.[29,12] While we acknowledge this study included a limited number of patients, it used a real-world dataset from our institution.

In summary, in this work we presented a novel deep neural network trained using hybrid inputs that was able to accurately classify patients with ICH into good and poor prognosis, based on hospital survival. Thus, we conclude that deep learning models have potential to be used for prognosis prediction of several illnesses including ICH, which could have a great impact on clinical decision-making.

Further, this research also suggests that the use of hybrid inputs, concatenating tabular data to the output of the convolutional part, is a promising approach for deep learning models.

ORCID

David Rodríguez González https://orcid.org/0000-0002-9160-5106

REFERENCES

1. Steiner T, Petersson J, Al-Shahi Salman R, et al. European research priorities for intracerebral haemorrhage. Cerebrovasc Dis. 2011;32:409-19.
2. van Asch CJ, Luitse MJ, Rinkel GJ, et al. Incidence, case fatality, and functional outcome of intracerebral haemorrhage over time, according to age, sex, and ethnic origin: a systematic review and meta-analysis. Lancet Neurol. 2010;9:167-76.
3. Steiner T, Al-Shahi Salman R, Beer R, et al. European Stroke Organisation (ESO) guidelines for the management of spontaneous intracerebral hemorrhage. Int J Stroke. 2014;9:840-55.
4. Patel SK, Saleh MS, Body A, et al. Surgical interventions for supratentorial intracranial hemorrhage: the past, present, and future. Semin Neurol. 2021;41:54-66.
5. Mendelow AD, Gregson BA, Rowan EN, et al. Early surgery versus initial conservative treatment in patients with spontaneous supratentorial lobar intracerebral haematomas (STICH II): a randomised trial. Lancet. 2013;382:397-408.
6. Caceres JA, Goldstein JN. Intracranial hemorrhage. Emerg Med Clin North Am. 2012;30:771-94.
7. Nakano T, Ohkuma H. Surgery versus conservative treatment for intracerebral haemorrhage—is there an end to the long controversy? Lancet. 2005;365:361-2.

8. Poirion OB, Jing Z, Chaudhary K, et al. DeepProg: an ensemble of deep-learning and machine-learning models for prognosis prediction using multi-omics data. Genome Med. 2021;13:112.

9. Schulz S, Woerl AC, Jungmann F, et al. Multimodal deep learning for prognosis prediction in renal cancer. Front Oncol. 2021;11:788740.

10. Chieregato M, Frangiamore F, Morassi M, et al. A hybrid machine learning/deep learning COVID-19 severity predictive model from CT images and clinical data. Sci Rep. 2022;12:4329.

11. Arbabshirani MR, Fornwalt BK, Mongelluzzo GJ, et al. Advanced machine learning in action: identification of intracranial hemorrhage on computed tomography scans of the head with clinical workflow integration. NPJ Digit Med. 2018;1:9.

12. Ginat DT. Analysis of head CT scans flagged by deep learning software for acute intracranial hemorrhage. Neuroradiology. 2020;62:335-40.

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. Ye H, Gao F, Yin Y, et al. Precise diagnosis of intracranial hemorrhage and subtypes using a three-dimensional joint convolutional and recurrent neural network. Eur Radiol. 2019;29:6191-201.

15. Heit JJ, Coelho H, Lima FO, et al. Automated cerebral hemorrhage detection using RAPID. AJNR Am J Neuroradiol. 2021;42:273-8.

16. Mongan J, Moy L, Kahn CE Jr. Checklist for artificial intelligence in medical imaging (CLAIM): a guide for authors and reviewers. Radiol Artif Intell. 2020;2:e200029.

17. Pérez A, Menéndez P, Sanz P, et al. Head-CT 2D/3D images with and without ICH prepared for Deep Learning. In: DIGITAL.CSIC. Available from: https://doi.org/10.20350/digitalCSIC/14706. Accessed 12 Aug 2022.

18. González DR, Carpenter T, van Hemert JI, et al. An open source toolkit for medical imaging de-identification. Eur Radiol. 2010;20:1896-904.

19. Lee JY, Kim JS, Kim TY, et al. Detection and classification of intracranial haemorrhage on CT images using a novel deep-learning algorithm. Sci Rep. 2020;10:20546.

20. Sun W, Cai Z, Li Y, et al. Data processing and text mining technologies on electronic medical records: a review. J Healthc Eng. 2018;2018:4302425.

21. Iglesias LL, Bellón PS, Del Barrio AP, et al. A primer on deep learning and convolutional neural networks for clinicians. Insights Imaging. 2021;12:117.

22. Miotto R, Wang F, Wang S, et al. Deep learning for healthcare: review, opportunities and challenges. Brief Bioinform. 2018;19:1236-46.

23. Somani S, Russak AJ, Richter F, et al. Deep learning and the electrocardiogram: review of the current state-of-the-art. Europace. 2021;23:1179-91.

24. Heo SJ, Kim Y, Yun S, et al. Deep learning algorithms with demographic information help to detect tuberculosis in chest radiographs in annual workers' health examination data. Int J Environ Res Public Health. 2019;16:250.

25. Willemink MJ, Koszek WA, Hardell C, et al. Preparing medical imaging data for machine learning. Radiology. 2020;295:4-15.

26. Yi T, Pan I, Collins S, et al. DICOM Image ANalysis and Archive (DIANA): an open-source system for clinical AI applications. J Digit Imaging. 2021;34:1405-13.

27. Schönberger D. Artificial intelligence in healthcare: a critical analysis of the legal and ethical implications. Int J Law Inf Technol. 2019;27:171-203.

28. Flanders AE, Prevedello LM, Shih G, et al. Construction of a machine learning dataset through collaboration: the RSNA 2019 brain CT hemorrhage challenge. Radiol Artif Intell. 2020;2:e190211.

29. Voter AF, Meram E, Garrett JW, et al. Diagnostic accuracy and failure mode analysis of a deep learning algorithm for the detection of intracranial hemorrhage. J Am Coll Radiol. 2021;18:1143-52.