Comprehensive Hyperparameter Tuning to Enhance Deep Learning Performance for Intracranial Hemorrhage Classification in Head CT Scans

Mohammad Hoseyni
Department of Electrical Engineering
K. N. Toosi University of Technology
Tehran, Iran
mohammadhosini@email.kntu.ac.ir

Kasra Davoodi
Department of Electrical Engineering
K. N. Toosi University of Technology
Tehran, Iran
seyedkasra.davoodi@email.kntu.ac.ir

Fatemeh Pakdaman
Department of Electrical Engineering
K. N. Toosi University of Technology
Tehran, Iran
fatemeh.pakdaman@email.kntu.ac.ir

Amirhossein Nikoofard*
Department of Electrical Engineering
K. N. Toosi University of Technology
Tehran, Iran
a.nikoofard@kntu.ac.ir

Mahdi Aliyari-Shoorehdeli
Department of Electrical Engineering
K. N. Toosi University of Technology
Tehran, Iran
aliyari@eetd.kntu.ac.ir

***Abstract*—Intracranial hemorrhages (ICHs) pose a critical medical challenge with a high mortality rate, necessitating timely and accurate diagnosis. This study focuses on enhancing the performance of deep learning models for classifying brain hemorrhages in CT scan images using the PhysioNet dataset. A grid search methodology was employed to tune hyperparameters, particularly addressing the imbalance between ICH and healthy slices. The study utilized ResNet50 for hyperparameter tuning, achieving significant improvements in sensitivity and overall performance through undersampling, bootstrapping, augmentation, and weighted loss techniques. The ResNet50 model demonstrated remarkable performance at the slice-level scope, achieving a sensitivity of 0.94, a specificity of 0.91, an F1 score of 0.64, and an accuracy of 0.91. It outperformed the literature in both sensitivity and accuracy, despite utilizing a smaller and more constrained dataset. At the patient-level scope, our approach achieved even higher metrics with a sensitivity of 1.00, a specificity of 0.80, an F1 score of 0.86, and an accuracy of 0.88, surpassing the literature across all reported metrics.**

***Index Terms*—Intracranial Hemorrhage, Deep Learning, CT Scan, Classification, Patient-Level, Slice-Level, Grad-CAM, t-SNE**

I. Introduction

ICH accounts for a critical medical situation, with a 40% patient mortality rate despite offering medical treatments and services to patients in a timely manner. [1] It is important to note that increasing life expectancy and inadequate blood pressure control have likely influenced the noticeable increase in hospital admissions over the past decade [2]. An AI assistant tool holds substantial potential to save numerous lives through faster and more accurate diagnoses by aiding physicians and radiologists. [3]. Hemorrhages may occur either within the brain tissue (intra-axial) or within the skull but outside the brain tissue (extra-axial). While both intra-axial and extra-axial hemorrhages carry significant clinical implications [2]. [3]. Despite the fact that these disorders are to be first recognized by symptoms they show, accurate diagnosis is mostly made depending on the imaging techniques [4]. In cases of traumatic brain injury (TBI) the most common imaging method of assessing the severity of brain hemorrhages is Computed Tomography (CT) [5].

CT scans exclusive properties as availability and the short time that it takes to acquire a data record, make them prioritized and preferred as the first-line diagnostic tool [6]. The utilization of CT imaging for the detection of ICH is recognized as a non-invasive and effective technique. Hemorrhages are distinguishable on non-contrast CT scans due to the slightly higher density of blood relative to other brain tissues and bones [3]. The traditional diagnosis technique of ICH is made by radiologists' visual inspection of CT scans and manually estimating the size of the hematoma and any midline shift while the critical part of the clinical treatment is the assessment of acute ICH cases [5], [7]. Initial CT scan interpretations are performed by junior radiologists, radiology trainees, or emergency physicians to provide immediate care to patients with acute conditions. Senior or more experienced radiologists subsequently review these initial interpretations to ensure accuracy and thoroughness [3]. However, the diagnostic process requires a significant amount of time, making the constant availability of a trained radiologist crucial. Furthermore, CT scans must be frequently evaluated in an emergency department, even outside regular working hours [3], [5]. The importance of early diagnosis in ICH cannot be overstated, as nearly half of ICH-related mortalities occur within the first 24 hours [2]. Delays in diagnosis can arise due to increased imaging utilization and distractions caused by non-interpretive tasks, resulting in turnaround times for non-contrast brain CT scans in emergency departments ranging from 1.5 to 4 hours. Such delays significantly impact patient care, as acute worsening from hemorrhage expansion often manifests

within the initial hours of symptom onset [1]. Computer-aided monitoring of acute neurological events in cranial imaging shows promise for improving radiology workflow, potentially reducing treatment times and workload while improving patient outcomes [4]. Therefore, an automated AI triage tool for head CT scans could prove highly advantageous in managing queues in busy trauma care settings or assisting decision-making in remote locations lacking immediate access to a radiologist [6].

Advances in computer vision techniques, such as deep learning, have demonstrated substantial potential for extracting clinically relevant information from medical images [7]. Recently, significant promise has been shown by artificial intelligence (AI) in the domain of medical imaging. Some studies have attempted to detect abnormalities in head CT scans, including ICH, using deep learning and machine learning methods [3]. Deep learning is distinguished by its ability to identify meaningful patterns in raw data without explicit instructions, provided it has sufficient examples. However, challenges such as the need for large datasets present substantial obstacles to the development and clinical implementation of medical deep learning systems [8]. Numerous deep learning methods have been proposed in recent years for classifying critical imaging findings which have demonstrated remarkable generalization capabilities.

To mention other works that have attempted to classify medical images on ICH, a 2D deep learning approach named RADnet for automated brain hemorrhage detection on CT scans was described by Grewa et al. [5], achieving a recall of 0.89 and an F1 score of 0.85 in a patient-wise manner. Rajagopal et al. [9] reported a precision of 0.9521 and an F1 score of 0.9463 using Hybrid Deep Neural Networks on the RNSA open dataset [10] in 2023 in which the decision it made on each slice. A review of various deep learning methodologies was conducted by Neethi et al. [11], utilizing various datasets, including the PhysioNet [12] and RSNA ICH datasets. They used the RSNA ICH dataset which contains 847000 slices for training deep learning models and assessed its performance on the PhysioNet ICH dataset. They achieved a recall of 0.76 and an F1 score of 0.67 with a ResNet50-V2 on the PhysioNet dataset. Among works on the PhysioNet ICH dataset, Ganeshkumar et al. [13] reported a recall of 0.89 and an F1 score of 0.91 using CycleGAN for data augmentation where slice-wise training and testing method was employed, and Kyung et al. [14] proposed a patient-level supervised multi-task aiding representation transfer learning network (SMART-Net) to achieve an F1 score of 0.73. wrong

Although some papers report high metrics in the literature review, there are some important problems in their training and assessment process. Since a CT scan is a 3D imaging modality, each CT image contains many slices thus splitting data into train set and test set should be done in a patient-level manner. Splitting data in a slice-level manner leads to an increase in the correlation of train and test sets, which causes inaccurate results on the test set. Moreover, due to the small size of the PhysioNet ICH dataset, using k-fold cross-validation is vital to evaluate the model performance in different situations and obtain more stable results. Additionally, due to the imbalance problem, using the accuracy metric to calculate the performance of the model is not preferable and other metrics like sensitivity, specificity, precision, and F1 score are important, but some papers don't report some of them.

In this research, the hyperparameters for training a classification model on PhysioNet have been tuned by doing a grid search among different ratios of augmentation, undersampling, and weighted loss which led to obtaining similar results with those research that used many more data for training the same model. Moreover by doing a 5-fold cross-validation technique significantly higher sensitivity has been achieved in both patient-level and slice-level scope.

This paper is organized into five sections. Section II details the dataset and the methodology used to train the deep learning model. Section III evaluates the model's performance. Section IV provides a discussion, and Section V concludes the paper.

II. Materials and Method

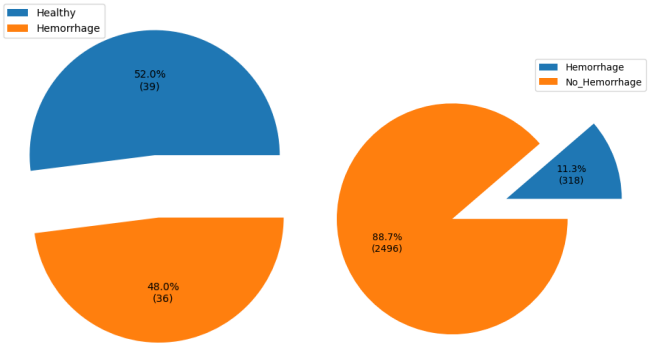
In this section, a comprehensive overview of the dataset employed in the study is provided. The specific characteristics and inherent challenges associated with this dataset are detailed. Furthermore, the methodologies and strategies implemented to effectively address these challenges are outlined, ensuring the robustness and reliability of the experimental results.

A. Dataset

The dataset titled "Computed Tomography Images for ICH Detection and Segmentation" [12], which is publicly available on PhysioNet [15], was utilized in this study. The CT scans were collected between February and August 2018 at Al Hilla Teaching Hospital, Iraq. This dataset comprises 82 non-contrast CT scans, each containing 34 slices with a 5 mm slice thickness. Of the subjects, 56% (46 patients) were male, and 44% (36 patients) were female. The dataset includes individuals ranging in age from 1 day to 72 years, with an average age of 27.8 ± 19.5 years. The broad age range impacts the skull's scale and shape observed in the CT scans, and this variation could affect the performance of the deep learning model. Out of the 82 patients, 36 were diagnosed with ICH. For data labeling, each slice was reviewed by two radiologists to determine the presence of hemorrhage or fractures [16].

As illustrated in Figure 1, the dataset consists of 2,814 slices, including 2,496 healthy slices and 318 slices with hemorrhage. This distribution exhibits a significant imbalance problem at the slice-level. However, at the patient-level, this imbalance problem is not observed. The imbalance at the slice-level is attributable to the fact that in patients with hemorrhage, many slices are healthy; in other words, hemorrhage occurs in only a few slices of a patient with ICH.

During data collection, 7 CT scans from patients aged 59 to 65 were missed. Kyung et al. suggest that the CT scans of patients 58 and 79 exhibit poor image quality according to radiologist opinions and should be eliminated



(a) (b)

Fig. 1: Distribution of slices (a) and patients (b) in the dataset.

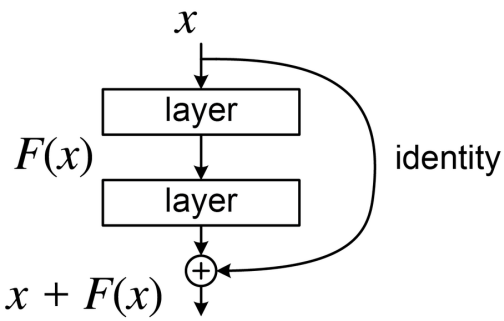


Fig. 2: Building block of residual learning

[14]. Consequently, 73 CT scans in NIFTI format are currently available. The CT scan images conform to the Hounsfield Unit scale, thus a windowing operation should be applied to the slices. The brain window is the target window in this research, with the window level set to 40 and the window width set to 120 [12].

B. Methodology

In this section, the detailed procedures for methodology are presented. First, the chosen deep learning model and the reasons for its selection are discussed. Next, methods for addressing unbalanced data are elaborated upon. Thirdly, the training configurations are described and Lastly, the decision policy is explained.

1) Model Selection: As noted in the preceding section, the dataset exhibits considerable imbalance at the slice level. This observation informed our decision to opt for a model with a higher chance of generalizability rather than a complex one with an extensive number of parameters. Our objective was to identify effective hyperparameters independent of the deep learning model while mitigating the risk of overfitting. Hence, ResNet50, which is a benchmark and common model, was selected as the main model for hyperparameter tuning.
ResNet50, introduced by He et al. [17], features a key innovation in its use of residual blocks, which address the problem of vanishing gradients in very deep neural networks, as illustrated in Figure 2. These residual blocks introduce skip connections that bypass one or more layers, allowing the network to learn residual mappings. This structure helps to maintain the flow of gradients through the network, making it easier to train deeper models by reducing the degradation problem. The skip connections enable the model to retain learned information from earlier layers, effectively enhancing its performance and stability in training deeper architectures. This design not only improves the generalizability of the model but also helps in achieving better accuracy with fewer training epochs.

2) Unbalance Reduction: Improving the model's performance significantly depends on addressing the imbalanced ratio between healthy and hemorrhage slices. Four main approaches were considered: undersampling, augmentation, bootstrap, and weighted loss. Undersampling involves reducing the number of healthy slices to increase the hemorrhage-to-healthy ratio. Conversely, augmentation and bootstrap focus on increasing the number of hemorrhage slices, thereby elevating the hemorrhage-to-healthy slice ratio. A crucial consideration before trying each of the explained methods is the distribution of hemorrhage slices within a patient. All slices of each patient were examined to identify the most frequently occurring bleeding slices. This process was repeated for all patients and led to finding the distribution of hemorrhage over slice number. undersampling, augmentation, and bootstrap were applied on slices based on the mentioned distribution. The fourth method is the weighted loss in the training procedure. Due to the significant difference in the amount of healthy and ICH slices, each training slice was assigned a weight based on its class frequency. Slices from the hemorrhage class were given higher weights to increase their influence during model training, thereby preventing the model from being biased toward the majority class. weighted loss was applied in two ways. First, the loss for ICH slices was incrementally increased by factors of 2, 4, 6, 8, and 10. Conversely, the weight of healthy slices was decreased by factors of 0.75, 0.5, and 0.25. All the aforementioned methods were integrated into an extensive grid search. This computationally intensive approach aimed to find the set of hyperparameters in order to improve the model's performance.

3) Training Configurations: The data was partitioned with approximately 80% allocated to the training set, while the remaining 20% was reserved for the test set. Importantly, this partitioning was conducted at the patient level, ensuring that any slices from a single patient were exclusively assigned to either the training set or the test set. This isolation approach prevented any overlap of patient data between the training and test sets. During the splitting procedure, both the training and test sets maintained the same ratio between healthy and ICH patients as observed in the original dataset. Moreover, 5-fold cross-validation was applied to the training set. Cross-entropy was considered as the loss function in the training procedure and Adam was selected as the optimizer of the training. Finally, the deep learning model is trained on slices of CT scan images. The training process was carried out using an NVIDIA T4 Tensor Core GPU and 16GB of RAM. It is

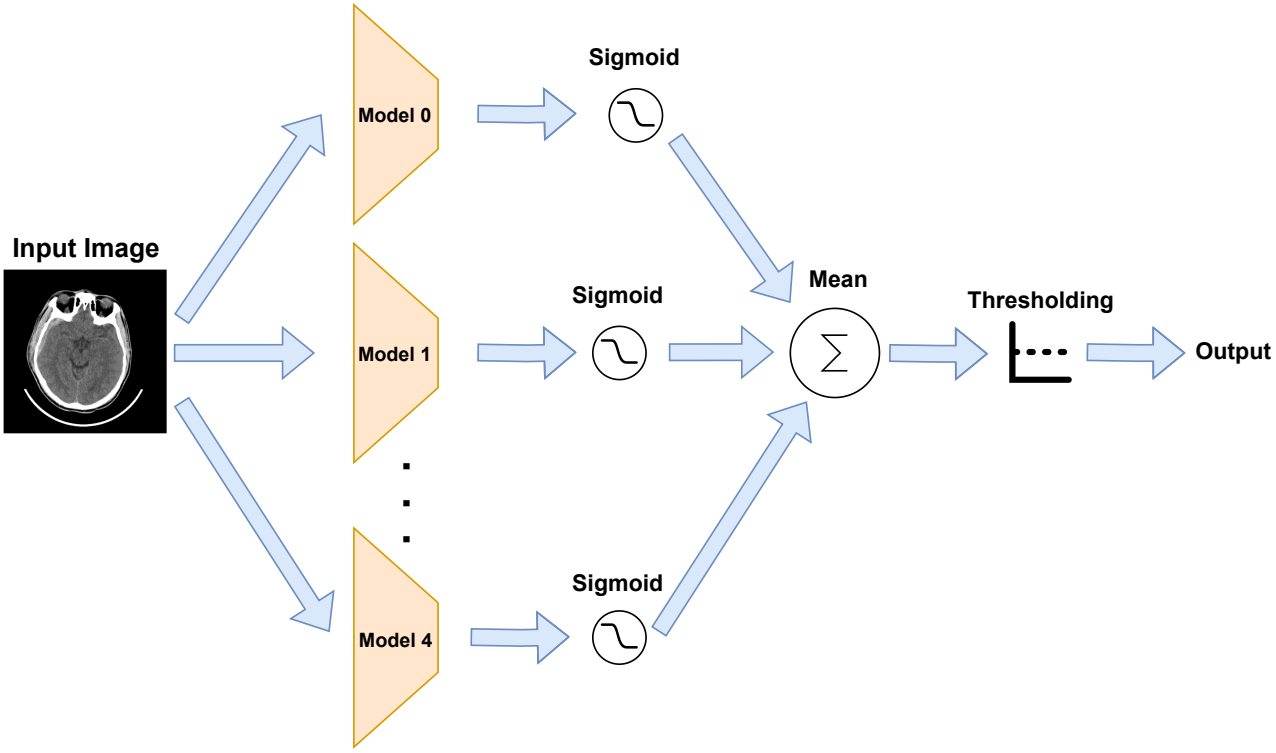


Fig. 3: voting structure for 5-fold cross-validation

also important to note that ResNet50, along with other models employed in this research, were all pre-trained on the widely recognized ImageNet [18] dataset.

4) Decision Policy: As noted above, 5-fold cross-validation was applied to each training iteration. For each fold within any given training, a validation fold was designated, and a graph was plotted for the F1 score over the classification threshold to identify the highest F1 score. This process was conducted for all 5 folds, resulting in 5 graphs. The average of these graphs was then computed, and the threshold corresponding to the highest F1 score on this mean graph was reported as the final threshold. Results from this average procedure will be referred to as the "voting model" throughout the remainder of the paper.

III. Results

In summary, a comprehensive analysis of the results encompassing slice-level and patient-level classifications is presented in this paper. The policy for determining the final hyperparameters in grid search involved calculating the mean F1 score and identifying the configuration that achieved the highest F1 score. Our best results were obtained when the augmentation of slices with ICH generated 3.3 times the synthetic data, and the weighted loss for ICH slices was set to be four times greater than that for normal slices.

Table I presents the results of training the ResNet50 model and details of each fold in the training process for the slice-level scope. As exhibited in the table, fold 0 achieves the best metrics among the other folds, reaching an F1 score of 0.62. This indicates that the distribution of data in the training and validation sets of fold 0 is most similar to that of the test set. Conversely, fold 4 attains the lowest metrics on the test set, while the metrics of the other folds are more consistent with each other.

Figure 4 indicates the F1 score over the threshold graph for

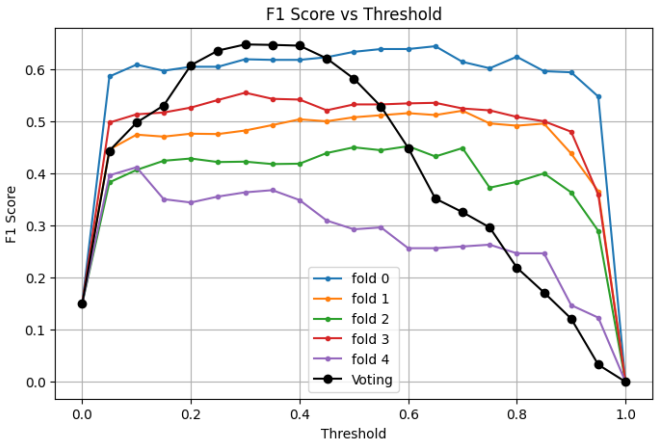


Fig. 4: F1 score over threshold on test set

different folds and the voting technique. Using all 5 folds together in a voting policy leads to a significant improvement in the F1 score and other metrics, particularly sensitivity, which is crucial for deep learning models in medical applications. The sensitivity metric measures the proportion of ICH slices correctly identified; thus, our model achieves a sensitivity of 0.94, a remarkable achievement in comparison with the

literature. For example, Neethi et al. [19] trained a ResNet-50-V2 on the RSNA [10] dataset, which has 847,000 CT scan slices, much larger than the PhysioNet dataset. Their trained model was then evaluated on the PhysioNet dataset, and they obtained metrics comparable to ours. This demonstrates that with well-tuned hyperparameters, the effect of a large amount of data can be reduced, resulting in significantly higher sensitivity and accuracy.

Higher sensitivity means that our model can diagnose ICH slices better, and higher accuracy, along with other metrics reported by Neethi et al., indicates that our model can predict more true negatives than theirs. Figure 5 shows the confusion matrix results, indicating the absolute number of slices diagnosed by the model.

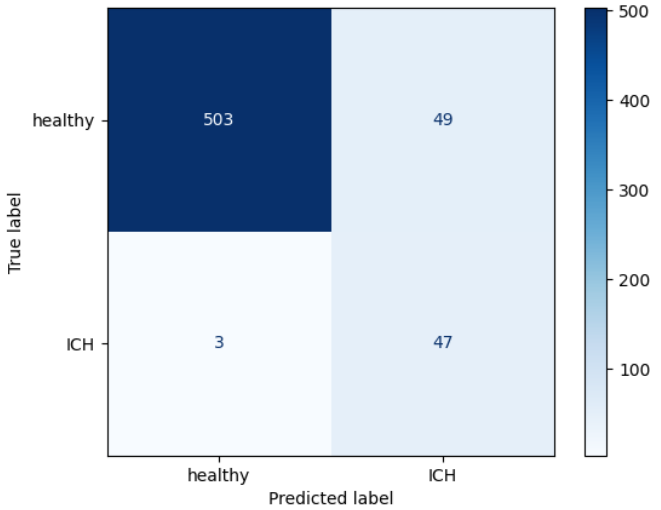


Fig. 5: Slice-level confusion matrix

TABLE I: Slice-level metrics of models

Model	Sensitivity	Specificity	Precision	F1	Accuracy
Resnet50 Fold 0	0.78	0.93	0.51	0.62	0.91
Resnet50 Fold 1	0.68	0.90	0.38	0.49	0.88
Resnet50 Fold 2	0.60	0.89	0.33	0.42	0.86
Resnet50 Fold 3	0.86	0.89	0.41	0.55	0.89
Resnet50 Fold 4	0.32	**0.96**	0.42	0.36	0.90
Neethi et al. [19]	0.76	-	**0.69**	**0.67**	0.54
Resnet50 Voting	**0.94**	0.91	0.49	0.64	**0.91**

Figure 6b demonstrates the t-SNE technique, providing a 2D representation of the features extracted from a ResNet-50 model trained on the ImageNet dataset, concluding that the extracted features are highly non-linear. In contrast, Figure 6a shows the 2D representation of features extracted from a ResNet-50 model trained on the PhysioNet dataset, with the result exhibiting an acceptable level of linear separability. Ganeshkumar et al. [13] reported a highly non-linear t-SNE graph from features extracted from their model, whereas the 2D representation obtained in our study shows an acceptable level of linear separability. This indicates that the features extracted from the ResNet-50 model possess high semantic value.

Additionally, to empirically validate the configuration obtained in this research, three other models—VGG16, MobileNet, and Inception—have been trained with the aforementioned hyperparameters and the 5-fold cross-validation technique. The results of these models, as exhibited in Table II, and the comparison of metrics for models trained on the original dataset without augmentation and weighted loss, with the results of models trained with the aforementioned configuration, indicate that the findings in this research could improve the performance of deep learning models.

TABLE II: Empirical assessment of proposed configuration for slice-level on other benchmark models

Model	Config	Sensitivity	Specificity	Precision	F1	Accuracy
ResNet 50	✓	0.94	**0.91**	**0.49**	**0.64**	**0.91**
ResNet 50	×	0.76	0.78	0.23	0.36	0.77
VGG-16	✓	0.92	0.88	0.41	0.56	0.88
VGG-16	×	0.88	0.76	0.25	0.39	0.77
MobileNet-V2	✓	0.98	0.82	0.33	0.50	0.84
MobileNet-V2	×	**1**	0.68	0.22	0.36	0.71
Inception-V3	✓	0.64	0.90	0.36	0.46	0.88
Inception-V3	×	0.88	0.78	0.27	0.41	0.79

Figure 7 illustrates the results of the Grad-CAM [20] technique, which is utilized to enhance the interpretability and explainability of the ResNet-50 model. The first row in this image displays the original radiography CT scan with a brain window. The second row highlights the regions activated by the classification model to identify hemorrhagic lesions. The last row presents the segmentation mask provided in the PhysioNet dataset. Based on the output of Grad-CAM, the functionality of ResNet-50 in the localization of hemorrhagic lesions is determined.

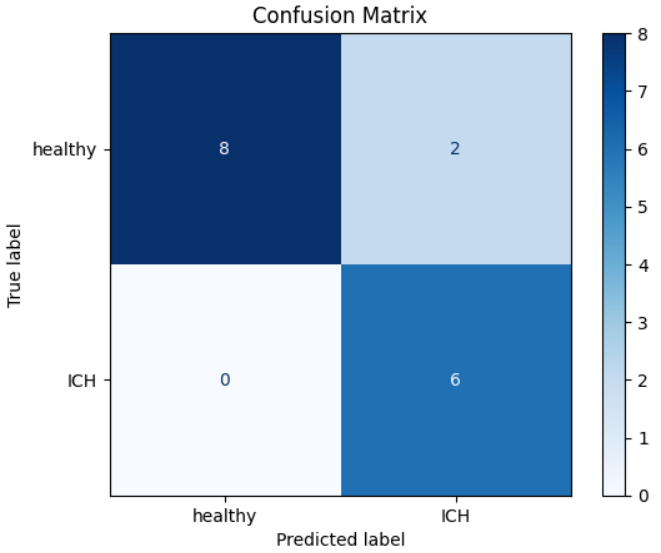
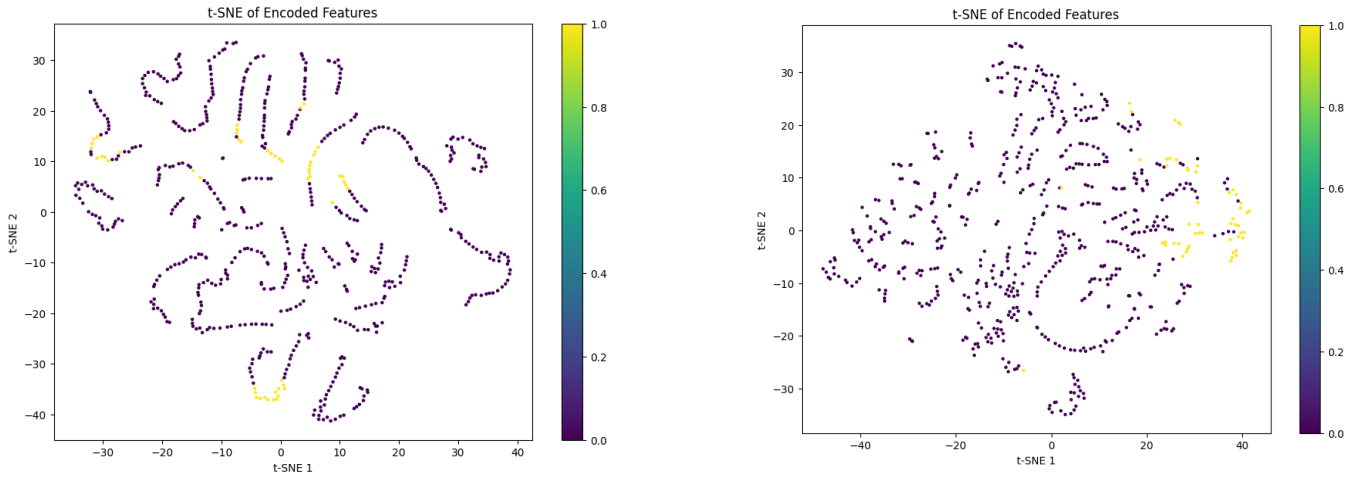


Fig. 8: Patient-Level confusion matrix

Table III presents the results of training the ResNet-50 model, detailing the performance of each fold in the training process for the patient-level classification. The prediction of ICH at the patient level is based on slice-level predictions, where a patient is considered to have ICH if at least one slice of their CT scan is identified as ICH. Consequently, the results from fold 0 to fold 3 are identical, reflecting their consistent performance in the slice-level classification. Additionally,



(a) (b)

Fig. 6: 2D representation of test set with t-SNE algorithm. a) t-SNE representation after training the model. b) t-SNE representation before training the model

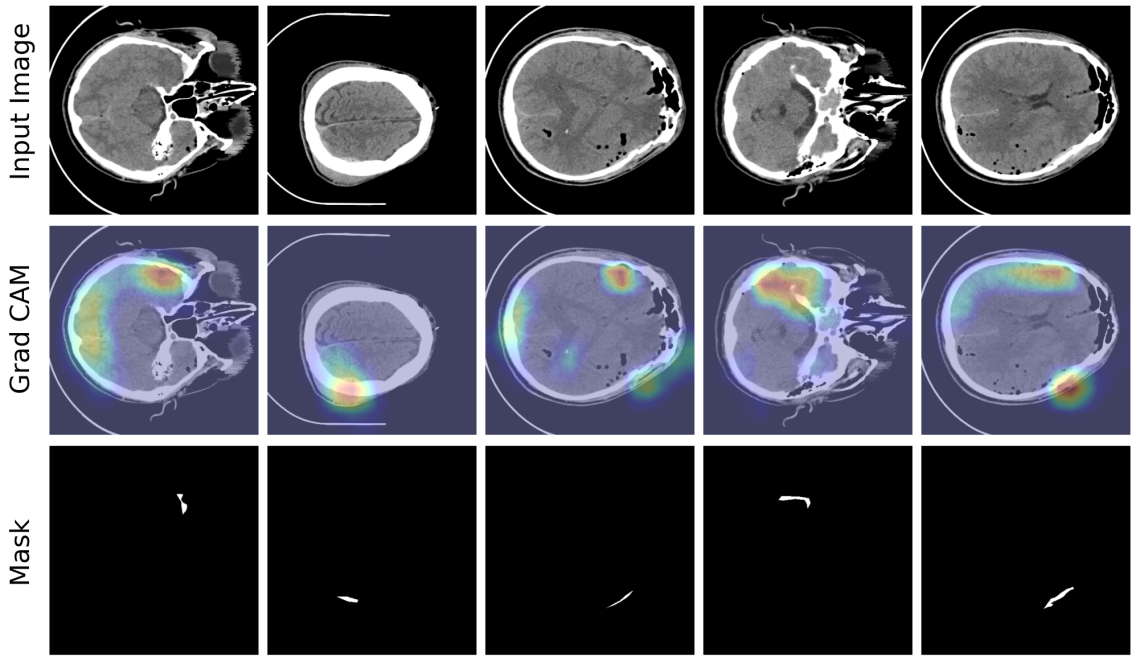


Fig. 7: Grad-CAM result to generate more interpretability and explainability for classification model

Kyung et al. [14] conducted patient-level classification on the PhysioNet dataset, achieving a sensitivity of 0.97, specificity of 0.74, and an F1 score of 0.84. Our results surpass theirs in all metrics, demonstrating the effectiveness of our approach. Specifically, our model achieves superior sensitivity, specificity, and F1 score, indicating a more accurate and reliable performance in detecting ICH at the patient level.Figure 5 indicates the performance of the model on 16 radiography CT scan images of patients included in the test set. This figure provides a detailed visual representation of how the model classifies each patient, demonstrating its ability to accurately identify cases of ICH.

TABLE III: Patient-level metrics of models

Model	**Sensitivity**	**Specificity**	**Precision**	**F1**	**Accuracy**
ResNet 50 Fold 1	1.00	0.60	0.60	0.75	0.75
ResNet 50 Fold 1	1.00	0.60	0.60	0.75	0.75
ResNet 50 Fold 2	1.00	0.60	0.60	0.75	0.75
ResNet 50 Fold 3	1.00	0.60	0.60	0.75	0.75
ResNet 50 Fold 4	0.83	**0.80**	0.71	0.77	0.81
Kyung et al. [14]	0.97	0.74	-	0.84	-
ResNet 50 Voting	**1.00**	**0.80**	**0.75**	**0.86**	**0.88**

IV. Discussion

This study aimed to enhance the performance of ICH binary classification on a small, unbalanced dataset by conducting an extensive grid search for hyperparameter tuning. As previously mentioned, ResNet50 was chosen as the core model for this study due to its generalizability and moderate complexity. Given the limitations of this dataset, our primary challenge was to overcome the imbalance ratio between ICH and healthy slices as much as possible.

To address the imbalance problem, a grid search was applied on distribution-based augmentation, distribution-based undersampling, and bootstrapping along with the weighted loss function. It's important to note that although undersampling improves the ratio between ICH and healthy slices, it can also leads to additional limitations in slice variations by removing various slices in an already small dataset. Bootstrap sampling is also susceptible to overfitting due to its inability to create sufficient variety in the slices. As expected, the best results in the grid search were obtained by applying augmentations alongside weighted loss. The results for weighted loss hyper-parameters indicated that increasing the weights of ICH slices yielded better outcomes than decreasing the weights of healthy slices. All of the results shown in Table II and Table III incorporate a 3.3x augmentation along with a four-time increase in the weights of ICH slices loss. Using these hyperparameters, the most significant improvement was observed in ResNet50, the primary model used in this study. However, due to its generalizability, this configuration also showed improvements across other models, as indicated in Table II.

It is important to emphasize that an ICH classifier is an assistant tool to provide second opinions for doctors and physicians, thus higher sensitivity over precision is preferred due to its lower risk to patients' health. It can be observed from Table I that our proposed voting model achieved significantly higher slice-level sensitivity and accuracy, with a slightly lower F1 score compared to Neethi et al in [19] for slice-level scope. In patient-level classification, the voting model achieved a sensitivity of 1.00, specificity of 0.80, precision of 0.75, F1 score of 0.86, and accuracy of 0.88, surpassing Kyung et al. [14] in every reported metric.

V. Conclusion

In conclusion, it can be asserted that the proposed hyperparameters enhance the performance of deep learning models on the PhysioNet dataset. The main achievement of this research is the ability to extract richer features from the dataset. Future work is proposed to assess the proposed ratio between ICH slices and healthy slices on other datasets, to experimentally demonstrate that these ratios could increase the performance of deep learning models. Additionally, the use of more complex models on this dataset will be explored to determine if the results can be extended to more sophisticated models.

References

[1] Peter D Chang, Edward Kuoy, Jack Grinband, Brent D Weinberg, Matthew Thompson, Richelle Homo, Jefferson Chen, Hermelinda Abcede, Mohammad Shafie, Leo Sugrue, et al. Hybrid 3d/2d convolutional neural network for hemorrhage evaluation on head ct. *American Journal of Neuroradiology*, 39(9):1609–1616, 2018.

[2] Mohammad R Arbabshirani, Brandon K Fornwalt, Gino J Mongelluzzo, Jonathan D Suever, Brandon D Geise, Aalpen A Patel, and Gregory J Moore. Advanced machine learning in action: identification of intracranial hemorrhage on computed tomography scans of the head with clinical workflow integration. *NPJ digital medicine*, 1(1):9, 2018.

[3] Hai Ye, Feng Gao, Youbing Yin, Danfeng Guo, Pengfei Zhao, Yi Lu, Xin Wang, Junjie Bai, Kunlin Cao, Qi Song, et al. Precise diagnosis of intracranial hemorrhage and subtypes using a three-dimensional joint convolutional and recurrent neural network. *European radiology*, 29:6191–6201, 2019.

[4] Joseph J Titano, Marcus Badgeley, Javin Schefflein, Margaret Pain, Andres Su, Michael Cai, Nathaniel Swinburne, John Zech, Jun Kim, Joshua Bederson, et al. Automated deep-neural-network surveillance of cranial images for acute neurologic events. *Nature medicine*, 24(9):1337–1341, 2018.

[5] Monika Grewal, Muktabh Mayank Srivastava, Pulkit Kumar, and Srikrishna Varadarajan. Radnet: Radiologist level accuracy using deep learning for hemorrhage detection in ct scans. In *2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018)*, pages 281–284. IEEE, 2018.

[6] Sasank Chilamkurthy, Rohit Ghosh, Swetha Tanamala, Mustafa Biviji, Norbert G Campeau, Vasantha Kumar Venugopal, Vidur Mahajan, Pooja Rao, and Prashant Warier. Deep learning algorithms for detection of critical findings in head ct scans: a retrospective study. *The Lancet*, 392(10162):2388–2396, 2018.

[7] Weicheng Kuo, Christian Hane, Pratik Mukherjee, Jitendra Malik, and Esther L Yuh. Expert-level detection of acute intracranial hemorrhage on head computed tomography using deep learning. *Proceedings of the National Academy of Sciences*, 116(45):22737–22745, 2019.

[8] Hyunkwang Lee, Sehyo Yune, Mohammad Mansouri, Myeongchan Kim, Shahein H Tajmir, Claude E Guerrier, Sarah A Ebert, Stuart R Pomerantz, Javier M Romero, Shahmir Kamalian, et al. An explainable deep-learning algorithm for the detection of acute intracranial haemorrhage from small datasets. *Nature biomedical engineering*, 3(3):173–182, 2019.

[9] Manikandan Rajagopal, Suvarna Buradagunta, Meshari Almeshari, Yasser Alzamil, Rajakumar Ramalingam, and Vinayakumar Ravi. An efficient framework to detect intracranial hemorrhage using hybrid deep neural networks. *Brain Sciences*, 13(3):400, 2023.

[10] Adam E Flanders, Luciano M Prevedello, George Shih, Safwan S Halabi, Jayashree Kalpathy-Cramer, Robyn Ball, John T Mongan, Anouk Stein, Felipe C Kitamura, Matthew P Lungren, et al. Construction of a machine learning dataset through collaboration: the rsna 2019 brain ct hemorrhage challenge. *Radiology: Artificial Intelligence*, 2(3):e190211, 2020.

[11] AS Neethi, S Niyas, Santhosh Kumar Kannath, Jimson Mathew, Ajimi Mol Anzar, and Jeny Rajan. Stroke classification from computed tomography scans using 3d convolutional neural network. *Biomedical Signal Processing and Control*, 76:103720, 2022.

[12] Murtadha Hssayeni, M Croock, A Salman, H Al-khafaji, Z Yahya, and B Ghoraani. Computed tomography images for intracranial hemorrhage detection and segmentation. *Intracranial hemorrhage segmentation using a deep convolutional model. Data*, 5(1):14, 2020.

[13] M Ganeshkumar, Vinayakumar Ravi, V Sowmya, EA Gopalakrishnan, KP Soman, and Chinmay Chakraborty. Identification of intracranial haemorrhage (ich) using resnet with data augmentation using cyclegan and ich segmentation using segan. *Multimedia Tools and Applications*, 81(25):36257–36273, 2022.

[14] Sunggu Kyung, Keewon Shin, Hyunsu Jeong, Ki Duk Kim, Jooyoung Park, Kyungjin Cho, Jeong Hyun Lee, GilSun Hong, and Namkug Kim. Improved performance and robustness of multi-task representation learning with consistency loss between pretexts for intracranial hemorrhage identification in head ct. *Medical Image Analysis*, 81:102489, 2022.

[15] Ary L Goldberger, Luis AN Amaral, Leon Glass, Jeffrey M Hausdorff, Plamen Ch Ivanov, Roger G Mark, Joseph E Mietus, George B Moody, Chung-Kang Peng, and H Eugene Stanley. Physiobank, physiotoolkit, and physionet: components of a new research resource for complex physiologic signals. *circulation*, 101(23):e215–e220, 2000.

[16] Murtadha D Hssayeni, Muayad S Croock, Aymen D Salman, Hassan Falah Al-Khafaji, Zakaria A Yahya, and Behnaz Ghoraani. Intracranial hemorrhage segmentation using a deep convolutional model. *Data*, 5(1):14, 2020.

[17] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[18] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision (IJCV)*, 115(3):211–252, 2015.

[19] AS Neethi, Santhosh Kumar Kannath, Adarsh Anil Kumar, Jimson Mathew, and Jeny Rajan. A comprehensive review and experimental comparison of deep learning methods for automated hemorrhage detection. *Engineering Applications of Artificial Intelligence*, 133:108192, 2024.

[20] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, pages 618–626, 2017.