From Pixels to Diagnosis: Using Machine Learning to Classify Medical Image sequences

A. Badilla-Olivas, Enrique Vílchez-Lizano, Brandon Mora-Umaña, Kenneth Villalobos-Solís, Adrián Lara Petitdemange
Escuela de Ciencias de la Computación e Informática
Universidad de Costa Rica
San José, Costa Rica
{anthonny.badilla, enrique.vilchezlizano, brandon.moraumana, kenneth.villalobossolis, adrian.lara}@ucr.ac.cr

***Abstract*—This study compares video and image-based machine learning models for intracranial hemorrhage (ICH) classification from CT scan sequences, hypothesizing that incorporating temporal information improves diagnostic accuracy. A video transformer model (ViVit) was compared against a state-of-the-art image-based convolutional neural network (ConvNeXt). ViVit achieved a mean accuracy of 0.72 and recall of 0.62, outperforming ConvNeXt's 0.60 accuracy and 0.13 recall, despite a slightly lower precision (0.73 vs. 0.67). These results highlight the importance of temporal context for ICH classification and the potential of video models in medical imaging. Future work should explore alternative architectures and address data imbalance.**

***Index Terms*—Machine learning, computer vision, hemorrhage, classification, video.**

I. Introduction

Intracranial hemorrhage (ICH) is a disastrous disease. It refers to any bleeding within the intracranial vault [1], making early diagnosis critical for effective treatment management [2, 3]. To this end, several detection methods have been explored, with computed tomography (CT) and magnetic resonance imaging (MRI) being the most common ones [3].

Non contrast computed tomography (NCCT) is usually the preferred methodology for acute ICH diagnosis, since it is readily accessible and exhibits high sensitivity [4]. This method produces multiples images, called slices which display different regions of the brain. As a consequence, radiologists have to interpret multiple CT slices, to identify hematomas, perihematomal edemas or other related hemorrhage causes [3]. Hence, the whole process is time-consuming [5].

Machine learning models have emerged as an efficient alternative for reducing costs and time, enabling the automatic identification of brain lesions [6]. In the context of early detection of ICH, commonly used techniques include support vector machines (SVM) [3], random forests, k-nearest neighbors (KNN), multilayer perceptrons (MLP) [7], as well as convolutional neural networks (CNNs) in both 2D [8] and 3D forms [9]. While many deep learning approaches focus on segmentation tasks [2, 10], others are geared towards prediction or classification [11].

To the best of our knowledge, most studies use a slice per patient as input for the models. Grewal et al. [5] is the only study that makes predictions based on multiple images of the brain hemorrhage in this context, which uses a LSTM with a CNN. Since doctors have to analyze multiple images (slices) to diagnose the patient's condition [3], we were motivated by the models that consider several slices per patient to assess whether or not they produce more accurate results. This was further suggested by studies that have tried so in different medical areas like breast cancer [12, 13, 14] or cardiology [15].

In this article, we propose a novel comparison for classifying patients with intracranial hemorrhage using a video-based classification model and a state-of-the-art image model. We wanted to explore if incorporating a temporal dimension by using a video model would produce favorable findings. In order to maintain as much fairness as possible, both models will classify sequences of X-ray CT images captured from various sections of the brain. However, the image model is trained using all the images independently instead of sequences as input.

Our video model achieves a mean accuracy of 0.72 and a mean multiclass recall of 0.62, outperforming the state-of-the-art image model in the classification task. Although the video model has a lower mean precision, it demonstrates greater consistency in producing correct predictions compared to the image model. These results highlight the significance of the temporal dimension in image sequences and the potential of video models in medical applications.

The structure of this paper is as follows. We begin by introducing the relevant background, followed by a review of the related work in this field. Next, we present the methodology for comparing image models with video classification models, detailing the the experimental setup. Following that, we report and analyze the results, applying our chosen methodology to assess whether the data supports our initial hypothesis. Lastly, we contextualize our findings within the current state of the art, highlighting their relevance, and suggest directions for future research in this domain.

II. Background

In this section we define key background knowledge to understand the study.

Intracranial hemorrhage (ICH) is a type of hemorrhage that occurs in specific parts of the brain and is usually caused by accidents that involve brain lesions [16]. This condition can lead to increased intracranial pressure, which may cause severe neurological deficits, brain herniation, infarcts, rebleeds, vasospasms, seizures, and even death [17]. Rapid diagnosis and intervention are crucial for improving patient outcomes, as delayed treatment can result in irreversible brain damage or complications.

It is also important to clarify the terminology used in the literature regarding intracranial hemorrhage (ICH) and hemorrhage in general. While some studies specifically refer to ICH as synonymous with intracerebral hemorrhage [3, 18], others use it to denote the broader concept of intracranial hemorrhage as described by Seymour et al. [7].

Independently of the type of hemorrhage prognosis being studied, non-contrast computed tomography (CT) scans of the head are one of the most common imaging modalities to evaluate the patient's condition [17]. This is a cross-sectional imaging technique that offers extra diagnostic insights in cases where standard radiography is inadequate. Multiple images (called slices) are obtained from a single scan. They are created by an X-ray beam that rotates in a circular gantry around the object.

These images can form a 3D image, being the third dimension the space and time. The slices represent the movement of the gantry across the patient. As a consequence, some studies refer to CT Slices as 3D images.

III. Related work

In this section, we review and discuss the key research and prior work relevant to our study. To ensure clarity and structure, we have divided the review into three different parts: Intracranial hemorrhage image classification, state of the art image models and video models.

A. Intracranial hemorrhage

A significant number of studies have focused on the segmentation of hemorrhagic lesions. This serves as a foundational step toward volume estimation and outcome prediction. For instance, Yu et al. [10] demonstrates effective segmentation techniques using a UNet for identifying ICH in brain CT scans. Similarly, Yao et al. [19] present an automated approach that enhances the segmentation process, highlighting the importance of accurate delineation in improving clinical outcomes.

In terms of advanced imaging techniques leveraging multiple images (3D images or sequences), Lu et al. [20] provide a compelling argument for the application of data augmentation strategies by mirroring the original CT image, which can enhance model performance. Furthermore, models like the improved 3D U-Net and those utilizing squeeze-and-excitation blocks have demonstrated promising results in achieving precise segmentation and volume measurements [2].

Conversely, the classification of intracerebral and intracranial hemorrhages have been explored through various approaches. For example, Thabarsa et al. [3] illustrate how support vector machines (SVM) have been effectively utilized to classify intracerebral hemorrhages based on radiomic features. Additionally, Seymur et al. [7] showcase the application of multiple machine learning techniques, including random forests, K-nearest neighbors (KNN), support vector machines, multilayer preceptrons, naive bayes and logistic regression for predicting outcomes related to intracranial hematoma expansion.

Despite the progress made in classification using single images, there is a scarcity of studies focusing on the classification of ICH with multiple CT slices. The work done by Grewal et al. [5], highlights the potential of utilizing multiple images in achieving higher accuracy, primarily through recurrent neural networks such as Long Short-Term Memory (LSTM) models. In a similar vein, Zhou et al. [21], demonstrate the application of both 3D CNNs and Swin Transformers, as well as remarking the benefits of multi-view approaches in medical image analysis.

B. Image Models

In medical image classification deep learning has been used to try to speed up and automate the process of detecting a condition based on a patients medical images as wells as tasks like image segmentation [22]. Architectures like Convolutional Neural Networks and Transformers have played an important role in recent years as well as techniques such as transfer learning [22].

In this context, ConvNeXT has emerged as a state-of-the-art model for image classification [23]. Its applications in medical imaging are particularly noteworthy, especially within the context of brain imaging. For instance, Nizamli et al. [24] achieved up to 95.44% accuracy using ConvNeXT for brain tumor detection in CT and MRI scans. Similarly, Panthakkan et al. [25] used ConvNeXT to classify brain tumors with 99% accuracy. Furthermore, Sharma et al. [26] demonstrated its effectiveness in multiclass classification of Alzheimer's disease from brain MRI images, achieving an accuracy of 98.89%. Notably, all these studies utilized pretrained weights on ImageNet 1k for transfer learning.

Beyond brain imaging, ConvNeXT has shown promise in classifying images related to other conditions, such as diabetic retinopathy [27] and breast cancer [28] where it stands out the high effectiveness given that breast cancer screening produces a sequence of images and individual images were use to classify as positive or negative.

C. Video models

Most of the efforts related to video models seem to be focused on creating new architectures that resolve the current challenges or solving generic tasks like human action recognition.

Notable examples of these new architectures include ViT-based models like TimeSFormer, proposed by Bertasius et al. [29], and ViVit, proposed by Arnab et al. [30]. On the other hand, significant advancements have also emerged from autoencoder architectures that employ masking techniques,

such as VideoMAE, developed by Tong et al. [31]. More recently, Li et al. [32] created their videoMamba model by adapting linear-complexity operators, opening the field to more cost effective alternatives.

Based on our research, there have not been many applications of video models in medical context. The only example found was the study by Howard et al. [15] where they analyzed the potential of different CNN architectures to classify echocardiography ultrasound videos. In this study, the authors compared traditional image-based CNNs, image CNNs with a time-distributed layer, 3D CNNs, and 'two-stream' (spatial and temporal) CNNs. Their findings showed that a two-stream CNN achieved the lowest error rate, suggesting that incorporating the temporal dimension could enhance the accuracy of automatic classification [15].

IV. Methodology

In this section we discuss the followed methodology. It provides a detailed explanation of the approach taken to collect, analyze, and interpret data. This section is composed of the dataset description and preprocessing methods, models, metrics, experiments descriptions and the statistical analysis.

A. Dataset

The dataset [16] contained 5000 JPG-format head CT images from 82 patients, evenly split between 2500 brain window images and 2500 bone window images. Each patient had about 30 slices. Intracranial hemorrhage masks were available for 318 images, though not required for classification. Two CSV files provided hemorrhage diagnosis data and patient demographic/clinical data.

For this study, only binary classification was relevant: cases were classified as positive (any hemorrhage present) or negative (no hemorrhage). Exploratory analysis revealed 36 patients with hemorrhage and 46 without. Only brain window images were analyzed.

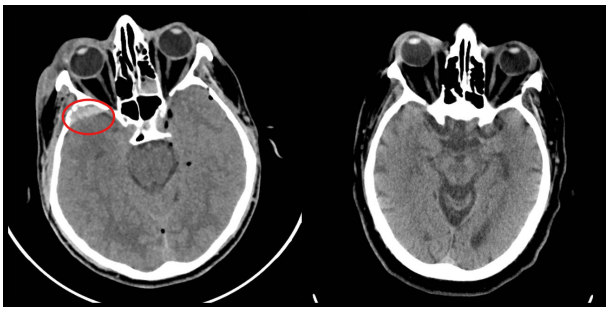

Fig. 1. Brain with hemorrhage vs brain without hemorrhage.

Since masks are not needed for classification, nor the bone images, we removed the mask images (which contained "HSE_Seg" in their name). Similarly, the whole bone folder was removed for each patient. Finally, we moved and separated patients into two folders, depending on whether they were positive or negative cases.

Fig. 2. Preprocessing flow.

B. Models

As previously mentioned, we used two models to compare their classification performance: a video model and an image model. The following subsections provide an overview of their architectures, model types, design, and other essential features and parameters.

For the video model, we selected a transformer-based architecture: ViVit [30]. This choice was motivated by the notable performance of transformers in single-image classification, the ability of these architectures to capture long-term dependencies through attention mechanisms, and the demonstrated effectiveness of ViVit as a modern, accessible model achieving promising results [30].

The specific ViVit model being used was Google's Vivit-b-16x2-kinetics400, which features a base backbone comprising 12 transformer layers, each with a 12-head self-attention block [30]. ViVit captures spatio-temporal information by structuring data into 'tubes' across three dimensions: height, width, and temporal, which in this case were set to 16, 16, and 2, respectively [30]. This model was pretrained on the Kinetics-400 dataset and only the weights of the final layer were modified during training.

The model was used with the following parameters, which yielded optimal performance: a validation proportion of 0.3, a learning rate of 0.000005, and a maximum of 20 epochs with early stopping to prevent overfitting.

For the image model, we selected a pure convolutional-based architecture: ConvNeXt [23]. This choice, similar to our choice of ViVit, was driven by the recognition that ConvNets represent the current state-of-the-art in image classification [23]. Moreover, this model not only demonstrated competitive performance against transformers in terms of accuracy and scalability, but also, to the best of our knowledge, had limited research evaluating its performance in ICT classification, reinforcing our decision [23].

The specific ConvNeXt model being used was Facebook's convenext-tiny-224. As its name suggests, this is a small-scale model designed to deliver good performance while minimizing computational expense. It was pretrained on the ImageNet-1k dataset at a resolution of 224x224, and only the weights of the final layer were modified during training.

Since ConvNeXt is an image model, training was conducted using individual images as input rather than entire sequences. For classification, however, the model processes the complete sequence, classifying it as positive if at least one image is identified as positive. This approach was chosen to minimize false negatives, which are typically more concerning and undesirable than false positives in a medical context.

The model was used with the following parameters, which yielded optimal performance: a validation proportion of 0.3, a

learning rate of 0.000005, and a maximum of 20 epochs with early stopping to prevent overfitting.

C. Metrics

The following metrics were used to evaluate the models efficacy:

1) Confusion Matrix

A confusion matrix is a table used to evaluate the performance of a classification model by comparing actual target values with the model's predictions. It summarizes outcomes as true positives (TP), where ICH is correctly predicted; false positives (FP), where ICH is predicted but not present; true negatives (TN), where ICH is correctly predicted as absent; and false negatives (FN), where ICH is present but not predicted.

2) Accuracy

It represents the ratio of correctly predicted instances to the total number of instances in the dataset. Mathematically, accuracy can be defined as:

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

3) Precision

It measures the accuracy of positive predictions made by the model, indicating the proportion of instances classified as positive that are actually positive. Mathematically, precision can be defined as:

$$\text{Precision} = \frac{TP}{TP + FP}$$

4) Recall

Recall, also known as sensitivity or true positive rate, is a metric used to evaluate the performance in scenarios where the focus is on correctly identifying positive instances. It calculates the proportion of actual positive instances that were correctly identified by the model. Mathematically, recall can be defined as:

$$\text{Recall} = \frac{TP}{TP + FN}$$

D. Experiments

After the initial preprocessing was done, additional data handling was performed as required for each model. The environment settings and parameters were then configured, followed by the execution of the experimental procedure and a statistical analysis of the results. The following segments detail the specifications of the experimental environment and provide a comprehensive description of the experimental scenarios that were conducted.

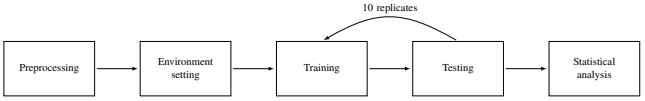

Fig. 3. Experiments flow.

The server used for running the experiments was composed of 2 Intel(R) Xeon(R) Gold 6240 CPU @ 2.60GHz, 2 NVIDIA Tesla V100S PCIE-32GB, 256Gb RAM, and 3TB storage. The server runned Clear Linux and used Slurm, a workload manager that allows users to schedule jobs and manage clusters of nodes. It also had access to a variety of software applications, including CUDA and Miniconda. For these experiments, a conda environment was set with pytorch.

Each replicate, or run, consisted of a training session followed by a testing session. In the training sessions, the dataset was randomly divided into a training subset and a validation subset, adhering to the proportions specified for each model. The model was then trained on the training subset for its designated number of epochs, optimizing its ability to classify ICT. In the testing session, the model's classification performance was evaluated using the validation subset, with all relevant metrics recorded. The experiment comprised a total of 10 replicates for model.

As previously mentioned, ViVit was trained using entire image sequences as input, aligning with its design as a video model capable of processing sequential data. In contrast, ConvNeXT was trained on individual images, as its structure as an image-based model does not support sequence-based training. This approach results in a larger training dataset for ConvNeXT compared to ViVit, which may had introduced bias. However, this strategy was maintained to ensure an objective comparison of their performance on this dataset and avoid introducing further experimental noise by manipulating data differently for each model.

E. Statistical analysis

To assess significant differences in model performance, the Shapiro-Wilk test was used to check normality, and Levene's test to check for homogeneity of variances. If both assumptions were satisfied, a t-test was conducted to compare the mean accuracies of the two models.

V. Results and analysis

This section presents the experimental results and provides an analysis to assess the performance of both models on the ICH sequence classification.

As discussed in the methodology, each model underwent 10 replicates, comprising a training phase and a testing phase, resulting in a total of 20 observations. It is important to note that the precision values from runs 2 and 7 of the ConvNeXt model were corrupted due to issues with the logging system and will therefore be considered 0 for the analysis.

Additionally, it is worth mentioning that multiclass precision and recall were used, as these metrics provide a more comprehensive view of both model's performance, evaluating their ability to classify both positive and negative cases.

First, the results of each model and metric were analyzed via a Shapiro-Wilk test for normality. Table I shows the results of the test. For ViVit, all metrics showed a possible normal distribution for an $\alpha = 0.05$. Nevertheless for ConvNeXt, precision wasn't normally distributed.

TABLE I
SHAPIRO-WILK TEST RESULTS.

Model	Metric	Result	P-value
ViVit	Accuracy	Passed	0.20
	Recall	Passed	0.91
	Precision	Passed	0.91
ConvNeXt	Accuracy	Passed	0.05
	Recall	Passed	0.16
	Precision	Failed	0.01

Note: Values are rounded to two decimal places.

In terms of the equality of variances, a similar situation to Shapiro-Wilk test happened. This can be observed in table II, which shows the Levene's test results. All metrics exhibited a possible equality except for the precision, that had a p-value lower than $\alpha = 0.05$.

TABLE II
LEVENE'S TEST RESULTS.

Metric	Result	P-value
Accuracy	Passed	0.69
Recall	Passed	0.25
Precision	Failed	0.02

Note: Values are rounded to two decimal places.

A t-test was conducted to assess the significant differences in the metrics of the models, as shown in table III. Also, the power for each metric is presented using a Cohen's d effect size. For both accuracy and recall, a Student's t-test was used, revealing a significant difference in the results. However, for precision, a Welch's t-test was performed due to unequal variances, despite the fact that the samples for ConvNeXt were not normally distributed. The results indicated no significant difference in precision.

TABLE III
T-TEST RESULTS.

Metric	Result	P-value	T-value	Power
Accuracy	Failed	0.00	4.20	1.00
Recall	Failed	0.00	7.82	1.00
Precision	Passed	0.68	0.43	0.54

Note: Values are rounded to two decimal places.

Since the precision samples for ConvNeXt were not normally distributed and the power of the t-test was reduced, we opted to perform a non-parametric test for all metrics to strengthen the validity of the subsequent analysis. The Mann-Whitney U test was conducted, and its results were consistent with those of the parametric t-test. These findings are presented in table IV.

TABLE IV
MANN-WHITNEY U TEST RESULTS.

Metric	Result	P-value	U-value
Accuracy	Failed	0.00	95.00
Recall	Failed	0.00	99.50
Precision	Passed	1.00	50.00

Note: Values are rounded to two decimal places.

Figure 4 shows a boxplot representing each model's performance on the accuracy. It is evident that ViVit outperformed ConvNeXt on this metric, exhibiting higher minimum, maximum, and median values. While both models exhibit similar variance, most of ViVit's observations are above 0.65, whereas the majority of ConvNeXT's observations fall below this threshold. This suggests that ViVit may have had a greater ability to accurately classify ICH sequences.

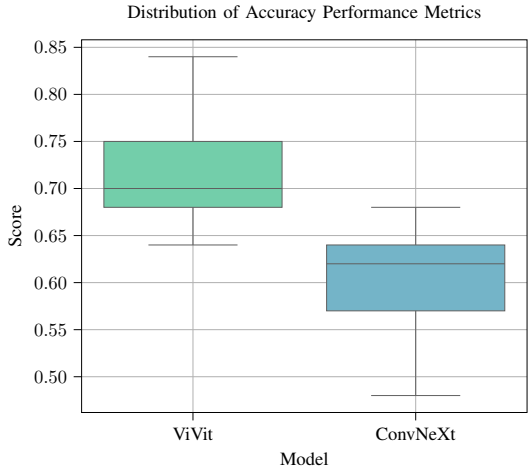

Fig. 4. Accuracy boxplots.

Similarly, Figure 5 presents the boxplot representation of each model's performance on precision. However, the results here are more complex. Both models achieved a maximum of score of 1, but ViVit's minimum score is significantly higher than ConvNeXt's, resulting in a much higher variance for the latter. While the median scores for both models were relatively close, ConvNeXt's median is notably higher.

This results suggests that although ConvNeXt may have had a greater potential for correct predictions, it also showed more variability and inconsistency in its results.

Once again, Figure 6 shows a boxplot representation of each model's performance, this time on the recall metric. The results are similar to those in Figure 4, with ViVit showing higher minimum, maximum, and median values than ConvNeXT. Although ConvNeXT exhibits lower variance, all its observations fall below 0.4, while most of ViVit's observations exceed this threshold. This indicates poorer performance by ConvNeXT, as the consistently low recall values may indicate a higher tendency for the model to misclassify sequences.

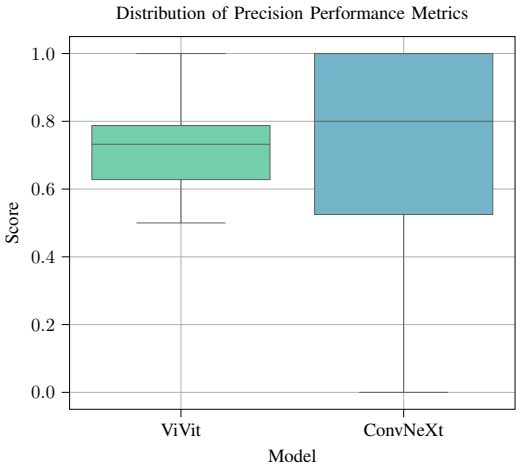

Fig. 5. Precision boxplots.

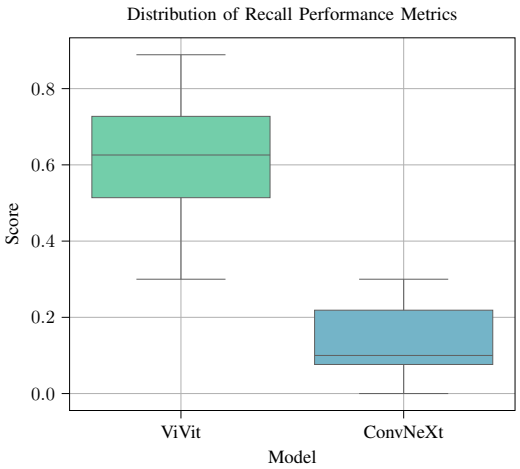

Fig. 6. Recall boxplots.

Table V provides a summary of the results shown in the boxplots, detailing the specific minimum and maximum values achieved by each model. Notably, Meanwhile, Table VI highlights the mean values, emphasizing ViVit's superiority across all three metrics. However, it is important to note that the mean scores were significantly influenced by outliers.

TABLE V
MIN AND MAX METRICS OF EACH MODEL.

Model	Accuracy		Precision		Recall	
	Min	Max	Min	Max	Min	Max
ViVit	**0.64**	**0.84**	**0.50**	**1.00**	**0.30**	**0.89**
ConvNeXt	0.48	0.68	0.00	**1.00**	0.00	0.30

Note: Values are rounded to two decimal places.

On the other hand, Figures 7 and 8 display the confusion matrices for ViVit and ConvNeXt, respectively. A key focus here is the number of false negatives, as these can lead to missed diagnoses, potentially overlooking the presence of ICH and resulting in fatal consequences for patients. ViVit significantly outperformed ConvNeXt in this regard, producing less than half the number of false negatives.

TABLE VI
MEAN METRICS OF EACH MODEL.

Model	Accuracy	Precision	Recall
ViVit	**0.72**	**0.73**	**0.62**
ConvNeXt	0.60	0.67	0.13

Note: Values are rounded to two decimal places.

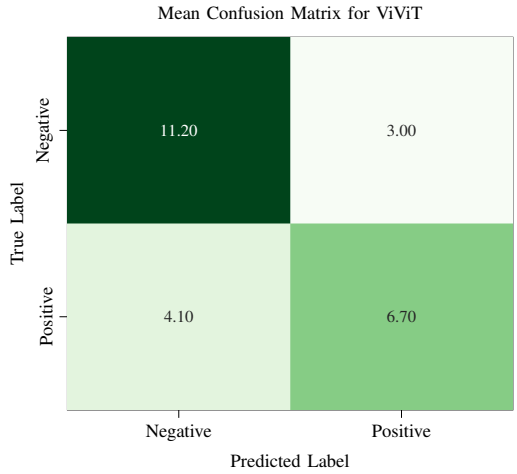

Fig. 7. Mean confusion matrix of ViVit.

Additionally, ViVit correctly predicted positive cases more than four times as often as ConvNeXt. However, it is also notable that ConvNeXt correctly predicted more negative cases than ViVit and produced six times fewer false positives.

VI. CONCLUSIONS

This section outlines the key findings from our experiments, highlights the main contributions and impact of this study on both ICH classification and the use of video models in medical contexts, and provides recommendations for future research aimed at building upon these findings.

ViVit's superior performance in accuracy and recall can be attributed to its ability to model temporal information inherent in video sequences. This temporal dimension likely played a significant role in ViVit's strong classification results, as it was able to capture the dynamics of ICH progression across time, which may be challenging for models that only analyze individual frames.

On the other hand, ConvNeXt, trained on individual images rather than image sequences, faced challenges due to the imbalanced nature of the dataset. The model was exposed to a disproportionate number of negative cases compared to positive ones, which likely contributed to its lower performance in terms of recall and overall misclassification of ICH cases. This imbalance could be alleviated by adopting methods such

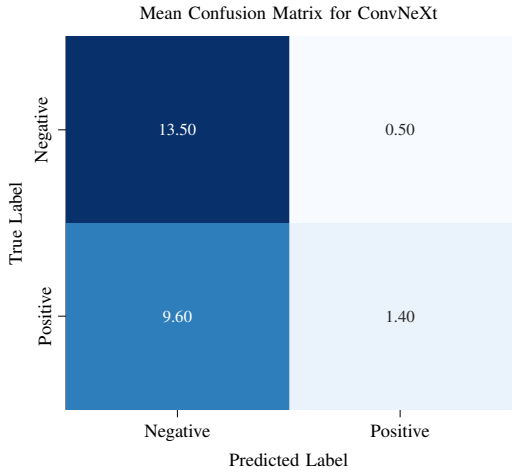

Fig. 8. Mean confusion matrix of ConvNeXt.

as data augmentation, oversampling, or using a loss function that accounts for class imbalance.

Even state-of-the-art image models like ConvNeXt can struggle with tasks requiring temporal understanding. ConvNeXt, optimized for single-image classification, is not inherently suited for image sequence analysis. This highlights a key limitation when applying standard image models to dynamic, time-dependent tasks like medical sequence classification. Future research may benefit from focusing on models designed for temporal information, such as video models or multi-view models.

To overcome the challenges observed in this study, future experiments should explore alternative approaches. One direction is to use ensemble models or multi-view architectures that can better handle the temporal and spatial complexities of medical image sequences. Additionally, incorporating larger and more diverse datasets with a greater number of patients would help to improve model generalization and robustness.

Given the potential of video models like ViVit, it would be worthwhile to repeat the experiment using other video-based models and compare their performance against more appropriate models designed for the task. An ensemble or multi-view approach might better suit the needs of medical image sequence classification, providing more robust results and reducing the inherent challenges posed by imbalanced data.

In summary, this study demonstrates the promise of using video models for medical image sequence classification, while also highlighting some limitations of current image-based models like ConvNeXt in this context. Future work should address data imbalance, leverage the temporal dimension of medical sequences, and explore alternative model architectures, such as multi-view or ensemble approaches, to further enhance performance in medical sequence classification tasks.

References

[1] J. A. Caceres and J. N. Goldstein, "Intracranial hemorrhage," *Emergency Medicine Clinics of North America*, vol. 30, no. 3, pp. 771–794, 2012, acute Ischemic Stroke.

[2] V. Abramova, A. Clèrigues, A. Quiles, D. G. Figueredo, Y. Silva, S. Pedraza, A. Oliver, and X. Lladó, "Hemorrhagic stroke lesion segmentation using a 3d u-net with squeeze-and-excitation blocks," *Computerized Medical Imaging and Graphics*, vol. 90, p. 101908, 2021.

[3] P. Thabarsa, S. Angkurawaranon, C. Madla, W. Vuthiwong, K. Unsrisong, and P. Inkeaw, "Classification of acute intracerebral hemorrhage using radiomics on brain computed tomography images," 2023, Conference paper, p. 299 – 304, cited by: 0.

[4] R. B. Domingues, C. Rossi, and C. Cordonnier, "Diagnostic evaluation for nontraumatic intracerebral hemorrhage," *Neurologic Clinics*, vol. 33, no. 2, pp. 315–328, May 2015.

[5] M. Grewal, M. M. Srivastava, P. Kumar, and S. 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)*, 2018, pp. 281–284.

[6] M. O. Khairandish, M. Sharma, and K. Kusrini, "The performance of brain tumor diagnosis based on machine learning techniques evaluation - a systematic review," in *2020 3rd International Conference on Information and Communications Technology (ICOIACT)*, 2020, pp. 115–119.

[7] S. E. Seymour, R. A. Rava, D. J. Swetz, A. Montiero, A. Baig, K. Schultz, K. V. Snyder, M. Waqas, J. M. Davies, E. I. Levy, A. H. Siddiqui, and C. N. Ionita, "Predicting hematoma expansion after spontaneous intracranial hemorrhage through a radiomics based model," vol. 12033, 2022, Conference paper, cited by: 3; All Open Access, Green Open Access.

[8] J.-W. Zhong, Y.-J. Jin, Z.-J. Song, B. Lin, X.-H. Lu, F. Chen, and L.-S. Tong, "Deep learning for automatically predicting early haematoma expansion in chinese patients," *Stroke and vascular neurology*, vol. 6, no. 4, pp. 610–614, 2021.

[9] J. Zhang, H. Zhang, L. Song, Y. Li, and P. Chen, "Improved 3d u-net model for precise brain hematoma segmentation and volume measurement," in *2023 8th International Conference on Intelligent Informatics and Biomedical Sciences (ICIIBMS)*, vol. 8, 2023, pp. 242–247.

[10] N. Yu, H. Yu, H. Li, N. Ma, C. Hu, and J. Wang, "A robust deep learning segmentation method for hematoma volumetric detection in intracerebral hemorrhage," *Stroke*, vol. 53, no. 1, pp. 167–176, Jan 2022, epub 2021 Oct 4.

[11] C. Ma, L. Wang, C. Gao, D. Liu, K. Yang, Z. Meng, S. Liang, Y. Zhang, and G. Wang, "Automatic and efficient prediction of hematoma expansion in patients with hypertensive intracerebral hemorrhage using deep learning based on ct images," *Journal of Personalized Medicine*, vol. 12, no. 5, p. 779, May 2022.

[12] S. Sarker, P. Sarker, G. Bebis, and A. Tavakkoli, "Mv-

swin-t: Mammogram classification with multi-view swin transformer," 2024.

[13] ——, "Mv-swin-t: Mammogram classification with multi-view swin transformer," 2024.

[14] H. Allaoui, Y. Alj, and Y. Ameskine, "Hybridmammonet: A hybrid cnn-vit architecture for multi-view mammography image classification," in *2024 IEEE 12th International Symposium on Signal, Image, Video and Communications (ISIVC)*, 2024, pp. 1–6.

[15] J. P. Howard, J. Tan, M. J. Shun-Shin, D. Mahdi, A. N. Nowbar, A. D. Arnold, Y. Ahmad, P. McCartney, M. Zolgharni, N. W. F. Linton, N. Sutaria, B. Rana, J. Mayet, D. Rueckert, G. D. Cole, and D. P. Francis, "Improving ultrasound video classification: an evaluation of novel deep learning methods in echocardiography," *Journal of Medical Artificial Intelligence*, vol. 3, no. 0, 2019.

[16] H. M, C. M, S. A, A. khafaji H, Y. Z, and G. B, "Computed tomography images for intracranial hemorrhage detection and segmentation," *PhysioNet*, 2020. [Online]. Available: https://physionet.org/content/ct-ich/1.3.1/

[17] S. Tenny and W. Thorell, "Intracranial Hemorrhage," in *StatPearls*. Treasure Island (FL): StatPearls Publishing, 2024.

[18] M. K. Nag, S. Koley, A. K. Sadhu, P. K. Dutta, B. Holsouser, S. Ashwal, and N. Ghosh, "A computer-aided tool for automatic volume estimation of hematoma using non-contrast brain ct scans," *Biomedical Physics and Engineering Express*, vol. 9, no. 4, 2023, cited by: 0.

[19] H. Yao, C. Williamson, J. Gryak, and K. Najarian, "Automated hematoma segmentation and outcome prediction for patients with traumatic brain injury," *Artificial Intelligence in Medicine*, vol. 107, p. 101910, 2020.

[20] L. Li, M. Wei, B. Liu, K. Atchaneeyasakul, F. Zhou, Z. Pan, S. A. Kumar, J. Y. Zhang, Y. Pu, D. S. Liebeskind, and F. Scalzo, "Deep learning for hemorrhagic lesion detection and segmentation on brain ct images," *IEEE Journal of Biomedical and Health Informatics*, vol. 25, no. 5, pp. 1646–1659, 2021.

[21] Z. Zhou, W. Chen, R. Yu, Y. Chen, X. Li, H. Zhou, Q. Fan, J. Wang, X. Wu, Y. Zhou, X. Zhou, and D. Guo, "He-mind: A model for automatically predicting hematoma expansion after spontaneous intracerebral hemorrhage," *European Journal of Radiology*, vol. 176, p. 111533, Jul 2024, epub 2024 May 25.

[22] R. Kumar, P. Kumbharkar, S. Vanam, and S. Sharma, "Medical images classification using deep learning: a survey," *Multimedia Tools and Applications*, vol. 83, no. 7, pp. 19 683–19 728, Feb. 2024, company: Springer Distributor: Springer Institution: Springer Label: Springer Number: 7 Publisher: Springer US.

[23] Z. Liu, H. Mao, C.-Y. Wu, C. Feichtenhofer, T. Darrell, and S. Xie, "A convnet for the 2020s," in *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022, pp. 11 966–11 976.

[24] Y. Nizamli, A. Y. Filatov, W. Fadel, and Y. A. Shichkina, "Accurate Anomaly Detection in Medical Images using Transfer Learning and Data Optimization: MRI and CT as Case Studies," in *2024 V International Conference on Neural Networks and Neurotechnologies (NeuroNT)*. Saint Petersburg, Russian Federation: IEEE, Jun. 2024, pp. 170–173.

[25] A. Panthakkan, S. M. Anzar, and W. Mansoor, "Unleashing the Power of EfficientNet-ConvNeXt Concatenation for Brain Tumor Classification," in *2023 15th Biomedical Engineering International Conference (BMEiCON)*. Tokyo, Japan: IEEE, Oct. 2023, pp. 1–5.

[26] G. Sharma, V. Anand, S. Malhotra, S. Kukreti, and S. Gupta, "NeuroSpectra: An Innovative Multi-Class Alzheimer's Disease Classification with ConvNeXt Transfer Learning Model," in *2023 International Conference on Artificial Intelligence for Innovations in Healthcare Industries (ICAIIHI)*. Raipur, India: IEEE, Dec. 2023, pp. 1–6.

[27] Devanshi, S. K. Baliarsingh, and P. P. Dev, "An Early diagnosis of diabetic retinopathy using ConvNeXt," in *2023 10th International Conference on Signal Processing and Integrated Networks (SPIN)*. Noida, India: IEEE, Mar. 2023, pp. 739–743.

[28] A. B. Reddy, B.-T. Pham, and J.-C. Wang, "Enhancing Breast Cancer Detection: A Novel Training Strategy and Batch Scheduler Method," in *2024 IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*. Niagara Falls, ON, Canada: IEEE, Jul. 2024, pp. 1–6.

[29] G. Bertasius, H. Wang, and L. Torresani, "Is space-time attention all you need for video understanding?" *CoRR*, vol. abs/2102.05095, 2021.

[30] A. Arnab, M. Dehghani, G. Heigold, C. Sun, M. Lucic, and C. Schmid, "Vivit: A video vision transformer," *CoRR*, vol. abs/2103.15691, 2021.

[31] Z. Tong, Y. Song, J. Wang, and L. Wang, "Videomae: Masked autoencoders are data-efficient learners for self-supervised video pre-training," *CoRR*, vol. abs/2203.12602, 2022.

[32] K. Li, X. Li, Y. Wang, Y. He, Y. Wang, L. Wang, and Y. Qiao, "Videomamba: State space model for efficient video understanding," 2024.