An Architecture based on CNNs and BiLSTMs for Slice-Level and Series-Level Intracranial Hemorrhage Identification in CT Scans

Daniel Henrique Comério[1], Karin Satie Komati[2], Thiago Oliveira-Santos[3], Filipe Mutz[4]
[1, 2]*Programa de Pós-Graduação em Computação Aplicada, Instituto Federal do Espírito Santo*, Serra, ES, Brasil
danielhcomerio@gmail.com[1], kkomati@ifes.edu.br[2]
[3, 4]*Departamento de Informática, Universidade Federal do Espírito Santo*, Vitória, ES, Brasil
todsantos@inf.ufes.br[3], filipe.mutz@ufes.br[4]

***Abstract*—This work proposes an architecture of neural networks for estimating the probability of intracranial hemorrhage and its subtypes in computed tomography images and series. The architecture consists of three stages, with the first being a CNN and the other two being BiLSTM recurrent networks. The first stage receives as input a CT image and returns the hemorrhages probabilities. The second stage improves these estimates using contextual information from neighboring images. The final stage integrates the predictions from all slices in order to provide an unified output for the series. Extensive experiments were performed using the datasets RNSA, CQ500, and PhysioNet for evaluating configurations of the architecture, improvements produced by each component, and generalization to new data. The best configuration uses the DenseNet-121 as backbone and achieved average accuracy, precision, recall and *f1-score* over datasets of 91%, 91%, 90% and 90% confirming the model's robustness and generalization.**

***Index Terms*—Convolutional neural networks, BiLSTM, Computed tomography, Intracranial hemorrhage**

I. Introdução

Hemorragias intracranianas (HICs) são uma condição médica grave que pode levar a sequelas graves ou até mesmo à morte se não tratadas de forma ágil. Estima-se que as HICs tenham uma incidência global de 24,6 por 100.000 pessoas por ano [1]. Essa condição ocorre quando há sangramento dentro do crânio e pode ser causada por uma variedade de fatores, incluindo trauma craniano, aneurismas, malformações arteriovenosas, hipertensão arterial e uso de drogas [2]. Estatísticas do Ministério da Saúde do Brasil indicam que as HICs são uma das principais causas de morte em adultos no país. Em 2018, foram registrados 197 mil atendimentos no Sistema Único de Saúde (SUS) relacionados à condição [3].

Exames de imagem, como tomografias computadorizadas (TCs), são comumente utilizados para diagnosticar HICs. No entanto, a interpretação desses exames por médicos radiologistas pode ser afetada por fatores como alta demanda de trabalho, falta de disponibilidade imediata e possíveis erros humanos [3]. Neste contexto, sistemas computacionais têm sido estudados com objetivo de auxiliar médicos na redução do tempo para diagnóstico de HICs. A título de exemplo, Arbabshirani et al. (2021) [4] mostraram que é possível reduzir o tempo para identificação de HICs usando modelos de aprendizado de máquina para ordenar a fila de TCs a serem analisados. No sistema proposto, a posição de um exame na fila depende da probabilidade de HICs atribuída pelos modelos (são avaliados primeiro exames com alta probabilidade de HICs) [4].

Buscando promover a pesquisa no tema, a Sociedade de Radiologia da América do Norte (*Radiological Society of North America* - RSNA) propôs em 2019 o *RSNA Intracranial Hemorrhage Detection Challenge*, um desafio em que times de pesquisadores competiam para desenvolver os algoritmos mais eficazes na tarefa de identificar e classificar HICs. Como parte da competição, foi construída uma base de dados pública e de larga escala contendo mais de 21 mil TCs, totalizando mais de 752 mil imagens (*slices*), cada um deles com anotações informando sobre a existência ou não de diferentes tipos de hemorragias [5].

A arquitetura utilizada pela equipe vencedora da competição [2] e por trabalhos posteriores com alta performance [1], [4] iniciam com a classificação das imagens das TCs usando redes neurais convolucionais e prosseguem com o uso de outros modelos de aprendizado de máquina para melhorar as predições por imagem usando dados adicionais como informações contextuais (e.g., predições feitas para as imagens vizinhas) e metadados associados às imagens (e.g., altura relativa da imagem). Por fim, a presença ou não de HICs nos exames é predita usando regras sobre as predições por imagem.

Partindo do pressuposto que inferências incorretas sobre uma ou mais imagens do exame podem levar à identificação incorreta de HICs nos exames, este trabalho propõe e avalia uma arquitetura inteiramente composta por redes neurais que, diferente dos trabalhos anteriores, utiliza uma rede *Bidirectional Long Short-Term Memory* (BiLSTM) [6] para integrar as análises por imagem e inferir a probabilidade de HICs e

*Os autores agradecem à FAPES e CAPES pelo PDPG (Programa de Desenvolvimento de Pós-Graduação - Parcerias Estratégicas nos Estados, processo 2021-2S6CD, FAPES nº132/2021). A professora Karin Komati agradece ao CNPq pela Bolsa de Produtividade DT-2 (308432/2020-7) e pelo projeto 407742/2022-0, também agradece à FAPES pelo Auxílio Taxa de Pesquisa (nº 293/2021) e pelo projeto nº1023/2022 P:2022-8TZV6.

seus subtipos nos exames.

Foi realizada uma extensa avaliação experimental para comparar a acurácia de diversas configurações de modelos e os ganhos trazidos por cada componente adicionado à arquitetura. Vale ressaltar que neste processo também foram avaliadas redes convolucionais diferentes daqueles usadas em trabalhos anteriores. Por fim, foram realizados experimentos para avaliar a capacidade de generalização para diferentes bases de dados, inclusive uma contendo TCs com características diferentes daquelas encontradas na base de treino. Resultados experimentais demonstraram o potencial da arquitetura proposta e em todas as bases de dados foram encontradas configurações com performances superiores ao estado-da-arte.

II. Trabalhos Correlatos

O aumento gradativo do número de HICs ao longo dos anos [7] em associação com a competição realizada pela RSNA, e a base de dados construída no evento, levaram à um aumento no número de trabalhos buscando o desenvolvimento de métodos para auxiliar no diagnóstico da doença pela análise de exames de imagens.

Grewal et al. (2018) [8] desenvolveram uma arquitetura de redes neurais profundas denominada RADnet para identificar automaticamente hemorragias intracranianas. Eles relataram valores de sensitividade e precisão de 88,64% e 81,25%, respectivamente, em uma base de dados privada de 77 tomografias cerebrais lidas por três radiologistas.

Salehinejad et al. (2021) [1] propuseram uma arquitetura híbrida neural e baseada em *gradient boosting* para calcular a probabilidade de existirem HICs epidurais, intraparenquimatosas, intraventriculares, subaracnóideas e subdurais em imagens de TC. As imagens eram utilizadas como entrada para CNNs SE-ResNeXt-50 e SE-ResNeXt-101 para obter uma estimativa inicial das probabilidades dos tipos de hemorragia para cada imagem. Como normalmente realizado em *ensembles*, foi calculada a média das probabilidades entre redes. Em seguida, um ensemble de três métodos baseados em *gradient boosting* foi utilizado para melhorar as predições por imagem usando dados contextuais das imagens vizinhas. Por fim, para produzir uma predição para o exame, os vetores de probabilidades de todas as imagens eram submetidas a uma série de limiares, encontrados utilizando otimização Bayesiana. Se pelo menos uma imagem superasse o limiar, o exame era classificado como contendo aquele tipo de hemorragia. O modelo alcançou 98,4% para área abaixo da curva ROC, 98,8% de sensitividade e 98,0% de especificidade na base de dados RSNA, e 95,4% para área abaixo da curva ROC, de 91,3% para sensitividade e de 94,1% para especificidade em uma base privada.

Wang et al. (2021) [2] propuseram a arquitetura vencedora da competição RSNA. Inicialmente, redes convolucionais são usadas para predizer a existência de HIC e seus subtipos nas imagens dos CTs. Em seguida dois estágios baseados em redes recorrentes buscam melhorar as predições usando metadados, informações contextuais das imagens vizinhas e média adaptativa de modelos. O treinamento foi realizado usando a base RSNA. As áreas abaixo da curva ROC no conjunto de teste desta base foram de 98,8% para as hemorragias intracranianas, 98,4% para as epidurais, 99,2% para as intraparenquimatosas, 99,6% para as intraventriculares, 98,5% para as subaracnóideas e 98,3% para as subdurais. A arquitetura alcançou ainda áreas abaixo da curva ROC de 96,4% e 94,9% nas bases de dados PhysioNet-ICH e CQ500, respectivamente.

Arbabshirani et al. (2021) [4] usaram um *ensemble* de redes neurais convolucionais para analisar imagens de TC e priorizar as listas de trabalho de radiologistas com objetivo de reduzir o tempo de diagnóstico de HICs. Na avaliação experimental, foi utilizada uma base de dados privada com 46.583 TCs coletadas de várias instalações da organização de saúde *Geisinger*. As métricas alcançadas no conjunto de teste foram de 84,6% para área abaixo da curva ROC, 73,0% para sensitividade e 80,0% para especificidade. Em uma avaliação após a implementação do sistema utilizou 347 exames e as métricas alcançadas foram de 84,0% de acurácia, 70,0% de sensitividade e 87,0% de especificidade. Além disso, o tempo médio até o diagnóstico foi reduzido de 512 para 19 minutos.

III. Método Proposto

Esta seção descreve as bases de dados utilizadas no trabalho, a arquitetura de redes neurais proposta e a metodologia de treinamento e avaliação da arquitetura.

A. Bases de Dados

Três bases de dados foram utilizadas no trabalho, a saber, as bases RSNA [5], CQ500 [9], [10] e PhysioNet [11]. O quantitativo de imagens e seus rótulos estão resumidos na Tabela I. As bases de dados RNSA e PhysioNet possuem anotações informando sobre a existência ou não de hemorragias e seus subtipos por imagem (*slice*). Já a CQ500 possui apenas anotações por exame. Contudo, Reis et al. (2020) produziram anotações por imagem informando sobre a existência de diferentes tipos de hemorragias e *bounding boxes* delimitando as lesões [10].

1) RSNA: A base RSNA foi criada durante a competição criada pela organização com mesmo nome e rotulada por uma equipe de mais de sessenta radiologistas [5]. A base RSNA foi disponibilizada em dois conjuntos, um de treino e outro de teste, sendo que neste último só estão disponíveis as imagens, sem os seus respectivos rótulos. Assim, neste trabalho, apenas os dados de treino foram utilizados, que consistem em 21.784 TCs, resultando em um total de 752.803 imagens na base (posteriormente, eles foram organizados em conjuntos de treino, validação e teste sem interseção de pacientes). Ao falar sobre a base de dados daqui em diante, estaremos nos referindo à este conjunto.

2) CQ500: A CQ500 foi criada pelo *Centre for Advanced Research in Imaging, Neurosciences and Genomics*, localizado em *New Delhi*, na Índia, usando dados de vários centros radiológicos da região. Cada imagem da base foi rotulada por três radiologistas com experiência de 8, 12 e 20 anos na interpretação de TCs de cabeça. A base CQ500 também foi disponibilizada em dois conjuntos contendo cada um 214 e

277 exames, totalizando 491 exames e 193.317 imagens de TC.

3) PhysioNet: A PhysioNet foi originada no *Al Hilla Teaching Hospital*, no Iraque, e apresenta anotações detalhadas incluindo, além da identificação de subtipos de hemorragia e fraturas cranianas, máscaras de segmentação binárias delimitando regiões de hemorragia em cada imagem. A base PhysioNet é composta por 75 exames de tomografia computadorizada de pacientes com lesões cerebrais traumáticas. A média é de 30 *slices* por exame, e o número total é de aproximadamente 2.814 imagens.

TABLE I
QUANTITATIVO DAS BASES DE DADOS

Tipo de Imagem	Bases de Dados		
	RSNA	**CQ500**	**PhysioNet**
Todas	752.803	193.317	2.814
Saudáveis	644.870	174.543	2.496
Hemorrágicas	107.933	18.774	318
Epidurais	3.145	131	173
Intraparenquimatosas	36.118	6.323	73
Intraventriculares	26.205	2.348	24
Subaracnóideas	35.675	9.590	18
Subdurais	47.166	6.391	56

B. Preprocessamento

A base de dados RSNA foi utilizada no treinamento do modelo, enquanto as bases CQ500 e PhysioNet foram usadas para avaliar a generalização para novos dados. A base RSNA foi dividida em três conjuntos: treino, validação e teste, em proporções de 70%, 20% e 10%, respectivamente. A divisão foi realizada garantindo que todos os exames de um paciente (e, portanto, as imagens que compõe os exames) estivessem no mesmo conjunto. Dois exames da base estavam corrompidos e foram descartados.

Aplicou-se um filtro denominado *windowing* [1], [12] à cada imagem dos CTs. Este filtro enfatiza diferentes tipos de tecidos e permite analisar aspectos específicos da imagem. O filtro consiste em uma operação de *clipping* sobre os valores dos pixels na escala de Houndsfield, seguida de uma renormalização para mapear os valores para o intervalo [0, 255]. O filtro foi aplicado três vezes com diferentes valores de limiar superior e inferior e as imagens que foram empilhadas em três canais, resultando em uma imagem equivalente à uma com canais de cores RGB. Seguindo a prática radiológica, os limiares foram definidos em termos do centro e largura da janela de recorte. O primeiro evidencia os ossos (*bone window*), o segundo a região subdural (*subdural window*) e o terceiro o cérebro (*brain window*), com valores de centro de 600, 80 e 40, e valores de largura de 2800, 200 e 80, respectivamente.

A Figura 1 ilustra imagens das 3 bases de dados após o processamento descrito. Note como as imagens da base CQ500 são mais ruidosas que aquelas das bases RNSA e PhysioNet. Isto se deve ao fato de que um equipamento com maior resolução no eixo axial foi utilizado na base CQ500. Tal

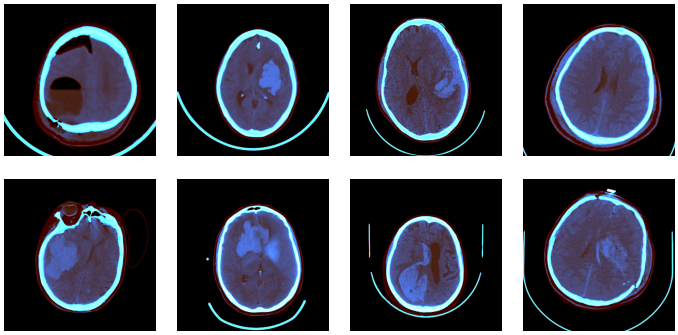

Fig. 1. Exemplos de slices de TC com hemorragia intracraniana das três bases de dados, após o pré-processamento. As duas primeiras colunas são de imagens da base RSNA, a terceira da CQ500 e a quarta da PhysioNet.

equipamento permite visualizar mais detalhes devido ao maior número de imagens por exame. Contudo, um efeito colateral negativo desta característica é a presença de ruído nas imagens.

Importante salientar que o número de *slices* por exame de cada uma bases de dados varia entre si, e até os exames de uma mesma base podem possuir quantidades diferentes. As bases RSNA e PhysioNet possuem uma média aproximada de 35 *slices* por exame. Já com relação à CQ500, alguns exames chegavam a possuir até 1.049 *slices*. Na construção das BiLSTMs de integração foi necessário aplicar uma técnica de redução para chegar-se aos 35. A técnica consistia em agrupar as predições desses *slices* em 35 grupos, respeitando-se a ordem axial. Grupos em que houvesse a predição de mais de um *slice*, eram substituídos por uma nova predição feita a partir do grupo, onde era vasculhado o maior valor para cada classe de hemorragia e inserido na nova predição. Já no caso dos exames com quantidade de *slices* inferior a 35, eram inseridos vetores contendo apenas zeros para completar.

C. Arquitetura Proposta

A arquitetura proposta é composta por três estágios em sequência, sendo que os dois primeiros produzem produzem predições por imagem, enquanto o terceiro produz predições por exame. A saída dos três estágios são iguais e consistem de 6 números, os 5 primeiros indicando as probabilidades de existirem HICs dos tipos epidurais, intraparenquimatosas, intraventriculares, subaracnóideas e subdurais e o último sendo a probabilidade de existir qualquer tipo de hemorragia. Cada valor pode ser interpretado como a saída de um classificador binário que tem como objetivo responder se existe ou não um dado tipo de hemorragia.

O primeiro estágio é dado por uma rede neural convolucional (CNNs) que recebe como entrada uma imagem preprocessada de uma TC. Para implementar este estágio, foram avaliadas os modelos DenseNet-121 e DenseNet-169 [13], ResNeXt–50 e ResNeXt–101 [14], e SqueezeNet-1.0 [15]. Todas estas redes foram pré-treinados na base de dados ImageNet [16]. A SqueezeNet foi escolhida por demandar menos recursos computacionais e ser uma alternativa economicamente vantajosa para organizações com recursos computacionais limitados.

No segundo estágio, as predições geradas pela CNN foram utilizadas como entrada para uma BiLSTM que tinha como objetivo de gerar uma nova predição aprimorada para cada imagem. Como em [1], foi utilizado um esquema de janela deslizante. Para cada imagem, a BiLSTM recebia como entrada uma sequência de nove vetores de probabilidades, sendo o do meio aquele que terá sua predição aprimorada, e os demais sendo referentes às 4 imagens abaixo e 4 acima (janela de tamanho 9). A melhoria nas predições acontece pela incorporação do informação contextual espacial entre imagens vizinhas no eixo axial. A título de ilustração, se as imagens acima e abaixo foram classificadas como tendo hemorragias com alta probabilidade e a do meio não, a sua predição pode ser atualizada para "concordar" com os vizinhos.

No terceiro estágio, uma segunda rede BiLSTM foi treinada para produzir uma predição para o exame como um todo. Ela recebe como entrada as predições aprimoradas produzidas na etapa anterior, integra estas informações e gera como saída as probabilidades do exame conter os diferentes tipos de HIC. Para treinar os dois primeiros estágios foram utilizadas anotações por imagem, enquanto que para treinar o terceiro estágio foram utilizadas anotações por exame. Assumimos que se pelo menos uma imagem possui um tipo de hemorragia, o exame também o contém.

Os mesmos hiperparâmetros foram utilizados para o treinamento dos três estágios. Os modelos foram treinados usando otimizador Adam por 20 épocas e taxa de aprendizado de 10^{-4}. A cada época, o conjunto de validação foi usado para calcular o *f1-score* para a saída que define se existe ou não hemorragia. O modelo relativo à época com maior *f1-score* foi retornado como resultado do processo de treinamento.

Além da avaliação das redes individuais, foram construídos *ensembles* dos modelos para cada uma das três fases. Para formar o *ensemble*, as predições das redes que o compunham foram calculadas e estratégias de sumarização foram utilizadas para produzir uma única saída a partir das respostas individuais. Os modelos foram treinados individualmente e depois usados para compor o *ensemble* sem treinamento conjunto.

Foram avaliadas três estratégias de sumarização: MaxVotos, Média e MaxProb. A estratégia MaxVotos retorna a classe mais votada pelos classificadores do *ensemble*. A estratégia Média calcula a média das saídas dos modelos, sendo similar a uma votação com pesos dados pela confiança das redes, e dando maior peso às predições dos modelos mais confiantes. A estratégia MaxProb retorna a classe com a maior probabilidade entre as saídas de todos os classificadores.

IV. Experimentos

Esta seção descreve os experimentos realizados e discute os resultados alcançados. O objetivo dos experimentos foi avaliar a arquitetura proposta e identificar as configurações com maior performance para incorporação em um sistema de diagnóstico assistido por computador. As métricas utilizadas foram acurácia (ACC), precisão (PRC), revocação (RVC) e *f1-score* (F1) da classe que define a presença ou ausência de hemorragia. Neste trabalho não avaliamos as métricas para os subtipos. Em alguns casos, a probabilidade de hemorragia para um dos subtipos era superior à da saída geral (que define se existe ou não HIC). Para acomodar estes casos, definimos que a probabilidade da saída geral seria dada pelo máximo entre o seu valor inicial e os valores dos subtipos. As principais métricas para a tarefa são a RVC que define o percentual de hemorragias identificadas e o F1 que considera também o número de falsos positivos.

TABLE II
Métricas das Redes Neurais Convolucionais

Modelo	Acurácia	Precisão	Revocação	F1-score
Redes Individuais				
DenseNet-121	94.64%	77.30%	89.55%	82.97%
DenseNet-169	**94.70%**	**77.35%**	89.99%	**83.19%**
ResNeXt-50	94.23%	75.53%	89.41%	81.88%
ResNeXt-101	94.30%	75.40%	**90.40%**	82.22%
SqueezeNet-1.0	93.38%	71.87%	89.69%	79.79%
Ensembles				
Média	**94.92%**	77.90%	90.91%	**83.90%**
MaxProb	94.79%	77.32%	**90.97%**	83.59%
MaxVotos	94.89%	**77.92%**	90.64%	83.80%

A. Comparação das CNNs

Este experimento avaliou a performance do primeiro estágio da arquitetura proposta, isto é, a performance das cinco CNNS listados na Seção III-C na tarefa de identificar HICs em imagens de TC. Para realização da comparação foi realizado o *fine-tuning* das redes pré-treinadas na ImageNet para a tarefa alvo. Os resultados deste experimento permitirão escolher qual das CNNs utilizar nos próximos experimentos.

A Tabela II compara as métricas das CNNs no conjunto de teste da base RNSA. As linhas representam os modelos e *ensembles*, enquanto as colunas representam as métricas. Os ensembles usaram as 5 CNNs. Como pode ser observado, o melhor *f1-score* dentre os modelos individuais foi alcançado pela DenseNet-169, com revocação e f1-*score* de 89.99% e 83.19%, respectivamente. Os *ensembles* com as três estratégias de sumarização levaram à pequenos aumentos das métricas, onde o *ensemble* de Média se destacou alcançando revocação e f1-*score* de 90.91% e 83.90%, respectivamente. Embora os modelos tenham alcançado uma alta acurácia, é importante lembrar que esta métrica é influenciada pelo desbalanceamento dos dados e a predominância de imagens sem HIC.

B. Melhoria de Predições usando Informação Contextual

Este experimento compara a melhoria na identificação de HICs por imagem usando a BiLSTM para incorporar informação contextual das predições vizinhas. Foram treinados e avaliados cinco modelos BiLSTM como modelos de correção de predições, um para cada CNN criada na etapa anterior.

A Tabela III apresenta as métricas obtidas. Como no caso anterior, linhas representam os modelos e *ensembles* utilizados, enquanto colunas representam as métricas. Comparando os resultados dos modelos de CNNs simples e seus *ensembles* com os modelos BiLSTM de correção, observamos melhoria significativa na métrica de precisão enquanto houve diminuição

TABLE III
MÉTRICAS DA BILSTM QUE OPERA SOBRE IMAGENS

Modelo	Acurácia	Precisão	Revocação	F1-score
Redes Individuais				
DenseNet-121	95.77%	87.46%	82.84%	85.09%
DenseNet-169	**95.81%**	**88.39%**	81.99%	85.07%
ResNeXt-50	95.43%	86.64%	81.16%	83.81%
ResNeXt-101	95.76%	86.43%	**84.07%**	**85.24%**
SqueezeNet-1.0	95.56%	86.80%	82.02%	84.34%
Ensembles				
Média	**96.01%**	88.42%	83.59%	**85.94%**
MaxProb	96.00%	**88.70%**	83.18%	85.85%
MaxVotos	95.98%	88.10%	**83.74%**	85.86%

TABLE IV
MÉTRICAS DOS MODELOS DE INTEGRAÇÃO

Modelo	Acurácia	Precisão	Revocação	F1-score
Redes Individuais				
DenseNet-121	94.15%	**93.46%**	92.42%	92.94%
DenseNet-169	**94.55%**	93.14%	**93.83%**	**93.49%**
ResNeXt-50	94.18%	94.27%	91.60%	92.92%
ResNeXt-101	94.21%	94.14%	91.83%	92.97%
SqueezeNet-1.0	94.00%	94.10%	91.31%	92.68%
Ensembles				
Média	94.52%	94.45%	92.27%	93.35%
MaxProb	94.40%	**94.77%**	91.60%	93.16%
MaxVotos	**94.55%**	94.32%	**92.50%**	**93.40%**

da revocação, porém o saldo dessas variações foi positivo resultando no aumento de *f1-score* devido ao maior equilíbrio entre as duas métricas. A BiLSTMs associada à ResNeXt-101 se sobressaiu tendo valores de acurácia, precisão, revocação e *f1-score* de 95.76%, 86.43%, 84.07% e 84.24%, respectivamente, apesar dos resultados dos demais modelos de correção terem sido bastante similares. Com relação aos *ensembles*, o de Média ainda permanece como a melhor técnica de sumarização, apesar do ganho de performance ser marginal em comparação com os modelos individuais.

Esses resultados demonstram que a utilização do contexto espacial dos *slices* foi eficaz em melhorar a taxa de acerto do modelo, refletindo em melhorias nas métricas de acurácia, precisão e *f1-score*. Esses resultados corroboram que usar os modelos BiLSTM para aprimorar as predições das CNNs é vantajoso.

C. Identificação de HICs a Nível de Exame

Este experimento avalia a performance do terceiro estágio da arquitetura que usa redes BiLSTM para integrar predições por imagem e produzir uma predição para o exame completo.

Foram treinados e avaliados cinco diferentes modelos BiLSTM para a tarefa de integração das predições, uma para cada rede do estágio anterior. Os resultados obtidos para cada um dos modelos BiLSTM de integração são apresentados na Tabela IV. As linhas representam os modelos BiLSTM de integração e seus *ensembles*, enquanto as colunas representam as métricas de desempenho. É importante frisar que os resultados desta tabela não é comparável com aqueles das tabelas anteriores, dado que esta traz resultados por exame, enquanto as outras trazem resultados por imagem.

Observou-se que os modelos identificaram hemorragias nos exames com alta precisão, revocação e *f1-score*. A DenseNet-169 destacou-se novamente com acurácia, precisão, revocação e *f1-score* de 94.55%, 93.14%, 93.83% e 93.49%, respectivamente, embora os outros modelos também apresentarem resultados similares. Quanto aos *ensembles*, MaxVotos foi superior, sem, contudo, melhorar significativamente as métricas em comparação aos modelos individuais. Os resultados sugerem que modelos de integração baseados em BiLSTM são eficazes na identificação de hemorragias intracranianas em exames.

Os modelos descritos até aqui foram aqueles com maior valor de *f1-score* no conjunto de validação. Porém, foram realizados experimentos com modelos de integração finais, após a última época do treinamento. O modelo baseado na DenseNet121 alcançou métricas de acurácia, precisão, revocação e *f1-score* de 94.68%, 92.17%, 95.32% e 93.72%. A revocação desse modelo no valor de 95.32% superou a do modelo campeão da competição da RSNA [2] que foi de 95.00%. Portanto, a arquitetura de modelo criada com o auxílio da técnica de integração foi capaz de elevar a performance, superando modelos do estado-da-arte nas métricas passíveis de comparação direta.

D. Generalização para Novas Bases de Dados

Tanto treinamento quanto avaliação foram realizados utilizando a base de dados RSNA e alcançaram *f1-score* superior a 93% nesta base. Com o intuito de avaliar se esta performance se mantém em face de novos dados, foram realizados experimentos usando as bases de dados CQ500 e PhysioNet.

As métricas para os modelos relativos à arquitetura completa (3 estágios) e os *ensembles* construídos usando estes modelos e as três estratégias de sumarização foram calculadas nas duas bases de dados. A Tabela V traz os resultados para os modelos que alcançaram os melhores *f1-score* em cada base. Linhas representam os melhores modelos para a base RSNA, CQ500 e PhysioNet, respectivamente. Colunas mostram as métricas para cada uma das bases avaliadas, onde a última, nomeada como "Média", corresponde à média aritmética simples dos resultados das métricas do modelo para cada base. As redes DenseNet-169 e DenseNet-121 descritas na tabela correspondem ao modelo com os três estágios individuais, já o modelo MaxProb refere-se ao *ensemble* gerado utilizando-se todos os modelos BiLSTM de integração.

Ao analisar as métricas da Tabela V, é possível notar que os modelos tiveram performance superior nas bases RSNA e PhysioNet se comparado à CQ500. Além disso, apesar dos modelos terem resultados similares para a base RSNA, nas demais ocorre uma leve oscilação. O modelo que obteve os resultados mais consistentes para as três bases de dados foi a BiLSTM de integração baseada na rede DenseNet-121 com médias de acurácia, precisão, revocação e *f1-score* de 91%, 91%, 90% e 90%, respectivamente. É interessante comentar que o modelo DenseNet-121 obteve resultado para a métrica de revocação levemente superior ao ganhador da competição da RSNA [2] para a base PhysioNet, sendo 88.9% de revocação contra 88.7%. É importante salientar que o foco ao gerar os

TABLE V
COMPARAÇÃO DOS MELHORES MODELOS DE CADA BASE DE DADOS

Modelo	RSNA				CQ500				PhysioNet				Média			
	ACC	PRC	RVC	F1	ACC	PRC	RVC	F1	ACC	PRC	RVC	F1	ACC	PRC	RVC	F1
DenseNet-169	**95%**	93%	**94%**	**93%**	74%	65%	**91%**	76%	91%	97%	83%	90%	86%	85%	89%	86%
MaxProb	94%	**95%**	92%	**93%**	**84%**	**78%**	89%	**83%**	88%	87%	**89%**	88%	89%	86%	**90%**	88%
DenseNet-121	94%	94%	92%	**93%**	83%	**78%**	87%	**83%**	**95%**	**100%**	**89%**	**94%**	**91%**	**91%**	**90%**	**90%**

modelos deste trabalho era em maximizar a métrica de f1-score proporcionando um maior balanço entre precisão e revocação. Dados os resultados foi possível atestar a capacidade de generalização dos modelos gerados para além da base de treinamento.

V. CONCLUSÕES

Este estudo propôs e avaliou uma arquitetura neural com CNNs e BiLSTMs para a identificação de HICs em imagens e exames de TCs. A arquitetura inicia com uma CNN para estimar a probabilidade de hemorragias e seus subtipos em cada imagem. Em seguida, uma rede BiLSTM aprimora essas estimativas com dados contextuais das imagens adjacentes. Na etapa final, outra aplicação de BiLSTM integra essas predições para fornecer probabilidades por exame.

Resultados experimentais mostraram que o modelos alcançaram boa performance e foram capazes de generalizar para outras bases de dados. A arquitetura com melhor performance média entre bases utilizou como *backbone* uma DenseNet-121 e alcançou médias de acurácia, precisão, revocação e *f1-score* de 91%, 91%, 90% e 90%, respectivamente. A arquitetura proposta levou à resultados equivalentes ou superiores à de trabalhos do estado-da-arte. O uso de *ensembles* de modelos levou à ganhos pequenos de performance em relação aos modelos individuais nos três estágios.

Em trabalhos futuros, consideraremos a aplicação de técnicas de *domain adaptation* com o objetivo de aprimorar a generalização e a precisão diagnóstica dos modelos.

REFERENCES

[1] H. Salehinejad, J. Kitamura, N. Ditkofsky, A. Lin, A. Bharatha, S. Suthiphosuwan, H.-M. Lin, J. Wilson, M. Mamdani, and E. Colak, "A real-world demonstration of machine learning generalizability in the detection of intracranial hemorrhage on head computerized tomography," *Scientific Reports*, vol. 11, 08 2021.

[2] X. Wang, T. Shen, S. Yang, J. Lan, Y. Xu, M. Wang, J. Zhang, and X. Han, "A deep learning algorithm for automatic detection and classification of acute intracranial hemorrhages in head ct scans," *NeuroImage: Clinical*, 08 2021.

[3] Brasil and M. da Saúde, "Linha de cuidados em acidente vascular cerebral (avc) na rede de atenção às urgências e emergências," 2012.

[4] M. R. Arbabshirani, B. K. Fornwalt, G. J. Mongelluzzo, J. D. Suever, B. D. Geise, A. A. Patel, and G. 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*, vol. 1, no. 1, pp. 1–7, 2018.

[5] A. Flanders, L. Prevedello, G. Shih, S. Halabi, J. Kalpathy-Cramer, R. Ball, J. Mongan, A. Stein, F. Kitamura, M. Lungren, G. Choudhary, L. Cala, L. Coelho, M. Mogensen, F. Morón, E. Miller, I. Ikuta, V. Zohrabian, O. McDonnell, and J. Nath, "Construction of a machine learning dataset through collaboration: The rsna 2019 brain ct hemorrhage challenge," *Radiology: Artificial Intelligence*, vol. 2, p. e190211, 05 2020.

[6] A. Graves, S. Fernández, and J. Schmidhuber, "Bidirectional lstm networks for improved phoneme classification and recognition," in *International conference on artificial neural networks*, pp. 799–804, Springer, 2005.

[7] V. Feigin, M. Brainin, B. Norrving, S. Martins, R. Sacco, W. Hacke, M. Fisher, J. Pandian, and P. Lindsay, "World stroke organization (WSO): Global stroke fact sheet 2022," *International Journal of Stroke*, vol. 17, pp. 18–29, 01 2022.

[8] 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)*, pp. 281–284, IEEE, 2018.

[9] S. Chilamkurthy, R. Ghosh, S. Tanamala, M. Biviji, N. G. Campeau, V. K. Venugopal, V. Mahajan, P. Rao, and P. Warier, "Deep learning algorithms for detection of critical findings in head ct scans: a retrospective study," *The Lancet*, vol. 392, no. 10162, pp. 2388–2396, 2018.

[10] E. P. Reis, F. Nascimento, M. Aranha, F. M. Secol, B. Machado, M. Felix, A. Stein, and E. Amaro, "Brain hemorrhage extended (bhx): Bounding box extrapolation from thick to thin slice ct images," *PhysioNet*, vol. 101, no. 23, pp. e215–20, 2020.

[11] M. Hssayeni, "Computed tomography images for intracranial hemorrhage detection and segmentation (version 1.3.1). physionet," 2020.

[12] T. M. Buzug, "Computed tomography," in *Springer handbook of medical technology*, pp. 311–342, Springer, 2011.

[13] G. Huang, Z. Liu, and K. Weinberger, "Densely connected convolutional networks," p. 12, 08 2016.

[14] S. Xie, R. Girshick, P. Dollar, Z. Tu, and K. He, "Aggregated residual transformations for deep neural networks," pp. 5987–5995, 07 2017.

[15] F. Iandola, S. Han, M. Moskewicz, K. Ashraf, W. Dally, and K. Keutzer, "Squeezenet: Alexnet-level accuracy with 50x fewer parameters and ¡0.5mb model size," 02 2016.

[16] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *2009 IEEE conference on computer vision and pattern recognition*, pp. 248–255, Ieee, 2009.