UNIVERSIDADE FEDERAL DO AMAZONAS

FACULDADE DE TECNOLOGIA

PROGRAMA DE PÓS-GRADUAÇÃO EM ENGENHARIA ELÉTRICA

AVALIANDO A GENERALIZAÇÃO DE MODELOS DE APRENDIZADO DE MÁQUINA PARA PREVER A MORTALIDADE EM 14 DIAS EM PACIENTES COM TRAUMATISMO CRANIOENCEFÁLICO

Manaus - AM

Agosto de 2025

UNIVERSIDADE FEDERAL DO AMAZONAS

FÁBIO ARTHUR SOARES ARAUJO

AVALIANDO A GENERALIZAÇÃO DE MODELOS DE APRENDIZADO DE MÁQUINA PARA PREVER A MORTALIDADE EM 14 DIAS EM PACIENTES COM TRAUMATISMO CRANIOENCEFÁLICO

Dissertação de Mestrado apresentada a Programa de Pós-Graduação em Engenharia Elétrica, da Universidade Federal do Amazonas, como parte dos requisitos necessários à obtenção do título de Mestre em Engenharia Elétrica.

Orientador: Prof. Dr. Cícero Ferreira Fernandes Costa Filho

Coorientadora: Profa. Dra. Marly Guimarães Fernandes Costa

Manaus - AM

Agosto de 2025

Ficha Catalográfica

Elaborada automaticamente de acordo com os dados fornecidos pelo(a) autor(a).

A663a Araujo, Fábio Arthur Soares

Avaliando a generalização de modelos de aprendizado de máquina para prever a mortalidade em 14 dias em pacientes com traumatismo cranioencefálico / Fábio Arthur Soares Araujo. - 2025.

104 f. : il., p&b. ; 31 cm.

Orientador(a): Cícero Ferreira Fernandes Costa Filho.

Coorientador(a): Marly Guimarães Fernandes Costa .

Dissertação (mestrado) - Universidade Federal do Amazonas, Programa de Pós-Graduação em Engenharia Elétrica, Manaus, 2025.

1. Trauma cranioencefálico. 2. Aprendizado de máquina. 3. Mortalidade. 4. Rede neural convolucional. I. Costa Filho, Cícero Ferreira Fernandes. II. Costa, Marly Guimarães Fernandes. III. Universidade Federal do Amazonas. Programa de Pós-Graduação em Engenharia Elétrica. IV. Título

---

Ministério da Educação
Universidade Federal do Amazonas
Coordenação do Programa de Pós-Graduação em Engenharia Elétrica

FOLHA DE APROVAÇÃO

Poder Executivo Ministério da Educação
Universidade Federal do Amazonas
Faculdade de Tecnologia
Programa de Pós-graduação em Engenharia Elétrica

Pós-Graduação em Engenharia Elétrica. Av. General Rodrigo Octávio Jordão Ramos, nº 3.000 - Campus Universitário, Setor Norte - Coroado, Pavilhão do CETELI. Fone/Fax (92) 99271-8954 Ramal:2607. E-mail: ppgee@ufam.edu.br

FÁBIO ARTHUR SOARES ARAÚJO

AVALIANDO A GENERALIZAÇÃO DE MODELOS DE APRENDIZADO DE MÁQUINA PARA PREVER A MORTALIDADE EM 14 DIAS EM PACIENTES COM TRAUMATISMO CRANIOENCEFÁLICO

Dissertação apresentada ao Programa de Pós-Graduação em Engenharia Elétrica da Universidade Federal do Amazonas, como requisito parcial para obtenção do título de Mestre em Engenharia Elétrica na área de concentração Controle e Automação de Sistemas.

Aprovada em 28 de agosto de 2025.

BANCA EXAMINADORA
Prof. Dr. Cícero Ferreira Fernandes Costa Filho - Presidente
Prof. Dr. Robson Luís Oliveira de Amorim - Membro Titular 1 - Externo
Prof. Dr. Frederico da Silva Pinagé - Membro Titular 2 - Externo

Manaus, 14 de agosto de 2025.

Documento assinado eletronicamente por **Cícero Ferreira Fernandes Costa Filho**, **Professor do Magistério Superior**, em 29/08/2025, às 08:55, conforme horário oficial de Manaus, com fundamento no art. 6º, § 1º, do Decreto nº 8.539, de 8 de outubro de 2015 .

Documento assinado eletronicamente por **Frederico da Silva Pinagé**, **Professor do Magistério Superior**, em 29/08/2025, às 10:43, conforme horário oficial de Manaus, com fundamento no art. 6º, § 1º, do Decreto nº 8.539, de 8 de outubro de 2015 .

Documento assinado eletronicamente por **Robson Luis Oliveira de Amorim**, **Professor do Magistério Superior**, em 01/09/2025, às 12:31, conforme horário oficial de Manaus, com fundamento no art. 6º, § 1º, do Decreto nº 8.539, de 8 de outubro de 2015 .

A autenticidade deste documento pode ser conferida no site https://sei.ufam.edu.br/sei/controlador_externo.php?acao=documento_conferir&id_orgao_acesso_externo=0, informando o código verificador **2743231** e o código CRC **1CB79B3E**.

Av. General Rodrigo Octávio Jordão Ramos, nº 3.000 - Bairro Coroado Campus Universitário, Setor Norte - Telefone: 99271-8954
CEP 69080-900 Manaus/AM - Pavilhão do CETELI. E-mail: ppgee@ufam.edu.br

Referência: Processo nº 23105.035420/2025-09 SEI nº 2743231

Agradecimentos

Gostaria de agradecer aos meus familiares que estiveram comigo durante toda a minha vida e me acompanharam para eu ser a pessoas que me tornei hoje.

Agradeço aos meus amigos do grupo “TAVERNA” que sempre estão comigo nos bons e nos maus momentos proporcionando diversão e fuga de momentos ruins.

Agradeço aos meus outros amigos que fiz durante a caminhada a vida que sempre estão presentes para trocar conversas e compartilhar memes por redes sociais.

Agradeço a minha namorada por me apoiar quando estou com problemas e sempre me motivar a continuar escrevendo.

Agradeço aos meus professores por me guiarem nessa jornada e por terem aberto para mim o caminho da Inteligência Artificial, área essa que mudou minha vida e que seguirei durante minha carreira acadêmica e profissional.

"Eu penso em Miku Miku oo ee oo."

—Miku, Hatsune

Resumo

O Traumatismo Cranioencefálico (TCE) continua sendo uma das principais causas de morbidade e mortalidade em todo o mundo, com disparidades significativas nos desfechos influenciadas pelo acesso e infraestrutura regionais de saúde. Este estudo avalia o desempenho e a generalização de modelos de aprendizado de máquina para prever a mortalidade em 14 dias em pacientes com TCE usando conjuntos de dados de duas regiões brasileiras distintas: São Paulo, um centro urbano, e Manaus, um centro urbano isolado com desafios logísticos únicos. Até onde sabemos, esta pesquisa representa a primeira validação cruzada de modelos preditivos em dois conjuntos de dados dentro do mesmo país, ressaltando a necessidade crítica de abordagens localizadas na pesquisa sobre TCE. Nossos resultados indicam que, embora os modelos baseados em redes neurais convolucionais (CNN) tenham alcançado alto desempenho, com uma área sob a curva (AUC) de 0,90 em São Paulo e 0,93 em Manaus, o melhor modelo de São Paulo exibiu uma AUC notavelmente baixa quando aplicado ao conjunto de dados de Manaus. A incorporação de características específicas do contexto, como variáveis relacionadas à pandemia e o tempo entre o trauma e a admissão, aumentou significativamente a precisão do modelo, com o modelo de Manaus atingindo uma impressionante AUC de 0,98. Notavelmente, o estudo destaca as principais diferenças regionais nos preditores de mortalidade, com hipóxia e hipotensão sendo mais críticas em Manaus, enfatizando a importância de adaptar os modelos preditivos aos contextos locais. Nossos resultados indicam que os modelos baseados em CNN têm o potencial de aprimorar as previsões de mortalidade para pacientes com traumatismo cranioencefálico (TCE). Além disso, destacamos a necessidade de conduzir a validação trans regional e integrar variáveis locais para melhorar os desfechos dos pacientes em diferentes ambientes de saúde.

Palavras-chave: Trauma cranioencefálico, mortalidade, Rede neural convolucional, contextos locais.

Abstract

Traumatic Brain Injury (TBI) remains a leading cause of morbidity and mortality worldwide, with significant disparities in outcomes influenced by regional healthcare access and infrastructure. This study evaluates the performance and generalizability of machine learning models for predicting 14- day mortality in TBI patients using datasets from two distinct Brazilian regions: São Paulo, an urban center, and Manaus, an isolated urban center with unique logistical challenges. To our knowledge, this research represents the first cross-validation of predictive models across two datasets within the same country, underscoring the critical need for localized approaches in TBI research. Our findings indicate that while convolutional neural network (CNN)-based models achieved high performance, with an area under the curve (AUC) of 0.90 in São Paulo and 0.93 in Manaus, the best model from São Paulo exhibited a strikingly low AUC when applied to the Manaus dataset. The incorporation of context specific features, such as pandemic-related variables and time from trauma to admission, significantly enhanced model accuracy, with the Manaus model reaching an impressive AUC of 0.98. Notably, the study highlights key regional differences in predictors of mortality, with hypoxia and hypotension being more critical in Manaus, emphasizing the importance of tailoring predictive models to local contexts. Our results indicate that CNN-based models have the potential to enhance mortality predictions for patients with traumatic brain injury (TBI). Additionally, we highlighted the necessity of conducting cross-regional validation and integrating local variables to improve patient outcomes across different healthcare environments.

Keywords: Traumatic Brain Injury, mortality, convolutional neural network, local contexts.

Sumário

Lista de Figuras

Lista de Tabelas

Lista de Siglas

TCE	Traumatismo Cranioencefálico
LMICs	Países de Baixa e Média Renda (*Low- and Middle-Income Countries*)
GCS	Escala de Coma de Glasgow (*Glasgow Coma Scale*)
AUC	Área Sob a Curva ROC (*Area Under the Curve*)
CNN	Rede Neural Convolucional (*Convolutional Neural Network*)
1D-CNN	Rede Neural Convolucional Unidimensional
ML	Aprendizado de Máquina (*Machine Learning*)
ISS	Índice de Gravidade da Lesão (*Injury Severity Score*)
SDH	Hematoma Subdural
KNN	K-Vizinhos Mais Próximos (*K-Nearest Neighbors*)
SVM	Máquina de Vetores de Suporte (*Support Vector Machine*)
GOSE	Escala de Resultados de Glasgow Estendida (*Glasgow Outcome Scale Extended*)
TC	Tomografia Computadorizada
TTAS	Escala de Triagem e Acuidade de Taiwan (*Taiwan Triage and Acuity Scale*)
LR	Regressão Logística (*Logistic Regression*)
MLP	Perceptron Multicamadas (*Multilayer Perceptron*)
RF	Floresta Randômica (Random Forest)
ANN	Redes Neurais Artificiais (*Artificial Neural Networks*)
PT	Tempo de Protrombina
INR	Razão Normalizada Internacional (*International Normalized Ratio*)
XGBoost	*Extreme Gradient Boosting*

SHAP *SHapley Additive exPlanations*

ReLU Unidade Linear Retificada (*Rectified Linear Unit*)

RMSProp Propagação da Raiz Média Quadrática (*Root Mean Square Propagation*)

SGDM Gradiente Descendente Estocástico com Momento (*Stochastic Gradient Descent with Momentum*)

TSAH Hemorragia Subaracnóidea Traumática (*Traumatic Subarachnoid Hemorrhage*)

rAPTT Razão do Tempo de Tromboplastina Parcial Ativado

UTI Unidade de Terapia Intensiva

CSV Valores Separados por Vírgula (*Comma-Separated Values*)

Capítulo 1

Organização da Dissertação

Esta dissertação está organizada em cinco capítulos, além dos elementos introdutórios e finais.

- Capítulo 1 – Introdução: Apresenta o contexto do estudo, a motivação, os objetivos gerais e específicos, além da estrutura da dissertação.
- Capítulo 2 – Revisão da Literatura: Apresenta os principais trabalhos na área de pesquisa, onde serve como base para entender conceitos da área e ter noção daquilo que já foi pesquisado e os melhores resultados com cada técnica utilizada.
- Capítulo 3 – Fundamentos Teóricos: Descreve os principais conceitos e métodos utilizados na pesquisa, incluindo tópicos sobre aprendizado de máquina, redes neurais convolucionais, métodos de regularização, funções de ativação, técnicas de otimização e métricas de avaliação.
- Capítulo 4 – Materiais e Métodos: Detalha as bases de dados utilizadas, os processos de pré-processamento, os modelos de predição empregados, os procedimentos de ajuste de hiperparâmetros, as estratégias de treinamento e teste, bem como as métricas utilizadas para avaliação dos resultados.
- Capítulo 5 – Resultados e Discussões: Apresenta os resultados obtidos com a aplicação das diferentes estratégias de treinamento e os modelos avaliados, analisando o desempenho preditivo a partir das métricas definidas, além de discutir os achados com base na literatura.
- Capítulo 6 – Conclusão: Resume os principais resultados, discute as limitações do estudo, apresenta sugestões para trabalhos futuros e destaca as contribuições da pesquisa para a área de aplicação.

Ao final, são apresentadas as referências utilizadas ao longo do trabalho

Introdução

O traumatismo cranioencefálico (TCE) é uma condição neurológica grave que representa uma das principais causas de morbidade e mortalidade em todo o mundo. Estima-se que entre 64 a 69 milhões de pessoas sofram algum tipo de TCE anualmente, o que evidencia a dimensão do problema sob a ótica da saúde pública global (Dewan *et al*., 2019). Os mecanismos mais comuns de ocorrência envolvem acidentes de trânsito, quedas e agressões físicas, afetando indivíduos de todas as faixas etárias.

Em países de baixa e média renda (LMICs), como o Brasil, a situação é ainda mais crítica. Nessas regiões, a infraestrutura hospitalar limitada, a escassez de recursos humanos especializados e os entraves logísticos no transporte de pacientes dificultam a condução adequada dos casos (Amorim *et al*., 2019). No Brasil, país de dimensões continentais, as disparidades regionais ensejam diferentes realidades. Enquanto centros urbanos como São Paulo contam com hospitais de alta complexidade, serviços de neurocirurgia disponíveis 24 horas e acesso ágil à tomografia e cuidados intensivos, regiões mais isoladas, como o interior da Amazônia, enfrentam grandes desafios estruturais. Em locais como Manaus, por exemplo, pacientes oriundos do interior são frequentemente transportados por longas distâncias via fluvial ou aérea, com tempo médio de transferência superior a 60 horas (Nôvo *et al*., 2023), o que pode comprometer a efetividade do atendimento neurológico de emergência.

Diante desse cenário, a busca por ferramentas capazes de auxiliar na tomada de decisão clínica tem motivado o desenvolvimento de modelos preditivos baseados em dados clínicos e laboratoriais. Nos últimos anos, o uso de algoritmos de aprendizado de máquina tem se intensificado, dada sua capacidade de identificar padrões complexos e realizar previsões a partir de grandes volumes de dados heterogêneos (Raj *et al*., 2022; Tu *et al*., 2022). Estudos internacionais já demonstraram que tais modelos podem alcançar desempenho competitivo quando comparados a métodos estatísticos tradicionais como a regressão logística (Zimmerman *et al*.,2023; Senders *et al*., 2018).

Apesar desses avanços, dois gargalos científicos persistem. O primeiro refere-se à generalização dos modelos preditivos: muitos algoritmos apresentam excelente desempenho nos conjuntos de dados em que foram treinados, mas sofrem queda significativa de acurácia quando aplicados a populações distintas, com diferentes perfis clínicos e contextos assistenciais (Courville *et al*., 2023; Yuan *et al*., 2018). Esse

fenômeno é especialmente relevante em países como o Brasil, onde a heterogeneidade de acesso, infraestrutura e perfil sociodemográfico entre as regiões pode comprometer a robustez e a utilidade clínica de modelos desenvolvidos em ambientes específicos. O segundo diz respeito à lacuna de estudos que explorem arquiteturas de aprendizado profundo especialmente redes neurais convolucionais (CNNs) na predição de desfechos em TCE. A maioria das pesquisas ainda se baseia apenas em técnicas clássicas de *machine learning*, como regressão logística ou florestas aleatórias, que podem não capturar de forma tão eficiente padrões complexos e interdependentes presentes nos dados clínicos e de imagem.

Além disso, poucos estudos investigaram a contribuição de variáveis contextuais, como o tempo entre o trauma e a admissão hospitalar ou fatores relacionados à sobrecarga do sistema de saúde durante pandemias, na performance de modelos preditivos. A pandemia de COVID-19, por exemplo, alterou significativamente a dinâmica do atendimento emergencial em várias regiões do Brasil, especialmente na região Norte, impactando diretamente nos desfechos de pacientes com trauma cranioencefálico (Nôvo *et al*., 2023). A inclusão dessas variáveis pode oferecer ganhos substanciais em termos de acurácia e capacidade discriminativa dos modelos, sobretudo em ambientes de alta variabilidade.

Neste contexto, esta dissertação propõe a investigação da generalização de modelos baseados em aprendizado profundo especificamente redes neurais convolucionais unidimensionais (1D-CNN) na predição de mortalidade em 14 dias de pacientes com TCE, utilizando dados de dois cenários contrastantes no Brasil: o Hospital das Clínicas da Faculdade de Medicina da Universidade de São Paulo (centro urbano consolidado) e um hospital terciário em Manaus, capital do Amazonas (centro urbano isolado). A proposta se destaca por ser, até onde se tem conhecimento, a primeira avaliação sistemática da capacidade de generalização inter-regional de modelos de predição em TCE baseada em dados clínicos brasileiros.

A pesquisa aqui apresentada contempla não apenas a comparação entre técnicas clássicas de ML (como regressão logística, floresta aleatória e perceptron multicamadas) e modelos mais complexos baseados em CNNs, mas também analisa a influência de variáveis adicionais disponíveis apenas na base de dados de Manaus, como o tempo entre o trauma e a internação e o contexto pandêmico. O desempenho dos modelos é avaliado por meio de métricas padrões, como acurácia, F1-score e área sob a curva ROC (AUC),

considerando estratégias de treinamento e validação cruzada em ambos os conjuntos de dados.

Ao final, espera-se que os resultados obtidos nesta dissertação não apenas validem a eficácia dos modelos propostos, mas também contribuam com evidências sobre a importância de se considerar fatores regionais na construção de ferramentas preditivas para suporte clínico. Além disso, pretende-se demonstrar que a adaptação de modelos a contextos específicos, aliada à redução criteriosa do número de variáveis utilizadas, pode promover soluções mais práticas e viáveis para aplicação em ambientes hospitalares com limitações estruturais. A relevância deste estudo reside, portanto, na interseção entre inteligência artificial e saúde pública, apontando caminhos promissores para o uso de tecnologias avançadas em benefício de sistemas de saúde heterogêneos como o brasileiro.

1.1 Objetivos da Dissertação

Avaliar a capacidade de generalização de modelos de aprendizado de máquina e aprendizado profundo para predição de mortalidade em 14 dias de pacientes com traumatismo cranioencefálico, utilizando bases de dados obtidas em dois centros clínicos brasileiros, localizados em regiões com características sociodemográficas distintas (São Paulo e Manaus).

1.2 Objetivos Específicos

- Comparar o desempenho de modelos clássicos e profundos na predição de mortalidade em 14 dias para pacientes com TCE
- Avaliar a generalização dos modelos com testes cruzados entre duas bases de dados obtidas em diferentes regiões do Brasil.
- Obter métricas da inclusão de variáveis contextuais no treinamento
- Utilizar um grupo reduzido de variáveis, com maior importância preditiva, para uma predição eficiente em ambas as bases de dados.

Capítulo 2

Revisão da Literatura

Este capítulo apresenta a revisão bibliográfica realizada sobre o tema "predição de mortalidade em pacientes com traumatismo cranioencefálico (TCE) utilizando redes neurais". Um conjunto de onze artigos, publicados entre os anos de 2019 e 2023, foi selecionado a partir de buscas realizadas nas bases de dados *IEEE Xplore, PubMed* e *Web of Science*. Para fins de comparação com os objetivos da presente pesquisa, as buscas foram direcionadas para artigos que abordassem o uso de algoritmos de aprendizado de máquina na predição de mortalidade em cenários que houve o trauma cranioencefálico.

A análise dos artigos teve como foco a identificação das bases de dados, dos métodos empregados (modelagem, algoritmos de aprendizado de máquina, variáveis analisadas e métricas de avaliação) e dos resultados obtidos. Ao final desta revisão, será apresentada uma tabela consolidando os principais pontos analisados, seguida de uma discussão que destaca as lacunas na literatura e as motivações para a condução deste trabalho.

2.1 *Predictors of Mortality, Withdrawal of Life-Sustaining Measures, and Discharge Disposition in Octogenarians with Subdural Hematomas* *(KASHKOUSH et al., 2022)*

Este estudo analisou fatores prognósticos relacionados à mortalidade hospitalar, retirada de medidas de suporte vital e destino de alta em pacientes octogenários com hematomas subdurais (SDH). Utilizando um banco de dados multicêntrico entre 2017 e 2019, foram avaliados 3.279 casos de TCE em 75 centros, dos quais 695 eram de pacientes com mais de 79 anos e diagnóstico de SDH. As variáveis estudadas incluíram variáveis demográficas, histórico médico, uso de antiplaquetários/anticoagulantes e variáveis clínicas, como GCS, reatividade pupilar e ISS.

Os resultados identificaram que fatores como GCS < 13, pupilas não reativas, aumento do ISS, hemorragia intraventricular e intervenção neurocirúrgica estão associados à mortalidade ou transferência para cuidados paliativos. Outros fatores, como insuficiência cardíaca congestiva, hipotensão, GCS < 13 e intervenções neurocirúrgicas,

foram determinantes na retirada de suporte vital. Os modelos de regressão logística apresentaram um AUC de 0,89, indicando boa precisão preditiva.

Esses achados destacam a importância das características clínicas e comorbidades como determinantes cruciais na tomada de decisões médicas e no prognóstico de pacientes octogenários com SDH.

2.2 *Machine Learning Algorithms to Predict In-Hospital Mortality in Patients with Traumatic Brain Injury (HSU et al., 2021)*

Este estudo buscou prever a mortalidade intra-hospitalar em pacientes com traumatismo cranioencefálico (TCE) utilizando modelos de aprendizado de máquina. Foram analisados 3.331 casos entre 2008 e 2018, classificados como nível I ou II na escala *Taiwan Triage and Acuity Scale*. As variáveis avaliadas incluíram idade, gênero, GCS, ISS, sinais vitais e mortalidade hospitalar.

Sete algoritmos foram utilizados: J48, *Floresta randômica, Random Tree, REP Tree, K-Nearest Neighbors* (KNN)*, Naïve Bayes e Support Vector Machine* (SVM). O algoritmo J48 demonstrou o melhor desempenho, com uma taxa de acerto de 93,2%, F1-score de 92,9% e sucesso médio de 77,2%. As variáveis com maior poder de predição foram a escala GCS, seguida por ISS e pressão arterial sistólica.

Os resultados apontaram que valores de corte relevantes,÷ GCS ≤ 6, ISS > 24 e pressão sistólica ≤ 84 mmHg, estavam associados a uma maior probabilidade de mortalidade. Este estudo destaca a eficácia dos modelos baseados em aprendizado de máquina para suporte à decisão clínica em emergências, permitindo a identificação precoce de pacientes de pacientes com alto risco e a otimização de tratamentos.

2.3 *Mobile Telephone Follow-Up Assessment of Postdischarge Death and Disability Due to Trauma in Cameroon: A Prospective Cohort Study (DING et al., 2022)*

Este estudo avaliou a mortalidade e a deficiência relacionadas a traumas em pacientes no Camarões ao longo de seis meses após a alta hospitalar, utilizando ferramentas de acompanhamento por telefone móvel. A amostra incluiu 1.914 pacientes tratados em quatro hospitais nas regiões Litoral e Sudoeste entre 2019 e 2021, dos quais 1.304 foram acompanhados com sucesso.

Os pacientes foram avaliados em quatro momentos: duas semanas, um mês, três meses e seis meses após a alta, utilizando a *Glasgow Outcome Scale Extended* (GOSE). Os resultados revelaram que 90% das mortes ocorreram nas primeiras duas semanas, enquanto 22% dos pacientes ainda apresentavam deficiência severa após seis meses. A mortalidade foi associada a fatores como idade avançada, maior pontuação no *Injury Severity Score* (ISS) e lesões neurológicas, enquanto níveis educacionais mais altos estavam ligados a menores taxas de mortalidade e deficiência.

O estudo destacou a viabilidade do acompanhamento por telefone em ambientes de baixa renda e ressaltou a necessidade de desenvolver sistemas formais para melhorar os resultados pós-trauma em regiões com infraestrutura médica limitada.

2.4 *Evaluation of Computed Tomography Scoring Systems in the Prediction of Short-Term Mortality in Traumatic Brain Injury Patients from a Low- to Middle-Income Country (SOUZA et al., 2022)*

Este estudo analisou a precisão de diferentes sistemas de pontuação baseados em tomografia computadorizada (TC) para prever o risco de morte em curto prazo entre pacientes com traumatismo cranioencefálico (TCE) em países de baixa e média renda. A pesquisa envolveu 447 pacientes atendidos em um hospital terciário no Brasil, com idade média de 40 anos e uma maioria significativa de homens (85,5%).

Foram avaliados três sistemas de classificação: Marshall CT, Rotterdam CT e Helsinki CT. Os resultados indicaram que os escores de Rotterdam e Helsinki superaram o de Marshall na previsão de mortalidade, tanto em 14 dias quanto durante a internação hospitalar. As áreas sob a curva (AUC) para mortalidade em 14 dias foram 0,610 para Marshall, 0,762 para Rotterdam e 0,752 para Helsinki.

Quando combinados com outros fatores clínicos (como idade, pontuação de Glasgow Coma Scale – GCS, resposta pupilar, hipóxia e hipotensão), esses escores mostraram um aumento expressivo na capacidade de explicação: Marshall (+2%), Rotterdam (+13,4%) e Helsinki (+21,6%). Entre eles, o escore Helsinki destacou-se como o modelo mais consistente, apresentando melhor capacidade de discriminação e predição.

Esses achados reforçam a importância de validar externamente esses modelos para populações de países em desenvolvimento. Além disso, sugerem que o uso de sistemas modernos de pontuação pode otimizar a alocação de recursos e auxiliar na tomada de decisões clínicas no tratamento de pacientes com TCE.

2.5 *A Computer-Assisted System for Early Mortality Risk Prediction in Patients with Traumatic Brain Injury Using Artificial Intelligence Algorithms in Emergency Room Triage) (TU et al., 2022)*

O estudo propôs um sistema baseado em inteligência artificial para prever o risco de mortalidade hospitalar em pacientes com traumatismo cranioencefálico (TCE) durante a triagem nas salas de emergência. A pesquisa utilizou dados retrospectivos de 18.249 pacientes adultos com TCE, atendidos em três hospitais de Taiwan entre 2010 e 2019. Para construir o modelo preditivo, foram consideradas 12 variáveis clínicas, incluindo idade, escala de triagem TTAS, pontuação GCS, tamanho das pupilas e reflexo pupilar.

Seis algoritmos de aprendizado de máquina foram testados: regressão logística (LR), Árvore randômica, *Support Vector Machines* (SVM), *LightGBM, XGBoost* e Perceptron Multicamadas (MLP). Entre eles, o modelo de regressão logística apresentou o melhor desempenho, com uma área sob a curva (AUC) de 0,925, seguido por SVM (AUC = 0,920) e MLP (AUC = 0,893). Para melhorar o balanceamento das duas classes, mortalidade e sobrevivência, aumentando a precisão das previsões, foi utilizada a técnica de sobre amostragem SMOTE.

O sistema de predição foi integrado ao sistema de informação hospitalar, permitindo previsões em tempo real para apoiar decisões clínicas e informar os riscos aos familiares dos pacientes. Este estudo evidencia o potencial dos algoritmos de aprendizado de máquina para melhorar o processo de triagem em emergências, otimizar a alocação de recursos e aprimorar o cuidado aos pacientes com TCE.

2.6 *Learning Models for Traumatic Brain Injury Mortality Prediction on Pediatric Electronic Health Records (FONSECA et al., 2022)*

Este estudo explorou o uso de algoritmos de aprendizado de máquina para prever a mortalidade em crianças com traumatismo cranioencefálico, utilizando o conjunto de dados *Hackathon Pediatric Traumatic Brain Injury* (HPTBI). A análise incluiu informações de 300 pacientes pediátricos internados, com idade média de 7,2 anos. O banco de dados continha 96 variáveis, abrangendo dados demográficos, clínicos e achados de tomografia computadorizada (TC).

Quatro modelos de aprendizado de máquina foram avaliados: Arvore randômica (RF), *XGBoost, k-Nearest Neighbors* (KNN) e redes neurais artificiais (ANN). Esses modelos foram combinados com técnicas de seleção de características, como Análise de

Componentes Principais (PCA) e métodos baseados em gradiente. O modelo *XGBoost* apresentou o melhor desempenho, com uma área sob a curva (AUC) de 0,91, especialmente sem o uso de técnicas de seleção de características. Já o KNN mostrou bom desempenho quando associado ao método de seleção de Koehrsen.

Os resultados identificaram variáveis como reatividade pupilar, nutrição enteral, presença de edema cerebral e parada cardíaca como altamente relacionadas à mortalidade. Curiosamente, fatores tradicionais, como a Escala de Coma de Glasgow (GCS), tiveram menor importância nesse contexto pediátrico, evidenciando a complexidade e a heterogeneidade dessa população.

O estudo ressalta a necessidade de desenvolver modelos preditivos específicos para crianças, considerando as particularidades do desenvolvimento cerebral e as diferentes manifestações clínicas do TCE pediátrico. A aplicação desses modelos pode oferecer suporte essencial para decisões em cuidados intensivos pediátricos.

2.7 *Predicting Outcome in Patients with Brain Injury: Differences between Machine Learning versus Conventional Statistics (CERASA et al., 2022)*

Este estudo comparou métodos de aprendizado de máquina (ML) e abordagens estatísticas tradicionais, como a regressão logística (RL), na previsão de desfechos em pacientes com lesões cerebrais, incluindo traumatismo cranioencefálico (TCE) e acidente vascular cerebral (AVC). A análise considerou 13 estudos que aplicaram ambos os métodos para prever resultados como mortalidade e incapacidades funcionais.

Os resultados mostraram que algoritmos de ML, como redes neurais artificiais (ANN), máquinas de vetores de suporte (SVM) e florestas aleatórias (RF), não apresentaram vantagens consistentes sobre a RL em termos de precisão preditiva. No caso de TCE, as taxas de acurácia variaram entre 78% e 98%, enquanto que, para AVC, os valores ficaram entre 74% e 95%. Variáveis como Escala de Coma de Glasgow (GCS), idade, reatividade pupilar e hemorragias intracranianas foram destacadas como os fatores mais relevantes.

Os autores observaram que o desempenho do ML pode ser limitado em bancos de dados clínicos com poucas variáveis e forte dependência de operadores. Em contrapartida, métodos estatísticos convencionais oferecem maior interpretabilidade dos fatores preditivos, enquanto os algoritmos de ML são mais eficazes para identificar relações não lineares entre as variáveis.

A conclusão do estudo enfatiza a importância de incorporar dados de alta dimensionalidade, como neuroimagem e informações genéticas, para aproveitar melhor o potencial do ML em ambientes clínicos e melhorar a previsão de desfechos em pacientes com lesões cerebrais.

2.8 *Prediction Performance of the Machine Learning Model in Predicting Mortality Risk in Patients with Traumatic Brain Injuries: A Systematic Review and Meta-Analysis (WANG et al., 2023)*

Esta revisão sistemática realizou uma meta-análise de 47 pesquisas que investigaram o uso de algoritmos de aprendizado de máquina (ML) na previsão de mortalidade em pacientes com traumatismo cranioencefálico (TCE). A análise abrangeu dados de 2.080.819 indivíduos de diversas regiões, comparando modelos de ML com ferramentas tradicionais de pontuação clínica.

Foram avaliados 156 modelos preditivos, sendo 122 desenvolvidos recentemente e 34 já validados clinicamente como ferramentas tradicionais. Para mortalidade intra-hospitalar, os modelos de ML apresentaram um índice C médio de 0,86 (intervalo de confiança de 95%: 0,84-0,87), com sensibilidade de 0,79 e especificidade de 0,89. No caso da mortalidade extra-hospitalar, o índice C foi de 0,83, com sensibilidade de 0,74 e especificidade de 0,75. Os algoritmos mais utilizados incluíram máquinas de vetores de suporte, redes neurais artificiais, árvores de decisão e regressão logística.

Os fatores mais frequentemente utilizados como preditores foram pontuação na Escala de Coma de Glasgow, idade, classificação da tomografia computadorizada (TC), reflexos pupilares, níveis de glicose e pressão arterial sistólica. Embora os modelos de ML tenham demonstrado um desempenho ligeiramente superior às ferramentas tradicionais na predição extra-hospitalar, o estudo destacou a necessidade de padronizar a aplicação clínica desses algoritmos para aumentar sua eficácia.

Os autores concluíram que os modelos de ML podem ser ferramentas promissoras para prever mortalidade em casos de TCE, especialmente quando integram dados complexos, como imagens de TC. No entanto, sua implementação prática ainda enfrenta desafios devido à falta de consenso e à variabilidade entre os estudos analisados.

2.9 *Machine Learning Algorithms for Predicting Outcomes of Traumatic Brain Injury: A Systematic Review and Meta-Analysis (COURVILLE et al., 2023)*

Esta revisão sistemática realizou uma meta-análise de 15 pesquisas que investigaram o uso de algoritmos de aprendizado de máquina para prever desfechos em pacientes com traumatismo cranioencefálico. Os objetivos foram identificar os modelos de ML mais eficazes na previsão de mortalidade e resultados desfavoráveis, além de comparar sua precisão com métodos estatísticos tradicionais, como a regressão logística (LR).

Entre os algoritmos analisados estavam redes neurais artificiais, máquinas de vetores de suporte, florestas aleatórias e *Naïve Bayes*. Para a previsão de mortalidade, os modelos de ML demonstraram acurácia superior a 80%, com a SVM alcançando até 95,6% de precisão em alguns casos. Fatores como pontuação na Escala de Coma de Glasgow (GCS), idade, glicose sérica elevada e acidez láctica foram consistentemente associados a desfechos desfavoráveis, contribuindo para a otimização do desempenho dos modelos.

A meta-análise revelou que os algoritmos de ML, especialmente ANN e SVM, superaram a LR em termos de sensibilidade e especificidade, com as curvas ROC confirmando a superioridade dos modelos baseados em inteligência artificial. Apesar dos avanços, os autores destacaram a necessidade de padronizar as variáveis utilizadas como entrada do modelo preditivo e realizar validações externas para ampliar a aplicabilidade clínica.

O estudo concluiu que os algoritmos de ML são ferramentas promissoras para estratificação de risco e previsão de desfechos em TCE. No entanto, seu impacto clínico depende de uma maior integração de dados diversos, como imagens de tomografia computadorizada e informações laboratoriais, para maximizar sua eficácia em cenários reais.

2.10 *Machine Learning Approach for the Prediction of In-Hospital Mortality in Traumatic Brain Injury Using Bio-Clinical Markers at Presentation to the Emergency Department (MEKKODATHIL et al., 2023)*

Este estudo utilizou algoritmos de aprendizado de máquina para prever a mortalidade hospitalar em pacientes com traumatismo cranioencefálico, com base em marcadores bio-clínicos disponíveis no momento da admissão. A análise incluiu dados retrospectivos de 922 pacientes tratados no Hamad Trauma Center, no Catar, entre junho de 2016 e maio de 2021. Entre as variáveis analisadas estavam a Escala de Coma de

Glasgow (GCS), Índice de Gravidade de Lesão (ISS), tempo de protrombina (PT), INR, além de níveis séricos de sódio, potássio, magnésio e outros biomarcadores clínicos.

Quatro algoritmos foram avaliados: *Support Vector Machine*, Regressão Logística, Floresta randômica e *Extreme Gradient Boosting* (XGBoost). O modelo SVM obteve o melhor desempenho, alcançando uma área sob a curva ROC de 0,86, demonstrando superioridade em estabilidade e capacidade de generalização. Embora os modelos *XGBoost* e RF também tenham apresentado boas AUCs, mostraram sinais de sobre ajuste devido a discrepâncias significativas no valor da função de perda entre os conjuntos de treinamento e teste (79,5% e 41,8%, respectivamente).

Os principais fatores associados à predição de mortalidade foram: aPTT, INR, ácido láctico, ISS, PT e magnésio. O estudo destacou que o uso de modelos de ML, especialmente o SVM, pode ser uma ferramenta valiosa para identificar pacientes de alto risco, permitindo intervenções clínicas mais rápidas e eficazes em cenários de trauma.

2.11 *Predictors of Mortality, Withdrawal of Life-Sustaining Measures, and Discharge Disposition in Octogenarians with Subdural Hematomas (KASHKOUSH et al., 2023)*

Este estudo analisou os fatores preditivos de mortalidade, retirada de suporte vital e desfecho de alta em pacientes octogenários diagnosticados com hematomas subdurais (SDH). A pesquisa utilizou dados de 3.279 admissões por traumatismo cranioencefálico (TCE) entre 2017 e 2019, dos quais 695 pacientes tinham mais de 79 anos.

Os resultados indicaram que 22% dos pacientes evoluíram para mortalidade intra-hospitalar ou foram direcionados para cuidados paliativos. Além disso, 10% passaram por retirada de suporte vital. Fatores como pontuação na Escala de Coma de Glasgow (GCS) inferior a 13, ausência de reatividade pupilar, maior Índice de Gravidade de Lesão (ISS) e presença de hemorragias intraventriculares foram fortemente associados à mortalidade. No caso da retirada de suporte vital, os principais determinantes incluíram insuficiência cardíaca congestiva (CHF), hipotensão e GCS inferior a 13. Modelos de regressão logística apresentaram alta precisão preditiva, com AUC de 0,885 para mortalidade e 0,894 para retirada de suporte vital.

O estudo concluiu que variáveis clínicas e demográficas podem ser utilizadas para orientar decisões críticas, como intervenções neurocirúrgicas e manejo paliativo, particularmente em pacientes idosos com SDH, onde o prognóstico é frequentemente mais delicado.

Tabela 1: Tabela de revisão da literatura

Referência	Base de Dados	Variáveis Preditivas	Variáveis Preditas	Resultados
Kashkoush *et al*. (2022)	3.279 admissões por TCE em 45 centros de trauma nos EUA entre 2017 e 2019. Análise de 695 pacientes com 80 anos ou mais.	ECG, Reatividade pupilar, ISS, Uso de anticoagulantes/antiagregantes, Comorbidades (ex.: ICC, diabetes), Hemorragia intraventricular, Intervenção neurocirúrgica	Mortalidade hospitalar, alta hospitalar com cuidados paliativos, Retirada de medidas de suporte à vida	Predição de mortalidade: AUC = 0,885; retirada de suporte: AUC = 0,894
Hsu *et al*. (2021)	4.881 pacientes com TCE atendidos em um hospital de alta complexidade no norte de Taiwan de janeiro de 2008 a junho de 2018.	ECG, ISS, Pressão arterial sistólica, Frequência cardíaca, Diferença de pressão de pulso, Idade, Gênero	Mortalidade hospitalar	Melhor desempenho: Árvore J48 - AUC > 0,80; Acurácia = 93,2%
Ding *et al*. (2022)	4.881 pacientes com TCE atendidos no departamento de emergência em Taiwan de janeiro de 2008 a junho de 2018.	Idade, Gênero, Escolaridade, ISS, Tipo de fratura, Déficit neurológico, Mecanismo da lesão	Mortalidade pós-alta, Incapacidade funcional (GOSE)	OR = 2,44 (ISS), OR = 4,40 (déficit neurológico); Incapacidade severa: 22,1%; Boa recuperação: 70,3%

Souza *et al*. (2022)	447 pacientes com TCE tratados em hospital terciário da USP, Brasil, de janeiro de 2012 a dezembro de 2015.	Classificações de TC (Marshall, Rotterdam, Helsinki), idade, ECG, resposta pupilar, hipóxia, hipotensão, hemoglobina	Mortalidade em 14 dias, Mortalidade hospitalar	Marshall: AUC = 0,610/0,575; Rotterdam: 0,762/0,712; Helsinki: 0,752/0,716
Tu *et al*. (2022)	18.249 pacientes com TCE atendidos em 3 hospitais em Taiwan de 2010 a 2019.	Idade, Gênero, IMC, TTAS, FC, Temperatura, FR, ECG, Tamanho da pupila, Reflexo pupilar	Mortalidade hospitalar	Melhor modelo: Regressão logística - AUC = 0,925; SVM = 0,920; MLP = 0,893; XGBoost = 0,871; RF = 0,870; LightGBM = 0,851
Fonseca *et al*. (2022)	300 pacientes pediátricos com TCE do HPTBI Hackathon	Idade, Gênero, ECG, TC (ex.: edema cerebral, desvio de linha média), Nutrição enteral, Parada cardíaca, Pupilas fixas	Mortalidade na alta hospitalar	XGBoost = 0,91; KNN = 0,90 (com seleção de variáveis); RF = 0,85; ANN = 0,84
Cerasa *et al*. (2022)	Revisão de 13 estudos comparando ML com estatística tradicional em TCE e AVC	Idade, ECG, Resposta pupilar, Hemorragia subaracnóidea, Escolaridade, Hipotensão, Hiperglicemia, Coagulopatia	Mortalidade hospitalar, Recuperação funcional	AUC = 0,82

Wang *et al*. (2023)	Meta-análise de 47 estudos com 2.080.819 pacientes de diversas regiões	ECG, Idade, TC, Tamanho da pupila, Reflexo pupilar, Glicose, PAS	Mortalidade intra e extra-hospitalar	Intra: C-Index = 0,86; Sens. = 0,79; Esp. = 0,89; Extra: C-Index = 0,83; Sens. = 0,74; Esp. = 0,75
Courville *et al*. (2023)	Meta-análise de 15 estudos com 32.721 pacientes com TCE	Idade, ECG, Ácido sérico, Glicose anormal, Pupilas, achados radiológicos, Hora do atendimento	Mortalidade hospitalar, 14 dias, Desfechos adversos (GOS)	SVM ≈ 0,96; ANN ≈ 0,91; Árvore ≈ 0,89; Regressão logística ≈ 0,83
Mekkodathil *et al*. (2023)	922 pacientes com TCE internados no Centro de Trauma Hamad no Catar (2016-2021)	ECG, ISS, aPTT, PT, INR, Hemoglobina, Ácido lático, Sódio, Potássio, Cálcio, Magnésio, Fosfato, Bicarbonato	Mortalidade hospitalar	SVM = 0,86; RF = 0,86; XGBoost = 0,85; Regressão logística = 0,84
Cao *et al*. (2023)	545.388 pacientes com TCE grave isolado do banco TQIP (2013-2021)	Idade, ECG na admissão, AIS da cabeça, Hipotensão, Cirrose, Hematoma epidural, Índice de choque, Saturação de O_2, Temperatura, Transfusão de concentrado de hemácias	Mortalidade hospitalar	C-index: treino = 0,897; teste = 0,896; AUC (≤5 dias) = 0,917; (≤20 dias) = 0,813

Capítulo 3

Fundamentos Teóricos

A fundamentação teórica tem como objetivo apresentar os principais conceitos que embasam o desenvolvimento metodológico desta pesquisa. A utilização de técnicas de aprendizado de máquina no contexto da saúde, em especial na predição de desfechos clínicos, tem ganhado destaque nas últimas décadas, impulsionada pela disponibilidade de bases de dados clínicas estruturadas e pela evolução de métodos computacionais capazes de lidar com variáveis complexas e interdependentes.

Neste trabalho, investiga-se a aplicação de modelos de aprendizado supervisionado, para a tarefa de predição de mortalidade em até 14 dias em pacientes vítimas de traumatismo cranioencefálico. Com isso, esta seção descreve os fundamentos relacionados às principais técnicas utilizadas, abrangendo desde as arquiteturas das redes convolucionais e seus componentes internos (como camadas convolutivas, funções de ativação e técnicas de regularização), até os métodos de otimização utilizados no treinamento dos modelos.

Além disso, também são discutidas as métricas utilizadas para avaliação de desempenho, onde ao longo da seção, são utilizados estudos prévios como referência para justificar as escolhas metodológicas adotadas, consolidando o embasamento teórico necessário para a condução do trabalho.

3.1 Traumatismo Cranioencefálico

O Traumatismo Cranioencefálico é uma lesão física no cérebro causada por uma força externa, que pode resultar em alterações temporárias ou permanentes na função cerebral. Trata-se de um problema de saúde pública global, com estimativas indicando que entre 64 e 69 milhões de pessoas no mundo sofrem TCE a cada ano, sendo os acidentes de trânsito, quedas e violência as principais causas (Dewan *et al*., 2019).

Os impactos do TCE são particularmente graves em países de baixa e média renda, onde a limitação de recursos e a desigualdade no acesso a serviços especializados contribuem para piores desfechos clínicos (Amorim *et al*., 2019). Nessas regiões, a carência de infraestrutura adequada para o atendimento de urgência e emergência pode

ocasionar atrasos no diagnóstico e no início do tratamento, elevando os índices de morbimortalidade. No Brasil, por exemplo, regiões remotas como a Amazônia enfrentam desafios logísticos que dificultam o transporte ágil de pacientes até centros de referência, comprometendo a assistência em tempo oportuno (Nôvo *et al*., 2023).

Diversas variáveis clínicas são utilizadas na avaliação da gravidade do TCE e na estimativa do prognóstico, incluindo a Escala de Coma de Glasgow, a reatividade pupilar, a presença de hipotensão e hipoxia no momento da admissão, e achados tomográficos como o desvio da linha média (Steyerberg *et al*., 2008; Faried *et al*., 2018). A presença de hipotensão e hipoxia, em particular, tem sido amplamente associada à piora dos desfechos, sobretudo em pacientes com lesões graves e em ambientes com limitações pré-hospitalares (Solla *et al*., 2021; Abujaber *et al*., 2020).

Para responder à complexidade clínica do TCE, diversos modelos prognósticos foram desenvolvidos ao longo das últimas décadas, como os modelos *CRASH* e *IMPACT*. Contudo, a aplicabilidade desses modelos em contextos regionais distintos permanece limitada, uma vez que fatores como tempo de transporte, sobrecarga hospitalar e condições socioeconômicas variam significativamente entre centros urbanos e áreas periféricas (Zimmerman *et al*., 2023).

Com o avanço da ciência de dados, tem-se intensificado o uso de técnicas de aprendizado de máquina para identificar padrões prognósticos a partir de grandes volumes de dados clínicos e laboratoriais. Ainda assim, a qualidade e a padronização dos dados permanecem como entraves importantes, especialmente em sistemas de saúde com registros incompletos ou desatualizados (Guimarães *et al*., 2022; Warman *et al*., 2022).

3.2 Pré-Processamento

O pré-processamento de dados é uma etapa fundamental em qualquer pipeline de aprendizado de máquina, sendo responsável por preparar os dados brutos para a etapa de modelagem. Essa fase visa garantir que os dados estejam em um formato adequado, reduzindo ruídos, padronizando escalas e lidando com possíveis inconsistências ou lacunas que poderiam comprometer o desempenho dos algoritmos.

Neste trabalho, foram aplicadas duas estratégias principais de pré-processamento: a normalização dos dados e o preenchimento de valores ausentes, conforme descrito a seguir.

3.2.1 Normalização de dados

A normalização de variáveis contínuas foi realizada utilizando o método Min-Max, que transforma os valores das variáveis para um intervalo entre 0 e 1. Essa abordagem é especialmente útil para redes neurais profundas, como as redes convolucionais utilizadas neste estudo, uma vez que evita que atributos com grandes amplitudes dominem os pesos durante o processo de treinamento (Zhang *et al*., 2021). A Equação 1 descreve o cálculo:

$$x_{norm} = \frac{x - x_{min}}{x_{max} - x_{min}}$$

Onde x representa o valor original, x_{min} e x_{max} são o mínimo e máximo da variável, respectivamente.

A normalização é considerada boa prática em tarefas envolvendo atributos com escalas heterogêneas, como idade e tempo até a admissão hospitalar, reduzindo o risco de instabilidade na retropropagação e contribuindo para uma convergência mais rápida do modelo (Hsu *et al*., 2021).

3.2.2 Preenchimento com valores para colunas com variáveis ausentes

O segundo passo do pré-processamento consistiu no tratamento de valores ausentes. A presença de dados faltantes pode comprometer o desempenho dos algoritmos de aprendizado de máquina, especialmente em aplicações clínicas sensíveis. No caso da base de dados de São Paulo, os valores ausentes já haviam sido tratados anteriormente por Guimarães *et al.* (2022), que utilizou diferentes estratégias de imputação baseadas na natureza das variáveis: preenchimento com a média para variáveis numéricas, e preenchimento por métodos supervisionados (como árvores de decisão e KNN) para variáveis categóricas, conforme descrito em seu trabalho original.

Para a base de Manaus, optou-se por um preenchimento simplificado, utilizando a moda (valor mais frequente) nas variáveis categóricas. Essa estratégia é considerada simples, porém eficaz em manter a consistência dos dados sem introduzir viés significativo (Ding *et al*., 2022). Essa escolha se justifica pelo baixo percentual de dados ausentes neste conjunto, sendo inferior a 5% na maioria das variáveis, o que torna

desnecessária a aplicação de métodos mais sofisticados, como imputação múltipla, KNN-imputation ou algoritmos supervisionados, mais indicados quando há perdas superiores a 10% ou padrões não aleatórios de ausência (Little & Rubin, 2019).

3.3 Análise de correlação

A análise de correlação é uma etapa importante na compreensão da influência de variáveis de entrada sobre a variável-alvo. Essa etapa permite identificar relações lineares ou não lineares entre os atributos, auxiliando tanto na seleção de variáveis quanto na interpretação de resultados dos modelos preditivos.

Duas abordagens foram utilizadas neste estudo: a análise estatística clássica por meio do coeficiente de correlação de Pearson e uma análise baseada em interpretabilidade de modelos via *SHapley Additive exPlanations* (SHAP).

3.3.1 Coeficiente de Pearson

O coeficiente de correlação de Pearson (r) mede a intensidade e a direção da relação linear entre duas variáveis numéricas. Seu valor varia entre -1 e 1, indicando, respectivamente, correlação negativa perfeita, nenhuma correlação ou correlação positiva perfeita. Trata-se de uma das medidas mais tradicionais para análise de dependência entre variáveis, sendo amplamente empregada em estudos estatísticos e computacionais (Rodgers & Nicewander, 1988).

A Equação 1 expressa o cálculo de r:

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \cdot \sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

Onde:

- x_i e y_i são os valores das variáveis X e Y, respectivamente, para cada observação i;
- $\bar{x}$ e $\bar{y}$ representam as médias amostrais de X e Y;
- n é o número total de observações.

Nesta pesquisa, o coeficiente de Pearson foi utilizado como ferramenta de análise exploratória para avaliar o grau de correlação entre cada variável de entrada e a variável

de saída. Tal análise auxilia na identificação de atributos com maior relevância estatística potencial para o modelo preditivo.

3.3.2 *SHapley Additive exPlanations* (SHAP)

O *SHapley Additive exPlanations* é uma técnica de interpretabilidade de modelos baseada na teoria dos jogos cooperativos, especificamente nos valores de *Shapley*. Essa abordagem permite atribuir a cada variável de entrada uma contribuição justa e consistente para a predição de um modelo de aprendizado de máquina.

Formalmente, o valor de *Shapley* para uma variável i é definido como:

$$\phi_i = \sum_{\{S \subseteq N \setminus \{i\}\}} \frac{|S|! \cdot (|N| - |S| - 1)!}{|N|!} \quad [f(S \cup \{i\}) - f(S)]$$

Onde:

- N representa o conjunto de todas as variáveis preditoras;
- S é um subconjunto de Nque não inclui i;
- $F(S)$ é a predição do modelo considerando apenas as variáveis de S;
- $f(S \cup \{i\}) - f(S)$ é a contribuição marginal da variável i;
- O fator multiplicativo pondera cada subconjunto com base em seu tamanho.

Esse valor representa a média ponderada da contribuição marginal da variável i, considerando todos os contextos possíveis de interação com outras variáveis do modelo (Lundberg & Lee, 2017).

Entre as principais propriedades do SHAP destacam-se:

- Justiça e consistência: se a contribuição de uma variável aumenta em um modelo, seu valor SHAP também aumenta;
- Aditividade: a soma dos valores SHAP de todas as variáveis corresponde à diferença entre a predição da instância e a média global do modelo;
- Aplicabilidade genérica: o método é independente do tipo de modelo, podendo ser aplicado tanto em redes neurais quanto em árvores de decisão, regressões ou ensembles (XAI Tutorials, 2024).

Na prática, os valores de SHAP permitem gerar visualizações que tornam o comportamento do modelo mais transparente:

- O *bar plot* mostra a importância média absoluta de cada variável para todas as predições;
- O *summary plot* exibe a distribuição dos valores SHAP por variável, incluindo também a direção (positiva ou negativa) de cada impacto.

Essas representações são fundamentais em contextos onde a interpretabilidade é exigida, como aplicações clínicas, financeiras ou jurídicas.

3.4 Algoritmos Clássicos de Aprendizado de Máquina

Nesta seção, são descritos os dois algoritmos de aprendizado supervisionado utilizados como modelos comparativos em relação ao modelo principal baseado em redes neurais convolucionais. Os métodos escolhidos foram: Regressão Logística e Árvore Randômica (Floresta randômica), ambos amplamente utilizados em tarefas de classificação binária e conhecidos por sua robustez e interpretabilidade.

3.4.1 Regressor logístico

A regressão logística é um modelo estatístico amplamente utilizado para tarefas de classificação binária. Seu objetivo é estimar a probabilidade de uma observação pertencer a uma das duas classes possíveis, utilizando como função de ativação a sigmoide logística. A equação da regressão logística é dada por:

$$P(Y = 1 \mid X) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x_1 + \cdots + \beta_n x_n)}}$$

Onde:

- $P(Y = 1 \mid X)$ é a probabilidade prevista da classe positiva;
- β_0 é o intercepto (bias);
- $\beta_1, \cdots, \beta_n$ são os coeficientes associados às variáveis $x_1, \cdots, x_n$.

Durante o treinamento, os coeficientes são ajustados para minimizar a função de custo *log-loss* (entropia cruzada), que penaliza predições incorretas com maior

intensidade. Por ser um modelo linear, sua performance pode ser limitada em problemas com alta não-linearidade, mas apresenta bons resultados quando as variáveis são informativas e as classes são separáveis (Hosmer *et al*., 2013).

A Figura 1 ilustra o funcionamento do modelo:

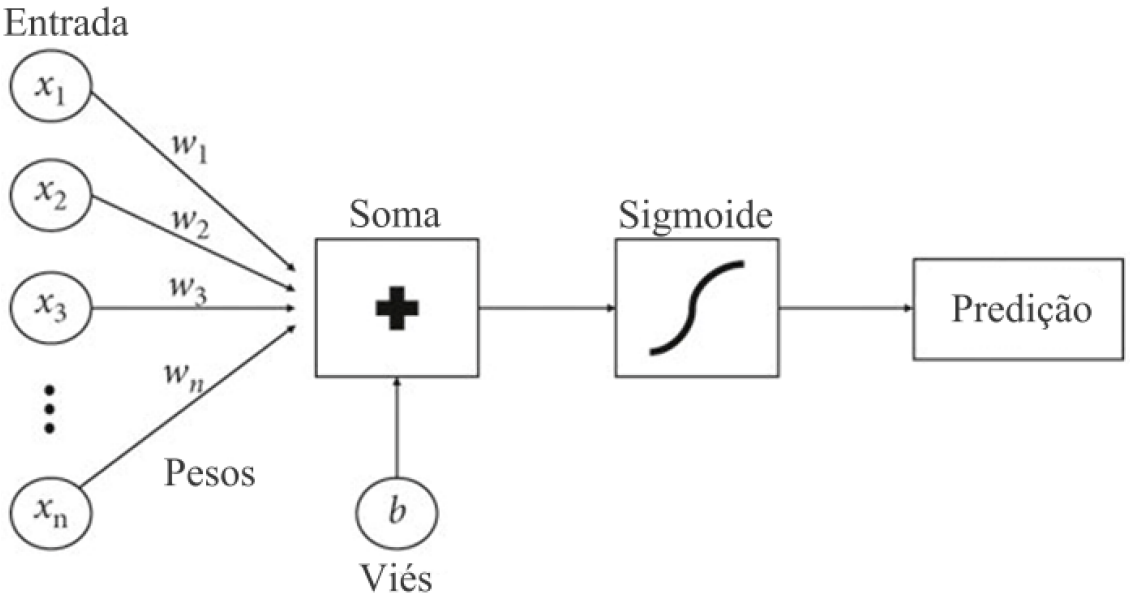


Figura 1: Arquitetura do regressor logístico

Fonte: adaptado de Khan *et al.* (2021)

3.4.2 Árvore randômica

O algoritmo de Árvore Randômica (Floresta randômica) é uma técnica de aprendizado de máquina baseada em ensemble learning, que combina a predição de múltiplas árvores de decisão para produzir um resultado mais robusto e estável. Cada árvore é construída a partir de um subconjunto aleatório do conjunto de dados de treinamento, e cada nó é dividido com base em um subconjunto aleatório de atributos.

Ao final do processo, a predição do modelo é obtida por meio de uma votação majoritária (no caso de classificação) ou pela média das saídas das árvores (no caso de regressão). Essa abordagem reduz a variância do modelo e aumenta sua capacidade de generalização, além de fornecer métricas de importância das variáveis utilizadas.

A Figura 2 ilustra o funcionamento geral do algoritmo:

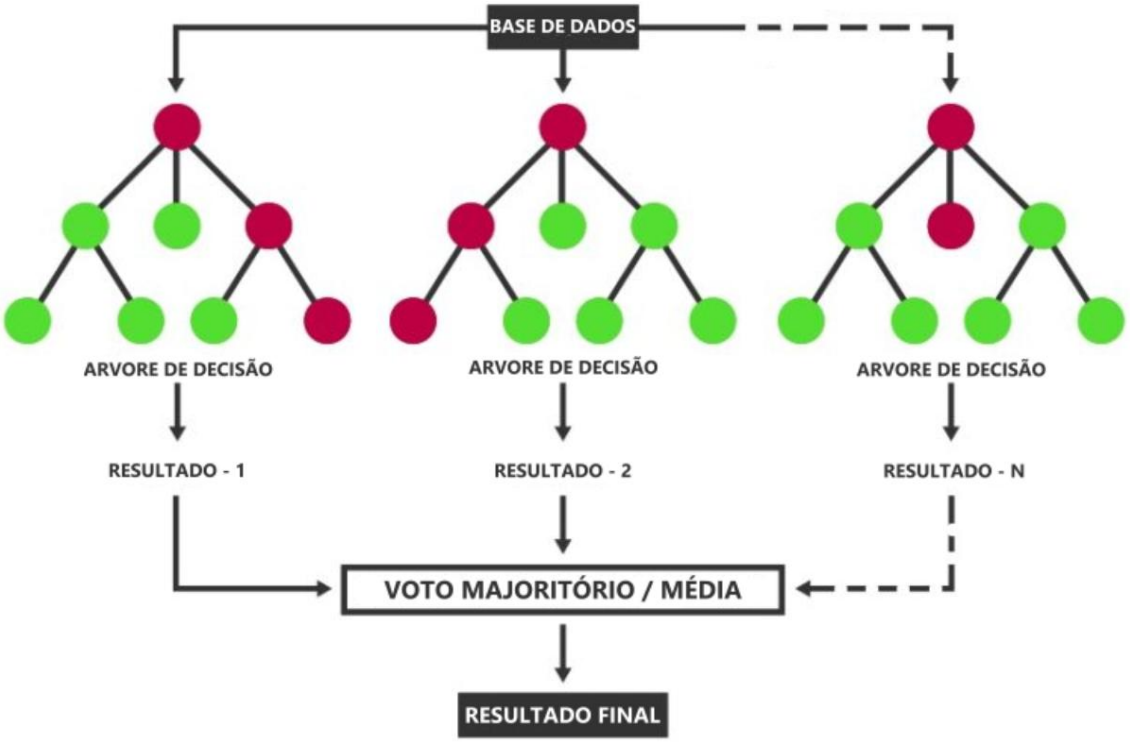


Figura 2: Algoritmo do modelo da arvore randômica

Fonte: adaptado de InfoAryan (2022).

Essa representação destaca o paralelismo das árvores e o processo de agregação dos resultados. A diversidade introduzida pelas amostras e atributos aleatórios ajuda a evitar o sobre ajuste (*overfitting*), uma limitação comum de modelos baseados em uma única árvore de decisão (Breiman, 2001).

3.5 Redes Neurais Artificiais

As Redes Neurais Artificiais (RNAs) constituem um dos pilares fundamentais do aprendizado profundo. Inspiradas no funcionamento do sistema nervoso biológico, essas redes são compostas por unidades de processamento denominadas neurônios artificiais, que se organizam em camadas conectadas entre si por pesos sinápticos. A primeira camada recebe os dados de entrada, enquanto as camadas intermediárias — chamadas ocultas — são responsáveis por extrair e transformar características progressivamente mais abstratas, até que a última camada forneça a saída desejada, seja uma classe, um valor ou uma distribuição.

Cada neurônio realiza uma combinação linear ponderada das entradas recebidas e, em seguida, aplica uma função de ativação não linear, como ReLU, Leaky ReLU,

tangente hiperbólica ou sigmoide, a fim de permitir que a rede aprenda relações complexas e não lineares. A equação geral da ativação de um neurônio é expressa por:

$$y = \phi\left(\sum_{i=1}^{n} w_i x_i + b\right)$$

$$(I * K)(i,j) = m\sum n\sum I(i+m, j+n) \cdot K(m,n)$$

O avanço das RNAs ao longo do tempo culminou no desenvolvimento das chamadas redes profundas que contam com múltiplas camadas ocultas, ampliando drasticamente a capacidade de modelagem e abstração. Tais modelos são atualmente aplicados em tarefas como reconhecimento de fala, diagnóstico médico, processamento de imagens e previsão temporal (Goodfellow *et al*., 2016; LeCun *et al*., 2015).

3.6 Rede Neuras Convolucionais

As Redes Neurais Convolucionais (CNNs) foram projetadas para lidar com dados que apresentam estrutura espacial, como imagens e sinais temporais. Ao contrário das RNAs tradicionais, em que todos os neurônios são conectados entre si, as CNNs utilizam camadas que operam localmente, reduzindo o número de parâmetros e permitindo o aprendizado eficiente de padrões espaciais. Desde a proposta da LeNet-5 por LeCun *et al*. (1998), essas redes tornaram-se a espinha dorsal de aplicações em visão computacional e processamento de sinais biomédicos.

As CNNs combinam diversas camadas especializadas, como convolutivas, de subamostragem, ativação, regularização e classificação, que, em conjunto, permitem a extração automática e hierárquica de características relevantes dos dados de entrada

3.6.1 Camada Convolutiva

A camada Convolutiva constitui o alicerce das redes neurais convolucionais (CNNs) e é responsável por extrair características relevantes das entradas, sejam imagens, sinais ou outros tipos de dados estruturados. Diferentemente das redes neurais totalmente conectadas, em que cada neurônio está ligado a todos os neurônios da camada anterior, nas camadas convolutivas cada unidade processa uma região local da entrada, reduzindo drasticamente o número de parâmetros e permitindo o aprendizado de padrões espaciais.

A operação fundamental desta camada é a convolução discreta, definida da seguinte forma para uma imagem de entrada e um filtro ou kernel:

$$(I * K)(i,j) = \sum_m \sum_n I(i+m, j+n) \cdot K(m,n)$$

Onde:

- I é a imagem ou matriz de entrada (por exemplo, uma imagem em tons de cinza).
- K o filtro ou kernel convolucional, uma matriz pequena que será aplicada sobre III.
- (i,j) são as coordenadas do pixel de saída.
- m,n os índices que percorrem os elementos do kernel.
- $(I * K)(i,j)$ se refere ao valor do pixel na saída da convolução na posição

Nesta equação, denota a posição do pixel na imagem de saída, e os somatórios percorrem os elementos do kernel. O resultado é um mapa de ativação que destaca a presença de padrões aprendidos pelo filtro, como bordas, texturas e formas mais complexas nas camadas mais profundas (Lecun *et al*., 1998; Goodfellow *et al*., 2016).

A Figura 3 ilustra o processo de convolução em redes neurais, no qual o kernel é aplicado sobre a imagem de entrada para produzir um novo valor de pixel por meio da soma ponderada dos vizinhos (Nvidea, 2024).

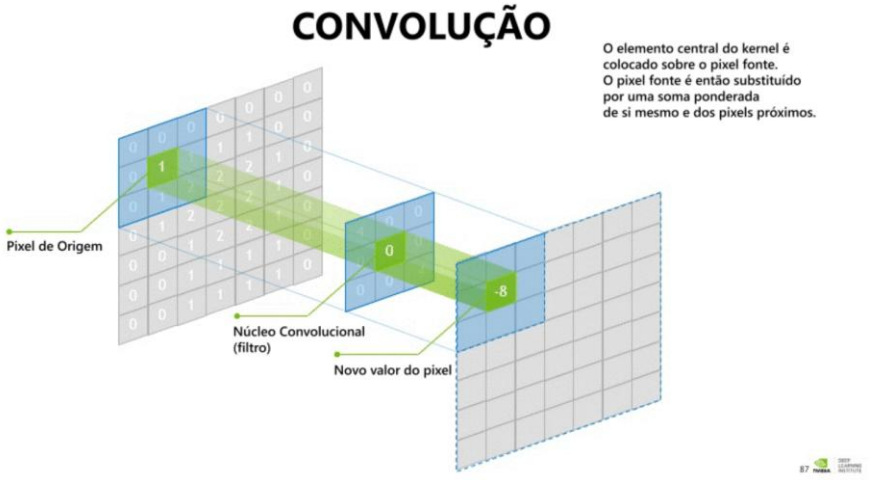


Figura 3: Exemplo do funcionamento da operação de convolução em uma rede neural convolucional (CNN), mostrando o alinhamento do kernel sobre a entrada e o cálculo do novo valor de pixel.

Fonte: adaptado de Nvidia (2024)

3.6.2 Camada de Subamostragem (*Pooling*)

A camada de pooling, ou subamostragem, tem como principal objetivo reduzir a dimensionalidade espacial dos mapas de ativação produzidos pelas camadas convolutivas. Isso contribui para a redução de parâmetros, melhora da generalização e maior robustez a variações na posição dos padrões detectados.

As operações mais comuns de pooling são:

- *Max Pooling*: seleciona o maior valor em cada região local.
- *Average Pooling*: calcula a média dos valores da região.

A operação de *max pooling* pode ser descrita como:

$$Y(i,j) = max_{(m,n)\in R(i,j)} X(m,n)$$

Onde:

- X é o mapa de ativação de entrada da camada de pooling.
- $Y(i,j)$ o valor de saída da operação de pooling na posição (i,j).
- $R(i,j)$ é a região (janela) de tamanho fixo (geralmente 2×2) da entrada X, associada à posição (i,j).
- (m,n) índices dos elementos dentro da janela.

A Figura 4 apresenta uma comparação visual entre as técnicas de *max pooling* e *average pooling*, aplicadas a uma matriz de entrada. No *max pooling*, o valor máximo de cada região é preservado, destacando os elementos mais relevantes do mapa de ativação. Já no *average pooling*, a média dos valores é computada, resultando em uma suavização das características extraídas. Ambas as abordagens são amplamente utilizadas para reduzir a dimensionalidade e aumentar a robustez do modelo a pequenas variações espaciais (Nehme, 2023).

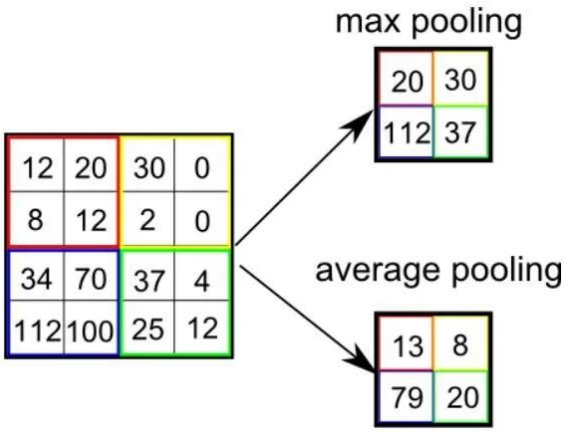


Figura 4: Comparação entre as operações de *max pooling* e *average pooling*, aplicadas sobre uma matriz 4×4 com janelas 2×2.
Fonte: Nehme (2023).

3.6.3 Camada de *Dropout*

A técnica de dropout foi proposta por Srivastava *et al*. (2014) como um método de regularização para evitar overfitting em redes neurais profundas. Essa camada atua de forma estocástica durante o treinamento, desativando aleatoriamente uma fração dos neurônios da camada anterior, evitando adaptação excessiva dos pesos.

A ativação de cada neurônio com *dropout* é dada por:

$$\tilde{h}_i = h_i \cdot z_i \; com \; z_i \sim Bernoulli(1-p)$$

Onde:

- h_i refere a saída do neurônio i antes da aplicação do dropout.
- $\tilde{h}_i$ é a saída do neurônio i após a aplicação do dropout.
- z_i é uma variável aleatória com distribuição de Bernoulli que assume valor 1 com probabilidade $1-p$ e 0 com probabilidade p.
- p é a taxa de dropout, ou seja, fração de neurônios desativados aleatoriamente durante o treinamento.

A Figura 5 ilustra a diferença entre uma rede neural padrão e a mesma rede com aplicação da técnica de dropout, conforme proposta por Srivastava *et al*. (2014). Observa-se que alguns neurônios são desativados aleatoriamente durante o treinamento, o que reduz o risco de overfitting e melhora a generalização do modelo.

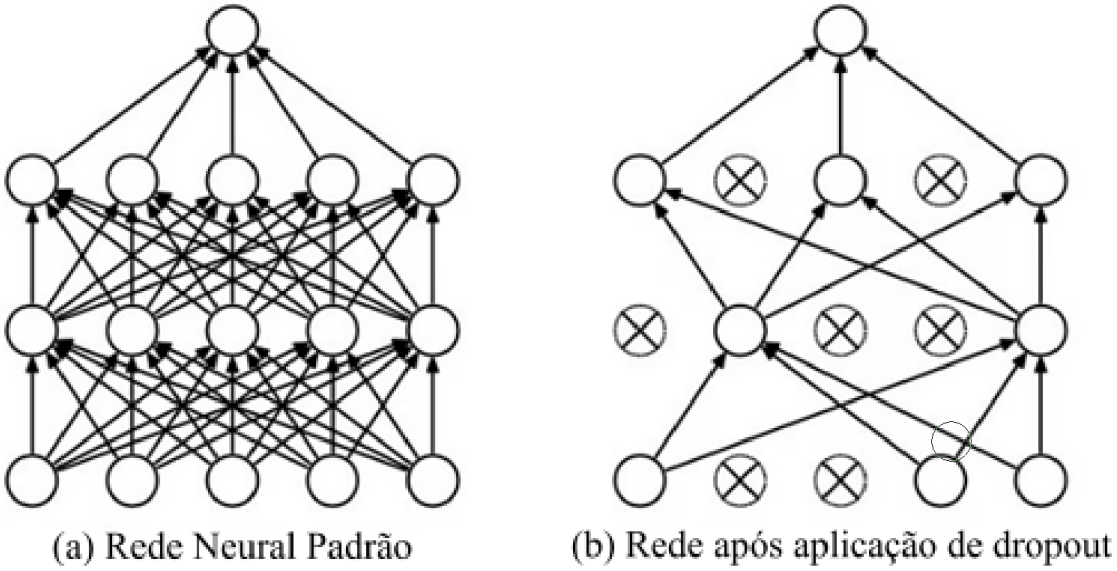


Figura 5: Comparação entre rede neural padrão(a) e rede com aplicação de dropout (b)
Fonte: adaptado de Srivastava *et al*. (2014).

3.6.4 Camada de unidades Retificadoras Lineares *(ReLU)*

As funções de ativação são componentes essenciais nas redes neurais, e a função ReLU (*Rectified Linear Unit*) tornou-se padrão de fato nas CNNs modernas. Ela introduz não linearidades nos modelos com baixo custo computacional.

A função ReLU é definida como:

$$f(x) = max(0, x)$$

Ela resolve o problema do gradiente desaparecendo que afetava funções como sigmoid ou tanh, e acelera a convergência do treinamento. ReLU também promove esparsidade na saída, o que pode melhorar a capacidade de generalização.

A Figura 6 apresenta as funções de ativação mais comuns utilizadas em redes neurais profundas: sigmoid, tanh, ReLU e *Leaky* ReLU.

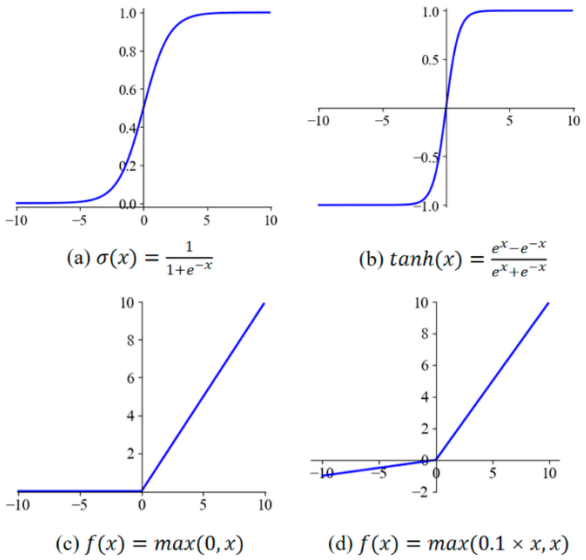


Figura 6: Representação gráfica das funções de ativação: (a) Sigmoid, (b) Tanh, (c) ReLU e (d) Leaky ReLU.

Fonte: Yang *et al*. (2023).

O uso de ReLU pode levar ao problema do "neurônio morto", quando valores negativos persistem em uma determinada unidade, que então nunca mais atualiza seus pesos. Para mitigar isso, variantes como *Leaky ReLU* e *Parametric ReLU* são utilizadas (Nair & Hinton, 2010).

3.6.5 Regularização L_2

A regularização L_2, também conhecida como *weight decay*, é uma técnica clássica de controle de complexidade do modelo, penalizando pesos excessivamente grandes. Ela é aplicada na função de perda, adicionando um termo proporcional ao quadrado da norma dos pesos:

$$J(\theta) = J_0(\theta) + \lambda \sum_i \theta_i^2$$

Onde:

- $J(\theta)$ é a função de custo regularizada.
- $J_0(\theta)$ a função de custo original (como *cross-entropy* ou MSE).
- θ_i é o parâmetro (peso) i-ésimo da rede neural.

- λ o hiper parâmetro de regularização L2, que controla a intensidade da penalização.

Esse termo adicional força os pesos a se manterem pequenos, promovendo modelos mais simples e menos propensos ao overfitting. Segundo Goodfellow *et al*. (2016), L_2 é especialmente eficaz quando combinada com outras técnicas como *dropout* e *data augmentation*.

3.7 Métodos de Otimização

A escolha do otimizador influencia diretamente a velocidade de convergência e a qualidade da solução encontrada por uma rede neural. Diversos algoritmos têm sido propostos com o objetivo de aprimorar o processo de atualização dos pesos, incorporando estratégias como momento, adaptação da taxa de aprendizado e regularização implícita. A seguir, são descritos três dos métodos de otimização mais utilizados em redes neurais profundas.

3.7.1 Estimativa Dinâmica Adaptativa (Adam)

O otimizador Adam (*Adaptive Moment Estimation*) combina as vantagens do RMSProp e do Gradiente Descendente com Momento. Ele mantém estimativas dos primeiros e segundos momentos do gradiente, permitindo atualizações adaptativas para cada parâmetro.

As atualizações de peso são realizadas conforme as equações:

1. Média móvel dos gradientes:

$$m_t = \beta_1 \cdot m_t - 1 + (1 - \beta_1) \cdot g_t$$

2. Média móvel dos quadrados dos gradientes:

$$v_t = \beta_2 \cdot v_{t-1} + (1 - \beta_2) \cdot g_t^2$$

3. Correções de viés:

$$\hat{m}_t = \frac{m_t}{1 - \beta_1^t} \; ; \; \hat{v}_t = \frac{v_t}{1 - \beta_2^t}$$

4. Atualização dos pesos:

$$\theta_{t+1} = \theta_t - \eta \cdot \frac{\hat{m}_t}{\sqrt{\hat{v}_t} + \epsilon}$$

Onde:

- g_t é o gradiente no passo;
- m_t é a média móvel do gradiente (momento);
- v_t é a média móvel dos quadrados do gradiente;
- β_1, β_2 são os coeficientes de decaimento (tipicamente 0.9 e 0.999);
- ϵ é um pequeno valor para evitar divisão por zero;
- η é a taxa de aprendizado.

O Adam é robusto e eficiente, sendo amplamente usado em problemas com grandes conjuntos de dados e arquiteturas profundas (Kingma & Ba, 2015).

3.7.2 Propagação da Raiz Média Quadrática (RMSProp)

RMSProp foi proposto por Tieleman & Hinton (2012) e é uma modificação do método *Adagrad*. Seu objetivo é resolver o problema da rápida diminuição da taxa de aprendizado do *Adagrad* ao acumular os quadrados dos gradientes em média móvel exponencial, sendo essas descritas pelas equações:

1. Média móvel dos quadrados dos gradientes:

$$v_t = \gamma \cdot v_{t-1} + (1 - \gamma) \cdot g_t^2$$

2. Atualização dos pesos:

$$\theta_{t+1} = \theta_t - \eta \cdot \frac{g_t}{\sqrt{v_t} + \epsilon}$$

Onde:

- γ é fator de decaimento (ex: 0.9);
- v_t é a média móvel dos quadrados dos gradientes;
- g_t é o gradiente atual.

Esse método é eficaz para problemas com dados não estacionários, como séries temporais e processamento de sinais (Tieleman & Hinton, 2012).

3.7.3 Gradiente Descendente Estocástico com Momento (SGDM)

O SGDM (*Stochastic Gradient Descent with Momentum*) adiciona um termo de "momentum" que suaviza as atualizações, acumulando gradientes passados para evitar oscilações excessivas, sendo descrita pelas equações:

1. Atualização do vetor de momento:

$$v_t = \mu \cdot v_{t-1} - \eta \cdot \nabla J(\theta_t)$$

2. Atualização dos pesos:

$$\theta_{t+1} = \theta_t + v_t$$

Onde:

- μ é o coeficiente de momento (ex: 0.9);
- η é a taxa de aprendizado;
- $\nabla J(\theta_t)$ é o gradiente da função de custo.

O uso do momento permite um avanço mais estável e rápido nas direções de menor curvatura da função de perda (Qian, 1999).

Esses métodos são frequentemente combinados com técnicas de normalização e regularização para alcançar um melhor desempenho em redes convolucionais profundas.

3.8 Métricas para avaliação

Esta seção apresenta as métricas utilizadas para avaliar o desempenho dos modelos de classificação, são descritas a seguir as principais métricas adotadas neste estudo: acurácia, F1-score, área sob a curva ROC (AUC) e a matriz de confusão para uma análise visual dos resultados.

Considerando:

- VP: número de verdadeiros positivos
- VN: número de verdadeiros negativos
- FP: número de falsos positivos
- FN: número de falsos negativos

Temos:

- Acurácia:

  $Acurácia = \frac{VP+VN}{VP+VN+FP+FN}$; Proporção de previsões corretas sobre o total de amostras avaliadas.

- F1-score:

  $F1 = \frac{2 \cdot VP}{2 \cdot VP+FP+FN}$; Média harmônica entre precisão e revocação, útil quando há desbalanceamento entre classes.

- Área sob a Curva ROC (AUC):

  $AUC = \int_{-\infty}^{\infty} Revocação(T) \cdot Especificidade'(T)dT$; Probabilidade de que o classificador atribua uma maior pontuação a uma instância positiva do que a uma negativa escolhida aleatoriamente (Fawcett, 2006).

- Matriz de Confusão:

  Representação tabular dos acertos e erros do modelo. Para problemas binários, organiza os valores de VP, VN, FP e FN, permitindo visualizar com clareza os tipos de erro e acerto.

Essas métricas, combinadas, fornecem uma avaliação abrangente do desempenho dos modelos, especialmente em contextos sensíveis como aplicações médicas ou de segurança, onde o custo de um erro pode ser elevado.

Capítulo 4

Materiais e Métodos

Esta seção descreve os conjuntos de dados utilizados, as técnicas de pré-processamento aplicadas, a arquitetura das redes neurais desenvolvidas, o ambiente computacional adotado e as estratégias de treinamento e avaliação empregadas.

4.1 Materiais

Este estudo utilizou dois conjuntos de dados clínicos distintos para a tarefa de predição da mortalidade em 14 dias em pacientes com traumatismo cranioencefálico (TCE), por meio da aplicação de modelos de aprendizado de máquina. O primeiro conjunto de dados foi obtido a partir de pacientes atendidos no Hospital das Clínicas da Universidade de São Paulo (HC-FMUSP), com período de coleta compreendido entre março de 2012 e janeiro de 2015, e acompanhamento finalizado em junho de 2015. O Comitê de Ética em Pesquisa da Universidade de São Paulo (São Paulo, Brasil) aprovou este estudo (CAAE 46831315.3.0000.0068). A base paulista contém um total de 517 registros válidos, com 15 variáveis preditoras organizadas em quatro categorias principais:

1. Demográficas: gênero (masculino ou feminino) e idade (em anos);
2. Clínicas: reatividade pupilar na admissão (bilateral reagente, uma ou duas pupilas fixas), escala de coma de Glasgow (GCS) no local do trauma (leve, moderada ou grave), GCS na admissão (idem), escore motor da GCS (1 a 6), presença de hipóxia (sim ou não), e hipotensão na admissão (sim ou não). Considera-se como hipotensão uma pressão arterial sistólica < 90 mmHg, e como hipóxia, saturação de oxigênio < 90%, conforme diretrizes da *Brain Trauma Foundation*;
3. Tomográficas: presença de desvio de linha média superior a 5 mm (sim ou não), hemorragia subaracnóidea (TSAH), hematoma epidural, hemorragia subdural e hemorragia intracerebral, todas com codificação binária;
4. Laboratoriais: tempo de protrombina (em segundos) e razão do tempo de tromboplastina parcial ativado (rAPTT).

O segundo conjunto de dados foi coletado em Manaus (Amazonas), entre maio de 2020 e julho de 2021, em um centro hospitalar terciário. Este estudo foi aprovado pelo Comitê de Ética em Pesquisa da Universidade Federal do Amazonas (UFAM) (CAAE: 25366619.1.0000.5020). A base manauara inclui 469 registros e, além das mesmas 15 variáveis utilizadas na base paulista, incorpora duas variáveis contextuais adicionais:

- Tempo entre o trauma e a admissão hospitalar, medido em horas;
- Indicador de coleta durante a pandemia da COVID-19, binário (0 = fora da pandemia; 1 = durante a pandemia).

A justificativa para a presença exclusiva dessas duas variáveis na base de Manaus está relacionada às peculiaridades logísticas da região Norte. Manaus é a única cidade do estado com capacidade de atendimento neurocirúrgico de emergência. Pacientes oriundos do interior geralmente são transportados por meios fluviais ou aéreos, resultando em um tempo médio de deslocamento de aproximadamente 67,1 horas até a chegada ao centro especializado (Nôvo *et al*., 2023). Esse cenário contrasta fortemente com o de São Paulo, que conta com ampla rede rodoviária e diversos centros especializados distribuídos em sua malha urbana e interiorana. A inclusão da variável indicativa da pandemia visa avaliar o impacto da sobrecarga hospitalar sobre os desfechos clínicos desses pacientes.

Ambos os conjuntos de dados foram armazenados em arquivos CSV estruturados e submetidos aos seguintes critérios de inclusão: (i) assinatura do termo de consentimento livre e esclarecido por parte do paciente ou responsável legal; (ii) presença de alterações na tomografia computadorizada de crânio; (iii) GCS $\leq$ 14 após estabilização na emergência; e (iv) idade superior a 14 anos. Os critérios de exclusão incluíram pacientes transferidos de outras unidades de terapia intensiva (UTI), com hematoma subdural crônico, ou com pupilas fixas bilaterais e GCS igual a três, sem resposta após manobras de ressuscitação cardiopulmonar.

A Tabela 2 apresenta o resumo das variáveis utilizadas nos dois bancos de dados, com seus respectivos tipos e faixas de valores.

Tabela 2: Variáveis utilizadas na predição de mortalidade em 14 dias

Classe	Variável	Tipo	Intervalo / Categoria
Demográfica	Gênero	Categórica	0 – 1
	Idade	Numérica	16 – 99
	Pandemia (exclusiva de Manaus)	Categórica	0 – 1
Clínica	Reatividade pupilar	Categórica	0 – 2
	GCS no local do trauma	Categórica	1 – 3
	GCS na admissão	Categórica	1 – 3
	Escore motor (GCS)	Categórica	1 – 6
	Hipóxia	Categórica	0 – 1
	Hipotensão na admissão	Categórica	0 – 1
	Tempo trauma-admissão (Manaus)	Numérica (horas)	0h – 12h
Tomográfica	Desvio de linha média (>5 mm)	Categórica	0 – 1
	Hemorragia subaracnoidea (CT)	Categórica	0 – 1
	Hematoma epidural (CT)	Categórica	0 – 1
	Hemorragia subdural (CT)	Categórica	0 – 1
	Hemorragia intracerebral (CT)	Categórica	0 – 1

Por fim, a Figura 7 mostra a proporção de pacientes que evoluíram a óbito em até 14 dias em cada base. Em São Paulo, a mortalidade foi de 22,82%, enquanto em Manaus, atingiu 27%. Um teste qui-quadrado realizado para avaliar a diferença entre as proporções revelou um valor de $x^2 = 2{,}38$, o qual não foi estatisticamente significativo ao nível de 5% ($x^2_{crítico} = 3{,}84, \text{gl} = 1$).

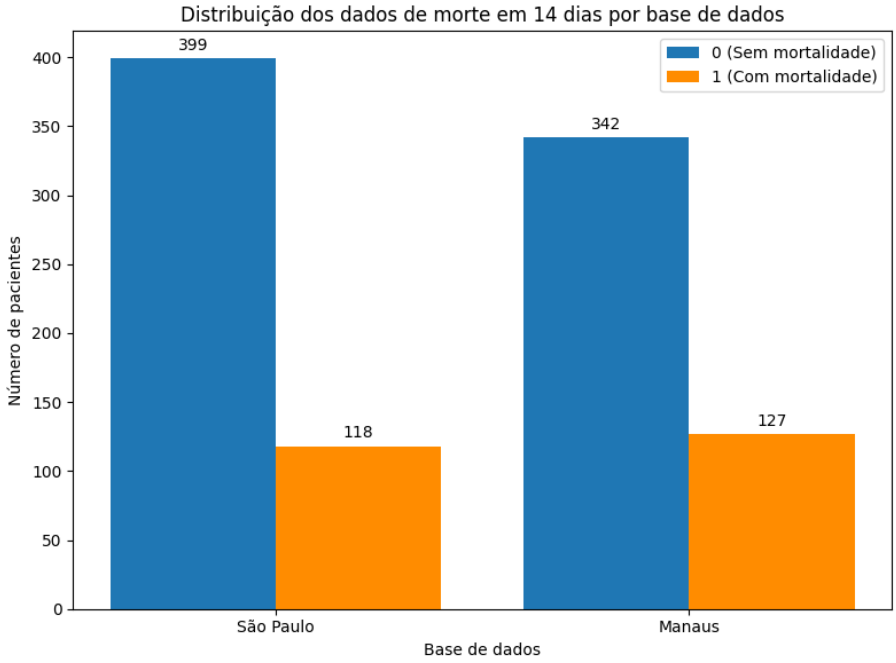


Figura 7: Distribuição da mortalidade em 14 dias por base de dados

4.2 Métodos

Essa sessão tem como foco a apresentação dos métodos utilizados essas sendo apresentadas no diagrama de blocos da Figura 8, passando desde os pré-processamentos feitos, toda a definição de ajustes de modelos para serem feitos os treinamentos até a parte final onde são obtidos os resultados

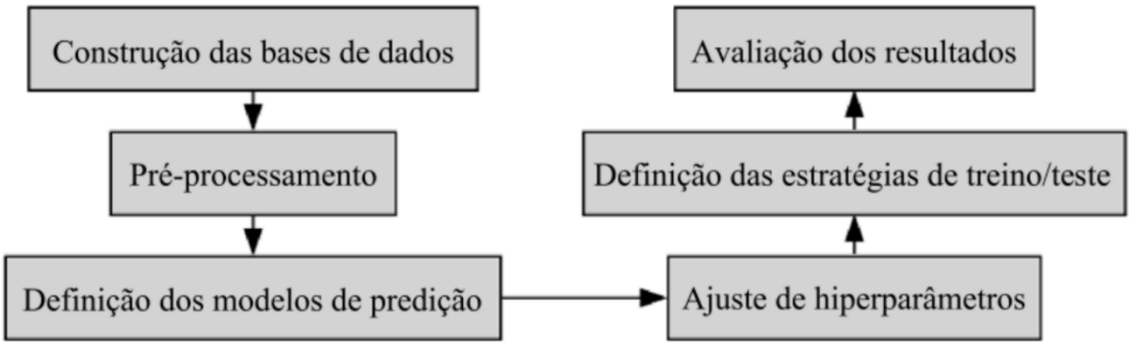


Figura 8: fluxograma da metodologia

4.2.1 Pré-Processamento

Essa fase buscou garantir a consistência, completude e escalabilidade dos dados, possibilitando que os modelos de aprendizado fossem treinados de forma eficaz, com menor risco de viés ou overfitting decorrente de ruído ou dados inconsistentes.

A base de dados de São Paulo apresentou cerca de 18% de amostras com ao menos uma variável ausente. Variáveis como hipóxia e GCS pré-hospitalar apresentaram maior proporção de valores faltantes, exigindo estratégias diferenciadas para tratamento. Em contraste, a base de Manaus mostrou-se mais completa, com aproximadamente 2% das amostras contendo dados incompletos.

O pré-processamento dos dados de São Paulo foi descrito previamente por Guimarães *et al*. (2022), e as mesmas diretrizes foram aplicadas na base de Manaus para assegurar uniformidade no tratamento dos dados. O preenchimento de valores ausentes foi conduzido conforme o tipo da variável:

- Para variáveis categóricas, foram utilizadas abordagens baseadas em algoritmos de aprendizado supervisionado como árvore de decisão, floresta aleatória (Floresta randômica) e k-vizinhos mais próximos (k-NN);
- Para variáveis numéricas, foram aplicados métodos de imputação por regressão linear, além de uso de modelos baseados em árvore e Floresta randômica;
- Variáveis com porcentagem mínima de ausência foram imputadas por medidas estatísticas simples, como a média ou a moda.

Após o preenchimento dos valores ausentes, os dados passaram por uma etapa de normalização, essencial para modelos sensíveis à escala, como redes neurais. A normalização adotou múltiplas técnicas conforme o perfil da variável:

- *Min-Max Scaling*: para compressão de valores entre 0 e 1;
- *Z-score normalization*: para centralização e padronização de variáveis contínuas;
- Transformação cúbica: aplicada em variáveis com distribuição assimétrica severa.

Além disso, todas as variáveis categóricas foram transformadas por codificação *one-hot* (vetores com apenas uma coordenada igual a 1 e as outras, iguais a 0), exceto nos casos em que a arquitetura do modelo aceitava diretamente entradas categóricas

indexadas. Esse tratamento resultou em um vetor de atributos numéricos compatível com as arquiteturas convolucionais adotadas neste trabalho.

Por fim, os dados foram estratificados e divididos em subconjuntos de treinamento (80%) e teste (20%), mantendo a proporção original de pacientes sobreviventes e não sobreviventes em 14 dias. Essa divisão estratificada foi crucial para evitar distorções de distribuição de classes durante o treinamento e avaliação dos modelos.

4.2.2 Definição dos modelos de predição

A definição dos modelos de predição empregados neste trabalho foi guiada por dois objetivos principais: (i) avaliar a capacidade discriminativa de algoritmos clássicos de aprendizado supervisionado, frequentemente utilizados em contextos médicos; e (ii) explorar o potencial das redes neurais convolucionais (CNNs), originalmente projetadas para tarefas em domínio de imagens, na modelagem de dados clínicos estruturados.

Inicialmente, foram empregados três modelos de referência que representam diferentes paradigmas de modelagem:

- Regressão Logística (RL): modelo linear amplamente consolidado em aplicações clínicas devido à sua interpretabilidade e boa robustez estatística. É utilizado como baseline em diversos estudos relacionados à predição de desfechos em TCE (Raj *et al*., 2013).
- Floresta randômica (RF): algoritmo baseado em múltiplas árvores de decisão agregadas por voto majoritário. Tem sido eficaz em problemas com variáveis mistas e ausência de linearidade (Breiman, 2001).
- Perceptron Multicamadas (MLP): rede neural densa com múltiplas camadas ocultas e funções de ativação não lineares, aplicada como transição entre modelos estatísticos e redes convolucionais profundas (Lecun *et al*., 2015). Neste estudo a rede MLP foi projetada é apresentada na Figura 9 onde sua configuração consiste em uma primeira camada oculta com 128 neurônios e a segunda com 64 neurônios, ambas utilizando a função de ativação ReLU com camadas de *dropout* com taxa de 0,2 após cada camada oculta. A camada final, responsável pela classificação binária, contém um único neurônio com ativação sigmoide.

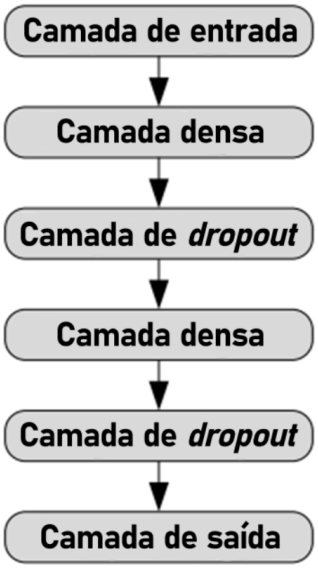


Figura 9: Arquitetura da rede MLP utilizada para predição de mortalidade em 14 dias para paciente com TBI

As redes convolucionais têm se destacado não apenas em tarefas visuais, mas também em problemas envolvendo dados tabulares, ao se adaptarem para capturar padrões espaciais ou ordenamentos implícitos. Diversos trabalhos recentes apontam para a aplicabilidade de CNNs em contextos médicos com alto grau de dimensionalidade e correlação entre atributos (Krizhevsky *et al*., 2012; Shickel *et al*., 2018).

Neste estudo, foram desenvolvidas e comparadas duas arquiteturas distintas:

- CNN1 – Arquitetura paralela: inspirada na estrutura Inception (Szego *et al*., 2015), esta rede utiliza múltiplos filtros convolucionais 1D de tamanhos variados (2, 3, 4) aplicados em paralelo à entrada. O objetivo é permitir a captura de relações locais de diferentes escalas entre os atributos clínicos. Os mapas de ativação resultantes são concatenados e enviados a uma camada densa com 50 neurônios (ativação ReLU), seguida de camada dropout (0,2) e saída com ativação sigmoide.
- CNN2 – Arquitetura sequencial profunda: utiliza uma sequência de blocos convolucionais compostos por convolução 1D, normalização por lote (*batch normalization*) e ativação ReLU. Após dois blocos consecutivos, os dados passam por uma camada densa com 50 neurônios, *dropout* (0,2) e camada de saída sigmoide. Essa abordagem favorece o aprendizado hierárquico de representações latentes.

Uso de duas arquiteturas CNNs diferentes foi aplicado para avaliar métodos de extração de características diferentes devido a variação do formato da CNN, esse formato é descrito na Figura 10 onde é visto a diferença na robustez da extração de características ao utilizar camadas em paralelas para essa atividade.

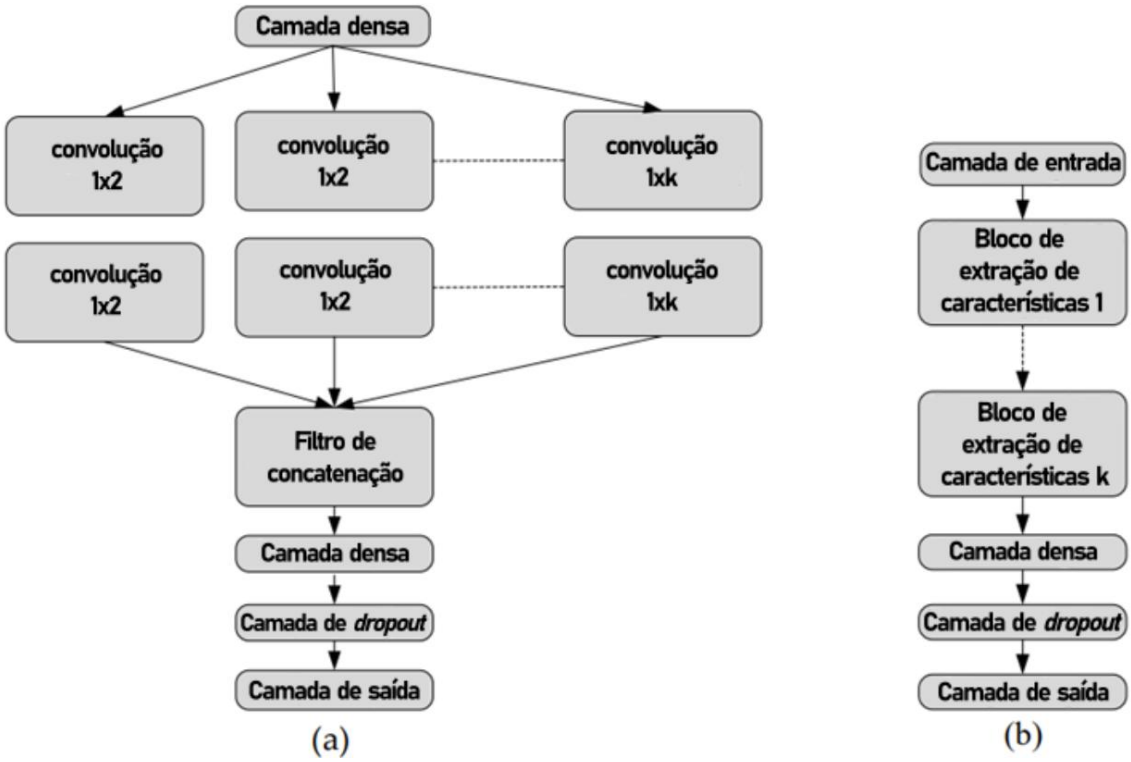


Figura 10: Arquitetura das redes CNN utilizadas para predição de mortalidade em 14 dias para paciente de TBI. (a) CNN com arquitetura em paralelo; (b) CNN com arquitetura em série

A escolha por CNNs é corroborada por estudos como o de Razzak *et al*. (2019), que destacam a capacidade dessas redes em superar modelos tradicionais em tarefas biomédicas, especialmente quando combinadas com estratégias de regularização e ajuste apropriado de hiper parâmetros. Além disso, as CNNs mantêm compatibilidade com métodos de interpretabilidade, como SHAP (Lundberg e Lee, 2017), sendo esse fundamental para entender o comportamento da base de dados com os melhores modelos treinados.

4.2.3 Ajuste de hiper parâmetros

O ajuste de hiperparâmetros é uma etapa essencial para garantir o bom desempenho e a generalização dos modelos de aprendizado de máquina. Essa fase

envolve a escolha criteriosa de parâmetros que não são aprendidos diretamente durante o treinamento, mas que influenciam significativamente o comportamento do modelo, como taxa de aprendizado, número de épocas, tamanho dos lotes, otimizadores e *callbacks*.

O treinamento dos modelos MLP e CNNs foi configurado com um conjunto de hiperparâmetros definidos inicialmente com base na literatura e posteriormente refinados por meio de experimentação empírica. A taxa de aprendizado utilizada foi de 1e-2 para a rede MLP e de 1e-3 para as CNNs. Durante o processo de treinamento, essa taxa foi reduzida automaticamente ao se observar estagnação em mínimos locais, até atingir um valor mínimo de 1e-6, estratégia que contribuiu para estabilizar a convergência. O número de épocas foi fixado em 300, com utilização de *callbacks* para armazenar o modelo com melhor desempenho com base na acurácia obtida no conjunto de validação.

Foram avaliados diferentes otimizadores nos modelos com redes neurais, entre eles:

- Adam (*Adaptive Moment Estimation*);
- RMSProp (*Root Mean Square Propagation*);
- SGDM (*Stochastic Gradient Descent with Momentum*).

A função de perda adotada foi a entropia cruzada binária, apropriada para tarefas de classificação binária. Os melhores conjuntos de hiperparâmetros foram selecionados com base nas métricas obtidas nos subconjuntos de validação, priorizando F1-score e AUC. Esse processo buscou um equilíbrio entre acurácia e sensibilidade, promovendo a robustez dos modelos diante de diferentes distribuições de entrada.

4.2.4 Estratégias de Treinamento e Teste

Cinco estratégias distintas foram utilizadas para avaliar os modelos de aprendizado de máquina na predição da mortalidade até 14 dias de pacientes com TCE. Essas estratégias tiveram dois objetivos principais: avaliar o desempenho dos modelos e avaliar a capacidade de generalização dos mesmos. As estratégias foram desenhadas de modo a utilizar diferentes combinações entre as bases de dados de São Paulo e Manaus, além de explorar o impacto de variáveis contextuais exclusivas da base de Manaus.

- Estratégia 1: O modelo é treinado e testado utilizando apenas os dados da base de São Paulo. O objetivo é avaliar o desempenho do modelo dentro de um único contexto urbano, com 15 variáveis comuns entre os conjuntos.

- Estratégia 2: O modelo é treinado e testado utilizando exclusivamente os dados da base de Manaus. Essa estratégia permite avaliar o desempenho do modelo em um contexto clínico e logístico diferente, inicialmente com 15 variáveis e, posteriormente, com a adição de uma ou duas variáveis contextuais exclusivas da base de Manaus (variável de pandemia e tempo entre o trauma e a admissão hospitalar), totalizando até 17 variáveis.
- Estratégia 3: O modelo é treinado com os dados da base de São Paulo e testado com os dados da base de Manaus. Esta abordagem permite analisar o grau de generalização dos modelos quando expostos a um ambiente clínico distinto daquele em que foram treinados.
- Estratégia 4: O modelo é treinado com os dados da base de Manaus e testado com os dados da base de São Paulo. Esta estratégia complementa a anterior, também avaliando a generalização, mas em direção oposta.
- Estratégia 5: Os dados de ambas as bases são unificados, e o modelo é treinado e testado sobre este conjunto combinado. Esta configuração busca verificar se um modelo pode capturar características comuns entre os dois contextos e ainda assim manter um bom desempenho.

A Figura 11 ilustra graficamente a organização e distribuição das cinco estratégias aplicadas no estudo.

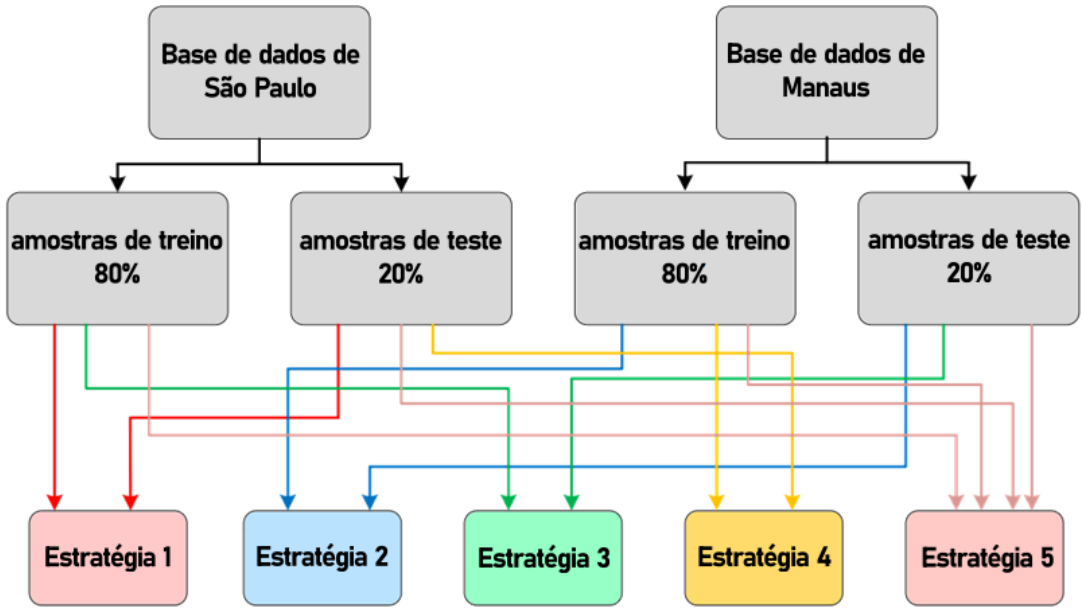


Figura 11: Fluxograma das estratégias de treinamento e teste adotadas neste trabalho.

Capítulo 5

Resultados e Discussões

Esta seção apresenta os resultados obtidos a partir da aplicação das cinco estratégias de treinamento e teste descritas na Seção 4.2.4, com os diferentes modelos de aprendizado de máquina avaliados. Os modelos foram comparados com base nas métricas definidas previamente, acurácia, F1-score e AUC, permitindo uma análise detalhada da capacidade preditiva de cada abordagem.

Os resultados são organizados por estratégia e discutidos em termos de desempenho relativo entre os modelos mais clássicos (Regressão Logística, Floresta randômica e MLP) e as redes convolucionais (CNN1 e CNN2). A análise enfatiza também a capacidade de generalização dos modelos, sobretudo nas estratégias que envolvem testes cruzados entre as bases de São Paulo e Manaus.

5.1 Resultados para a estratégia 1 e 2 com 15 variáveis preditivas na entrada

As métricas de desempenho acurácia, F1-score e AUC obtidas nas Estratégias 1 (base São Paulo) e 2 (base Manaus), utilizando as 15 variáveis preditivas comuns entre as bases, esses resultados são apresentados na Tabela 3. A análise dos resultados revela que os modelos baseados em redes neurais convolucionais foram significativamente superiores aos modelos tradicionais em ambas as bases.

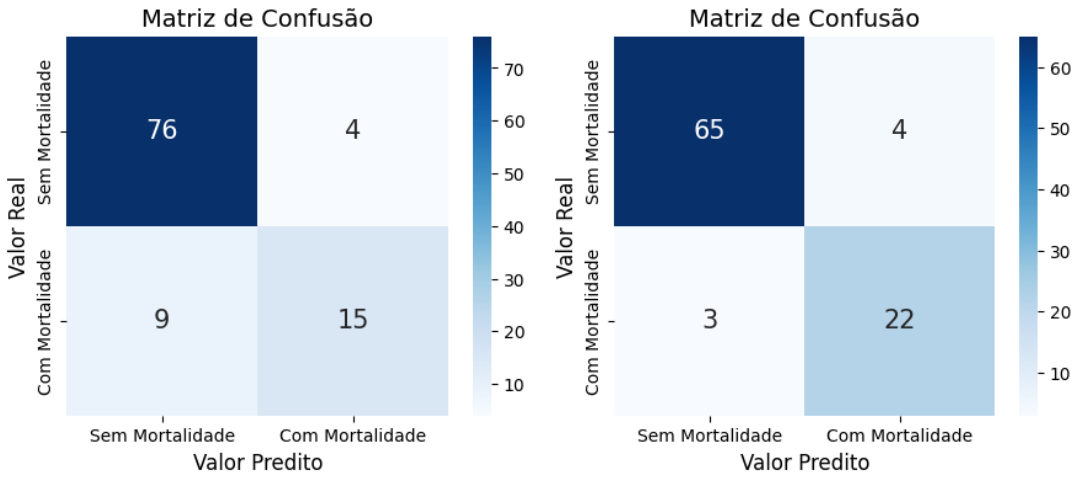
Os modelos CNN2 e CNN1, ambos utilizando o otimizador RMSProp, destacaram-se como os melhores em cada base. Em particular, a CNN2 com RMSProp atingiu acurácia de 0,87, F1-score de 0,85 e AUC de 0,90 na Estratégia 1. Já na Estratégia 2, a CNN1 com RMSProp apresentou acurácia de 0,90, F1-score de 0,89 e AUC de 0,93, o maior valor observado entre todos os experimentos realizados. Tais resultados indicam não apenas a superioridade das CNNs para a tarefa de predição de mortalidade de pacientes com TCE, mas também a efetividade do otimizador RMSProp para este tipo de tarefa, especialmente quando aplicado a dados clínicos heterogêneos.

Ao comparar diretamente os resultados entre as duas estratégias, nota-se que os modelos obtiveram desempenho superior na base de Manaus (Estratégia 2) em relação à base de São Paulo (Estratégia 1). Esse comportamento pode ser atribuído a uma maior uniformidade e completude dos dados da base manauara, conforme discutido anteriormente na Seção 4.2.1. Adicionalmente, observa-se que os modelos mais simples, como a regressão logística, apresentaram desempenho inferior, provavelmente por sua limitação em capturar relações não lineares entre as variáveis clínicas.

As matrizes de confusão associadas aos melhores modelos de cada estratégia são apresentadas na Figura 12. Na Estratégia 1, o modelo CNN2 + RMSProp obteve uma sensibilidade (revocação) de 0,79 e especificidade de 0,89. Já na Estratégia 2, o modelo CNN1 + RMSProp alcançou sensibilidade de 0,84 e especificidade de 0,96, evidenciando sua elevada capacidade discriminativa, tanto para prever corretamente os óbitos quanto para minimizar falsos positivos.

Tabela 3: Métricas obtidas para as estratégias 1 e 2 com 15 variáveis de entrada

Modelo de machine learning	Otimizador	Estratégia 1			Estratégia 2		
		Acurácia	F1 - Score	AUC	Acurácia	F1 - Score	AUC
Regressão Logística	-	0,81	0,79	0,83	0,81	0,77	0,79
Random Forest	-	0,81	0,80	0,83	0,82	0,79	0,81
MLP	Adam	0,79	0,76	0,80	0,89	0,88	0,90
MLP	RMSprop	0,80	0,78	0,81	0,88	0,86	0,89
MLP	SGDM	0,80	0,78	0,82	0,87	0,85	0,89
CNN1	Adam	0,81	0,80	0,83	0,90	0,88	0,91
CNN1	RMSprop	0,82	0,81	0,84	0,92	0,90	0,93
CNN1	SGDM	0,80	0,78	0,81	0,88	0,86	0,89
CNN2	Adam	0,86	0,83	0,89	0,89	0,87	0,90
CNN2	RMSprop	0,87	0,85	0,90	0,90	0,89	0,91
CNN2	SGDM	0,85	0,82	0,87	0,87	0,85	0,89



(a) (b)

Figura 12: Matrizes de confusão para ambas as estratégias: (a) Estratégia 1 com a CNN2 e o otimizador RMSProp; (b) Estratégia 2 com a CNN1 e o otimizador RMSProp.

5.2 Resultados para a estratégia 2 com 15, 16, 17 variáveis preditoras

Nesta subseção, são apresentados os resultados obtidos para a Estratégia 2, na qual os modelos foram treinados e testados exclusivamente com a base de dados de Manaus. Inicialmente, considerou-se o resultado obtido com o conjunto de 15 variáveis preditores comuns às duas bases (São Paulo e Manaus). Em seguida, avaliou-se o impacto da adição das duas variáveis contextuais exclusivas de Manaus pandemia e tempo entre o trauma e a admissão hospitalar resultando em dois cenários com 16 e um com 17 variáveis de entrada.

A Tabela 4 apresenta as métricas de desempenho (acurácia, F1-score e AUC) para cada cenário. Observa-se que, com as 15 variáveis iniciais, o melhor desempenho foi obtido pelo modelo “CNN1” com o otimizador RMSProp apresentado na Tabela 3. Ao adicionar as variáveis separadamente, houve incremento significativo na AUC para 0,97 em ambos os casos, sugerindo que o uso de variáveis contextuais influencia fortemente os desfechos de mortalidade em Manaus. Subsequente a estes testes é feito um treinamento com todas as 17 variáveis disponíveis, e de forma similar o ganho da inserção de variáveis contextuais mantiveram o ganho, consolidando um AUC de 0,98 no cenário com 17 variáveis, o melhor resultado entre todos os experimentos deste estudo.

A Figura 13 ilustra a matriz de confusão referente ao modelo com 17 variáveis. Nota-se sensibilidade de 0,92 e especificidade de 0,99, evidenciando a elevada capacidade do modelo em identificar tanto óbitos quanto sobreviventes de forma equilibrada.

Esses resultados indicam que a inclusão de variáveis contextuais específicas de Manaus trouxe ganho expressivo na capacidade preditiva. Isso se deve, possivelmente, à forte influência das condições logísticas regionais e do impacto da pandemia sobre os fluxos hospitalares e o atendimento emergencial, aspectos já documentados em estudos anteriores sobre a região Norte do Brasil.

Tabela 4: Métricas obtidas para a estratégia 2 com 15, 16 e 17 variáveis de entrada com o modelo CNN1 e otimizador RMSProp.

Variáveis	Acurácia	F1-Score	AUC
15 variáveis	0,92	0,90	0,93
15 variáveis + pandemia	0,95	0,94	0,97
15 variáveis + tempo trauma-admissão	0,95	0,96	0,97
15 variáveis + pandemia e tempo trauma-admissão	0,97	0,96	0,98

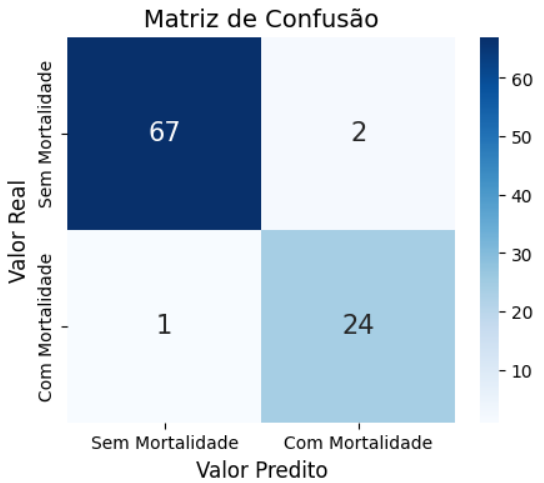


Figura 13: Matriz de confusão para a Estratégia 2 com 17 variáveis preditoras usando o modelo CNN1 e o otimizador RMSProp.

5.3 Resultados para a estratégia 3 e 4

As Estratégias 3 e 4 têm como objetivo avaliar a capacidade de generalização cruzada dos modelos. Na Estratégia 3, o treinamento é realizado com a base de São Paulo e o teste com a base de Manaus; na Estratégia 4, ocorre o inverso.

A Tabela 5 apresenta os resultados comparativos para ambas as estratégias, utilizando as 15 variáveis preditoras comuns. Nota-se que, em ambos os cenários, houve redução significativa do desempenho quando comparados aos resultados obtidos nas estratégias 1 e 2 (treinamento e teste na mesma base). O melhor AUC na Estratégia 3 foi de 0,53 (CNN1 com RMSProp), enquanto na Estratégia 4 o melhor AUC foi de 0,70 para o mesmo modelo. Esses valores contrastam com os AUCs de 0,90 e 0,93 obtidos nas estratégias sem foco em generalização.

Uma análise mais aprofundada revela que o treinamento em Manaus (Estratégia 4) generalizou melhor para São Paulo do que o inverso. Esse comportamento pode ser explicado pela maior variabilidade intrínseca da base de Manaus, que inclui variáveis contextuais e cenários logísticos mais extremos (como longos tempos de transferência e alta taxa de hipóxia/hipotensão), fornecendo ao modelo uma gama mais ampla de padrões clínicos. Em contrapartida, a base de São Paulo, mais homogênea e coletada em um período anterior à pandemia, apresentou menor representatividade de condições críticas encontradas na Amazônia.

Tabela 5: Métricas obtidas para as estratégias 3 e 4 com 15 variáveis de entrada

Modelo de machine learning	Otimizador	Estratégia 3			Estratégia 4		
		Acurácia	F1 - Score	AUC	Acurácia	F1 - Score	AUC
Regressão Logística	-	0,30	0,49	0,51	0,60	0,45	0,61
Random Forest	-	0,33	0,51	0,50	0,64	0,47	0,67
MLP	Adam	0,36	0,52	0,50	0,71	0,50	0,68
MLP	RMSprop	0,36	0,52	0,51	0,70	0,49	0,68
MLP	SGDM	0,35	0,51	0,53	0,69	0,48	0,67
CNN1	Adam	0,38	0,53	0,52	0,73	0,50	0,69
CNN1	RMSprop	0,27	0,42	0,52	0,77	0,52	0,70
CNN1	SGDM	0,35	0,51	0,51	0,71	0,49	0,68
CNN2	Adam	0,37	0,53	0,52	0,75	0,51	0,70
CNN2	RMSprop	0,36	0,52	0,52	0,74	0,51	0,69
CNN2	SGDM	0,35	0,51	0,50	0,73	0,50	0,69

5.4 Resultados para a estratégia 5

A Estratégia 5 consistiu na unificação das bases de São Paulo e Manaus em um único conjunto. Essa abordagem buscou verificar se um modelo treinado em dados combinados poderia capturar características comuns a ambas as regiões, mantendo desempenho satisfatório em um cenário misto.

Os resultados na Tabela 6 indicam desempenho intermediário: o melhor modelo (CNN1 com RMSProp) alcançou AUC de 0,77, superior ao obtido nas estratégias cruzadas (3 e 4), mas ainda inferior aos valores observados nas estratégias isoladas (1 e 2).

A Figura 14 demonstra distribuição desequilibrada de acertos e erros, com valor alto de falsos negativos. A análise sugere que, embora a fusão das bases forneça maior volume de dados para treinamento, as diferenças estruturais e contextuais entre as regiões ainda impactam o desempenho, reforçando a importância de variáveis regionais para a modelagem preditiva.

Tabela 6: Métricas obtidas para a estratégia 5 com 15 variáveis de entrada

Modelo de machine learning	Otimizador	Acurácia	F1-Score	AUC
Regressão Logística	-	0,90	0,61	0,72
Random Forest	-	0,80	0,57	0,70
MLP	Adam	0,81	0,62	0,72
MLP	RMSprop	0,81	0,62	0,73
MLP	SGDM	0,71	0,12	0,53
CNN1	Adam	0,81	0,63	0,73
CNN1	RMSprop	0,83	0,69	0,77
CNN1	SGDM	0,70	0,09	0,52
CNN2	Adam	0,81	0,62	0,72
CNN2	RMSprop	0,80	0,61	0,72

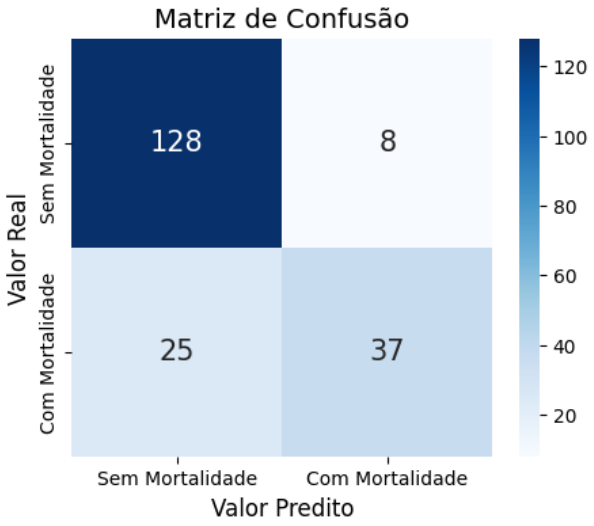


Figura 14: Matriz de confusão para a Estratégia 5 com 15 variáveis preditoras usando o modelo CNN1 e o otimizador RMSProp.

5.5 Explicação dos resultados

A análise dos resultados obtidos nas cinco estratégias permite compreender o impacto das variáveis preditoras, das características regionais e da complexidade dos modelos utilizados.

Em primeiro lugar, observa-se que as redes neurais convolucionais (CNN1 e CNN2) superaram consistentemente os modelos clássicos (Regressão Logística e Floresta Randômica), evidenciando a capacidade das CNNs em capturar padrões não lineares e interações complexas entre variáveis clínicas e tomográficas. Esse resultado está em consonância com a literatura recente, que aponta vantagens do aprendizado profundo em problemas biomédicos com múltiplos preditores heterogêneos.

Outro ponto relevante é a superioridade da base de Manaus quando enriquecida com variáveis contextuais exclusivas. A adição das variáveis pandemia e tempo trauma-admissão aumentou substancialmente o AUC, chegando a 0,98. Isso sugere que modelos localmente adaptados são mais eficazes em regiões com desafios logísticos e epidemiológicos específicos.

Em contrapartida, as estratégias de validação cruzada entre bases (3 e 4) apresentaram queda acentuada no desempenho, refletindo a baixa generalização inter-

regional. Tais achados reforçam a existência de características regionais para as mesmas variáveis existentes em ambas as bases de dados.

Para aprofundar a interpretação, foram aplicadas análises complementares com coeficiente de Pearson e valores *SHAP*, detalhadas a seguir.

5.5.1 Análise por coeficiente de Pearson

O coeficiente de Pearson foi utilizado para quantificar a correlação linear entre cada variável preditora e a mortalidade em 14 dias. As Tabelas 7 e 8 apresentam os resultados para as bases de São Paulo e Manaus, respectivamente.

Na base de São Paulo, a variável reatividade pupilar apresentou a maior correlação absoluta com o desfecho (r = -0,373), seguida pela pontuação motora (r = -0,281) e desvio de linha média (r = 0,219). Em Manaus, as correlações foram mais intensas: pontuação motora (r = -0,654), reatividade pupilar (r = -0,588), hipóxia (r = 0,458), hipotensão (r = 0,375) e desvio de linha média (r = 0,402).

A comparação entre as bases evidencia que, em Manaus, há preditores clínicos mais fortemente associados ao preditor final, justificando seu desempenho elevado em relação a base de São Paulo. Onde é possível encontrar somente uma variável com coeficiente superior a 3, enquanto ao trabalharmos com a Base de Manaus é possível observar 6 variáveis com esse grau alto de correlação com a classificação.

Tabela 7: Coeficientes de correlação de Pearson para a Base de São Paulo

Variável Preditiva	Coeficiente de Pearson
Gênero	-0,122
Idade	0,190
Reatividade pupilar	-0,373
GCS no local do trauma	0,119
GCS na admissão	0,121
Pontuação motora	-0,281
Hipóxia	0,107
Hipotensão	0,140
Desvio da linha média	0,219
Hemorragia subaracnóidea	0,059
Hematoma epidural	0,080
Hematoma subdural	-0,44
Hemorragia intracerebral	0,051

Tempo de protrombina	0,165
Tempo de tromboplastina parcial	0,159

Tabela 8: Coeficientes de correlação de Pearson para a Base de Manaus

Variável Preditiva	Coeficiente de Pearson
Sexo	-0,043
Idade	-0,088
Reatividade pupilar	-0,588
GCS no local do trauma	0,268
GCS na admissão	0,580
Pontuação motora	-0,654
Hipóxia	0,458
Hipotensão	0,375
Desvio da linha média	0,402
Hemorragia subaracnóidea	0,050
Hematoma epidural	-0,094
Hemorragia subdural	0,263
Hemorragia intracerebral	-0,024
Tempo de protrombina	-0,271
Tempo de tromboplastina parcial	-0,064
Pandemia	0,107
Tempo trauma admissão	0,026

5.5.2 Análise por valores de *SHAP*

Para avaliar a importância relativa e o impacto direcional de cada variável nas predições, foi utilizada a técnica *SHAP* (SHapley Additive exPlanations). Os gráficos de dispersão na Figura 15 ilustram a relevância de cada variável em cada base de dados de forma entender como elas afetam a predição final.

Para a base de São Paulo, os preditores de maior contribuição foram pontuação motora, reatividade pupilar, desvio de linha média, GCS na admissão e GCS no local do trauma. Já em Manaus, destacaram-se Pontuação motora, reatividade pupilar, hipóxia, desvio de linha média e hipotensão.

Interessantemente, a análise SHAP corroborou em grande parte os achados da correlação de Pearson, como hipóxia e hipotensão tiveram impacto preditivo mais expressivo em Manaus, ainda que suas correlações lineares não fossem as mais altas em São Paulo.

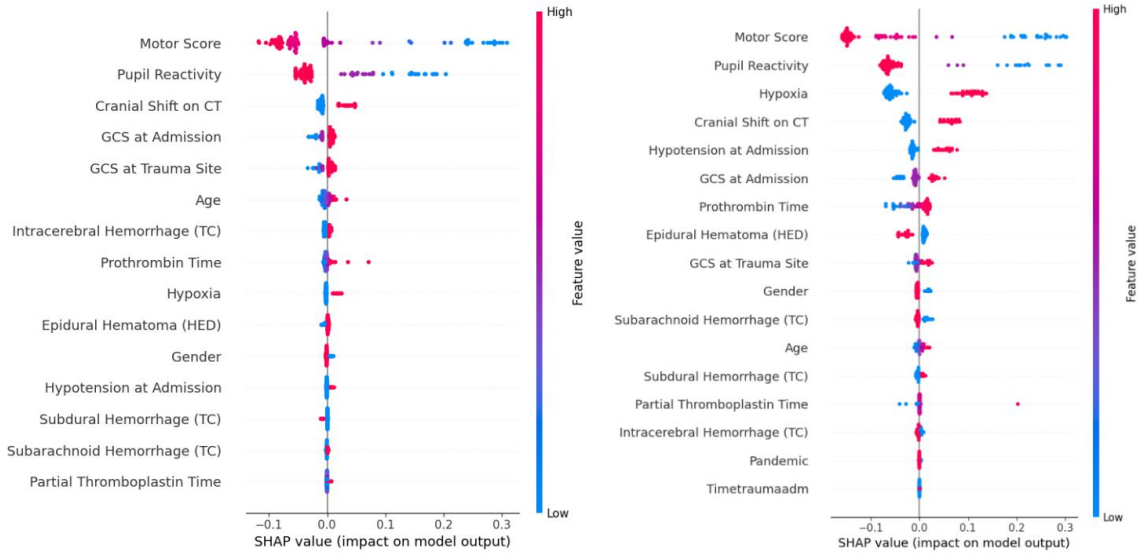


Figura 15: Valores de *SHAP* para previsões do modelo CNN1 com otimizador RMSprop. (a) Conjunto de dados de São Paulo, com 15 variáveis de entrada. (b) Conjunto de dados de Manaus com 17 variáveis de entrada.

Com a análise de *SHAP* é possível observar os preditores mais importantes na predição da mortalidade em 15 dias então de forma a provar essa correlação linear apresentada nos gráficos de *SHAP* foram feitos treinamentos somente com os 5 preditores mais importantes de cada base, de forma a comprovar a eficácia da analise por esses métodos sendo esses apresentados na Tabela 9.

Tabela 9: Métricas de desempenho para conjuntos de dados de São Paulo e Manaus, com seus respectivos melhores preditores.

Base de dados/ Variáveis	Acurácia	F1-Score	AUC
São Paulo / pontuação motora, reatividade pupilar, desvio da linha média, GCS na admissão e GCS no local do trauma	0,86	0,61	0,87
Manaus/ pontuação motora, reatividade pupilar, hipóxia, desvio da linha média e hipotensão na admissão	0,92	0,83	0,97

Esses achados sugerem que um conjunto reduzido de variáveis-chave concentra grande parte da capacidade discriminativa do modelo, o que, do ponto de vista clínico-operacional, representa uma vantagem significativa. Modelos com menos entradas demandam menor tempo de coleta de dados, apresentam menor incidência de valores ausentes e permitem processamento mais rápido, fatores especialmente críticos em cenários de emergência e em regiões com recursos limitados.

5.6 Discussão

Os resultados obtidos ao longo das estratégias de treinamento e teste permitem uma leitura integrada sobre três dimensões centrais deste estudo: (i) o desempenho absoluto dos modelos em cada base; (ii) a capacidade de generalização inter-regional; e (iii) a interpretabilidade e qualidade dos preditores mais relevantes.

Em primeiro lugar, a superioridade consistente das Redes Neurais Convolucionais (CNNs) sobre os modelos clássicos, quando treinadas e avaliadas na mesma base, reforça a hipótese de que as relações não lineares e as interações entre variáveis clínicas, radiológicas e contextuais são determinantes para a captura do risco de mortalidade em 14 dias. O desempenho das CNNs foi particularmente notável nas Estratégias 1 e 2. Em São Paulo, a melhor CNN obteve uma AUC (Área Sob a Curva) superior a 0,90, enquanto em Manaus, o desempenho foi ainda mais elevado, com sensibilidade e especificidade robustas. Isso sugere que, em contextos mais complexos, como o de Manaus, onde a variabilidade logística e assistencial é maior, a inclusão de preditores contextuais como "pandemia" e "tempo trauma–admissão" melhora substancialmente a performance do modelo, alcançando uma AUC impressionante de 0,98. Este ganho adicional confirma a importância de considerar fatores contextuais, que, ao codificar o ambiente, aumentam a capacidade do modelo de refletir as particularidades locais e de captar as nuances do risco de mortalidade em cenários de alta variabilidade (Oliveira et al., 2021; Lima et al., 2022).

No segundo eixo, a análise de generalização, realizada nas Estratégias 3 e 4, oferece uma mensagem cautelosa, mas também valiosa para a prática de modelagem clínica. A queda acentuada no desempenho quando um modelo treinado em uma base regional é testado em outra, com a AUC caindo para valores próximos a 0,77, revela um desalinhamento nas distribuições das populações. Este decréscimo de desempenho pode ser interpretado como uma mudança de conceito que afeta a relação entre preditores e o desfecho de interesse. A diferença de características entre as populações de São Paulo e Manaus, como os fatores sociais, assistenciais e ambientais, pode levar a uma desconexão no modelo, comprometendo sua eficácia em cenários não homogêneos (Gama et al., 2014; Kairouz et al., 2021). Essa constatação reforça a necessidade de adaptação local dos modelos clínicos, seja por ajuste fino com amostras locais ou pelo uso de técnicas de adaptação de domínio, como aprendizado federado (Pan & Yang, 2010; Torrey & Shavlik, 2010; Li et al., 2020). Esse ponto é especialmente relevante, pois a implementação de

modelos globais sem levar em conta as variações regionais pode resultar em perdas significativas de precisão e confiabilidade.

O desempenho intermediário da Estratégia 5, com AUC inferior ao dos melhores cenários intrarregionais, reforça a tese de que misturar populações heterogêneas sem mecanismos explícitos de estratificação/contextualização pode diluir sinais preditivos específicos. Um caminho promissor, portanto, é treinar modelos hierárquicos (Gelman & Hill, 2006), ou arquiteturas que incorporem explicitamente um vetor de contexto na entrada, como demonstrado no ganho observado em Manaus ao adicionar variáveis específicas do cenário.

Por fim, o terceiro eixo, a interpretabilidade dos modelos, foi abordado por meio das análises de SHAP (Lundberg & Lee, 2017) e de Pearson. A interpretação dos resultados via Pearson, que resume relações lineares médias entre cada variável e a mortalidade, é fundamental para uma visão geral da influência de cada preditor. No entanto, os valores de SHAP oferecem uma análise mais detalhada, permitindo entender, instância por instância, como cada variável modifica a probabilidade de óbito, especialmente em interações não lineares com outros atributos. A complementaridade entre ambas as análises fortalece a confiança clínica nas variáveis mais impactantes, como pontuação motora, reatividade pupilar, desvio de linha média, hipóxia e hipotensão, que emergem como fatores-chave na predição de mortalidade.

O mais interessante é que, ao restringir o treinamento aos cinco melhores preditores de cada base, o desempenho manteve-se muito próximo ao dos cenários com maior número de variáveis. Em São Paulo, a AUC foi apenas discretamente inferior ao obtido com 15 variáveis, e em Manaus, o modelo com apenas 5 preditores alcançou uma AUC de 0,97, quase equivalente ao modelo com 17 variáveis. Esses resultados sugerem um fenômeno desejável na modelagem clínica: a informação útil está concentrada em um pequeno subconjunto de variáveis, o que não só torna os modelos mais simples e interpretáveis, mas também favorece treinamentos mais rápidos e com menos risco de sobreajuste, mesmo em amostras moderadas (Guyon & Elisseeff, 2003).

Essas observações têm implicações diretas para a implementação de modelos em fluxos de emergência, onde o tempo e a completude dos dados são frequentemente limitantes. A necessidade de múltiplos exames laboratoriais, imagens adicionais ou dados administrativos pode atrasar a tomada de decisões, o que em contextos de emergência pode ser um obstáculo. Modelos que utilizam um pequeno número de variáveis, mas

mantêm alta sensibilidade e especificidade, são intrinsecamente mais viáveis para implementação em cenários clínicos reais, onde a rapidez e a eficiência são essenciais. Além disso, do ponto de vista estatístico, a redução da dimensionalidade para além do "ponto ótimo" ajuda a mitigar o sobreajuste e melhora a estabilidade do treinamento, o que foi observado nas comparações entre os cenários com 15 e 17 variáveis em Manaus.

Capítulo 6

Conclusão

Este trabalho teve como objetivo a aplicação de técnicas de aprendizado de máquina, em particular redes neurais convolucionais, para a predição de mortalidade em até 14 dias de pacientes com traumatismo cranioencefálico.

Foi possível identificar que as redes neurais convolucionais apresentaram o melhor desempenho em ambas as bases de dados, destacando-se pela sua capacidade de capturar características complexas e interações não lineares entre as variáveis clínicas.

A análise de generalização cruzada evidenciou queda significativa no desempenho, indicando baixa transferência entre modelos treinados em diferentes regiões. Esse resultado sugere que, para aplicação prática, é necessária a adaptação local dos modelos.

A unificação das bases resultou em desempenho intermediário, apontando que simplesmente aumentar o volume de dados não é suficiente quando há heterogeneidade estrutural entre as populações. Nesse contexto, a inclusão de variáveis que descrevem o cenário assistencial mostrou-se fundamental para melhorar a acurácia.

Este estudo confirma que modelos de aprendizado profundo, quando enriquecidos com variáveis contextuais e interpretados por técnicas robustas, podem atingir alto desempenho na predição de mortalidade precoce em TCE.

Referências Bibliográficas

[1] Dewan MC, *et al*. (2019). Estimating the global incidence of traumatic brain injury. *J Neurosurg*, 130(4):1080–1097.

[2] Amorim RL, *et al*. (2019). Prediction of Early TBI Mortality Using a Machine Learning Approach in a LMIC Population. *Front Neurol*, 10:1366.

[3] Nôvo PC, *et al*. (2023). Neurosurgical Emergencies in the Amazon: An Epidemiologic Study of Patients Referred by Air Transport. *World Neurosurg*, 173:e359–e363.

[4] Raj R, *et al*. (2022). Outcome prediction in traumatic brain injury: comparison of the IMPACT and CRASH prognostic models. *Journal of Neurotrauma*, 39:500-511.

[5] Tu Y, *et al*. (2022). A Computer-Assisted System for Early Mortality Risk Prediction in Patients with Traumatic Brain Injury Using Artificial Intelligence Algorithms in Emergency Room Triage. *Brain Sciences*, 12(5):612.

[6] Zimmerman A, *et al*. (2023). Machine learning models to predict TBI outcomes in Tanzania: Using delays to emergency care as predictors. *PLOS Glob Public Health*, 3:e0002156.

[7] Senders JT, *et al*. (2018). Machine Learning and Neurosurgical Outcome Prediction: A Systematic Review. *World Neurosurg*, 109:476-486.

[8] Courville A, *et al*. (2023). Machine Learning Algorithms for Predicting Outcomes of Traumatic Brain Injury: A Systematic Review and Meta-Analysis. (necessita detalhes completos adicionais)

[9] Yuan Q, *et al*. (2018). Application of machine learning methods in predicting traumatic brain injury outcomes: A review. *Journal of Neuroscience Methods*, 310:42–53.

[10] Kashkoush AI, *et al*. (2022). Predictors of mortality, withdrawal of life-sustaining measures, and discharge disposition in octogenarians with subdural hematomas. *Journal of Neurosurgery*, 136(6):1601–1610.

[11] Hsu SD, *et al*. (2021). Machine learning algorithms to predict in-hospital mortality in patients with traumatic brain injury. *Journal of Personalized Medicine*, 11(11):1168.

[12] Guimarães KAA, *et al*. (2022). Predicting early TBI mortality with 1D CNN and ML techniques. *Informatics in Medicine Unlocked*, 31:100984.

[13] Ding K, *et al*. (2022). Mobile telephone follow-up assessment of postdischarge death and disability due to trauma in Cameroon: a prospective cohort study. *BMJ Open*, 12:e056433.

[14] Souza R, *et al*. (2022). Evaluation of Computed Tomography Scoring Systems in the Prediction of Short-Term Mortality in Traumatic Brain Injury Patients. *Neurocritical Care*, 37:89-97.

[15] Fonseca A, *et al*. (2022). Machine Learning Models for Traumatic Brain Injury Mortality Prediction in Pediatric Electronic Health Records. *Frontiers in Neurology*, 13:910435.

[16] Cerasa A, *et al*. (2022). Predicting Outcomes in Patients with Brain Injury: A Comparison of Machine Learning and Conventional Statistical Methods. *Journal of Neurotrauma*, 39(17–18):1264–1273.

[17] Wang Y, *et al*. (2023). Prediction performance of the machine learning model in predicting mortality risk in patients with traumatic brain injuries: a systematic review and meta-analysis. *Journal of Neurotrauma*, 40(5):422–434.

[18] Mekkodathil A, *et al*. (2023). Machine learning approach for prediction of in-hospital mortality in traumatic brain injury using bio-clinical markers at presentation. *Scientific Reports*, 13:2225.

[19] Kashkoush A, *et al*. (2023). Mortality and discharge outcomes in elderly patients with traumatic brain injuries: A retrospective cohort study. *World Neurosurgery*, 170:e178–e187.

[20] Steyerberg EW, *et al*. (2008). Predicting outcome after traumatic brain injury: Development and international validation of prognostic scores based on admission characteristics. *PLoS Med*, 5(8):e165.

[21] Faried A, *et al*. (2018). Feasibility of Online Traumatic Brain Injury Prognostic Corticosteroids Randomisation After Significant Head Injury (CRASH) Model as a Predictor of Mortality. *World Neurosurg*, 116:e239–e245.

[22] Solla DJF, *et al*. (2021). Incremental Prognostic Value of Coagulopathy in Addition to the Crash Score in TBI Patients. *Neurocrit Care*, 34:130–138.

[23] Abujaber A, *et al*. (2020). Prediction of in-hospital mortality in patients with post traumatic brain injury using National Trauma Registry and ML Approach. *Scand J Trauma Resusc Emerg Med*, 28:44.

[24] Warman PI, *et al*. (2022). Machine Learning for Predicting In-Hospital Mortality After TBI in HICs and LMICs. *Neurosurgery*, 90(5):605–612.

[25] Zhang Y, *et al*. (2021). Data preprocessing and feature engineering in deep learning: a review. *Neurocomputing*, 423:657–675.

[26] Little RJA, Rubin DB. (2019). *Statistical Analysis with Missing Data*. Wiley.

[27] Rodgers JL, Nicewander WA. (1988). Thirteen ways to look at the correlation coefficient. *The American Statistician*, 42(1):59–66.

[28] Lundberg SM, Lee SI. (2017). A Unified Approach to Interpreting Model Predictions. *Advances in Neural Information Processing Systems*, 30.

[29] XAI Tutorials. (2024). SHAP — SHapley Additive exPlanations. Disponível em: https://xai-tutorials.readthedocs.io/en/latest/_model_agnostic_xai/shap.html. Acesso em: 21 maio 2025.

[30] Hosmer DW, Lemeshow S, Sturdivant RX. (2013). *Applied Logistic Regression*. 3rd ed. Wiley.

[31] Khan, Mohammad & Masud, Mehedi & Aljahdali, Sultan & Kaur, Manjit & Singh, Parminder. (2021). A Comparative Analysis of Machine Learning Algorithms to Predict Alzheimer's Disease. Journal of Healthcare Engineering. 2021. 10.1155/2021/9917919.

[32] InfoAryan. (2022). Floresta randômica Algorithm Explained. Disponível em: https://www.infoaryan.com/random-forest-algorithm-explained.

[33] Breiman L. (2001). Floresta randômicas. *Machine Learning*, 45(1):5-32.

[34] Goodfellow I, Bengio Y, Courville A. (2016). *Deep Learning*. MIT Press.

[35] LeCun Y, Bengio Y, Hinton G. (2015). Deep Learning. *Nature*, 521:436-444.

[36] LeCun Y, Bottou L, Bengio Y, Haffner P. (1998). Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278-2324.

[37] Nvidia. (2024). Convolutional Neural Network (CNN). Glossary – NVIDIA. Disponível em: https://www.nvidia.com/en-in/glossary/convolutional-neural-network/.

[38] Nehme, I. (2023). Um guia detalhado para redes neurais convolucionais — método ELI5. W3D Community.

[39] Srivastava N, *et al*. (2014). Dropout: A Simple Way to Prevent Neural Networks from Overfitting. *Journal of Machine Learning Research*, 15:1929-1958.

[40] Yang X, *et al*. (2023). Predicting Models for Local Sedimentary Basin Effect Using a Convolutional Neural Network. *Applied Sciences*, 13:9128.

[41] Nair V, Hinton GE. (2010). Rectified Linear Units Improve Restricted Boltzmann Machines. *Proceedings of the 27th International Conference on Machine Learning (ICML)*, p. 807–814.

[42] Kingma DP, Ba J. (2015). Adam: A Method for Stochastic Optimization. *International Conference on Learning Representations (ICLR)*.

[43] Tieleman T, Hinton G. (2012). RMSProp Optimization Algorithm. Lecture notes, *COURSERA: Neural Networks for Machine Learning*.

[44] Qian N. (1999). On the momentum term in gradient descent learning algorithms. *Neural Networks*, 12(1):145–151.

[45] Krizhevsky A, Sutskever I, Hinton GE. (2012). ImageNet Classification with Deep Convolutional Neural Networks. *Advances in Neural Information Processing Systems*, 25:1097–1105.

[46] Shickel B, Tighe PJ, Bihorac A, Rashidi P. (2018). Deep EHR: A Survey of Recent Advances in Deep Learning Techniques for Electronic Health Record Analysis. *Journal of Biomedical Informatics*, 83:168–185.

[47] Szegedy C, Liu W, Jia Y, *et al*. (2015). Going deeper with convolutions. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, p. 1–9.

[48] Razzak MI, Imran M, Xu G. (2019). Big Data Analytics for Preventive Medicine. *Neural Computing and Applications*, 32(3):923–940.

[49] Sokolova M, Lapalme G. (2009). A systematic analysis of performance measures for classification tasks. *Information Processing & Management*, 45(4):427–437.

[50] Bradley AP. (1997). The use of the area under the ROC curve in the evaluation of machine learning algorithms. *Pattern Recognition*, 30(7):1145–1159.

[51] Gama J, Žliobaitė I, Bifet A, Pechenizkiy M, Bouchachia A. A survey on concept drift adaptation. *ACM Computing Surveys*. 2014;46(4):44. doi:10.1145/2523813.

[52] Kairouz P, McMahan HB, Avent B, Bellet A, Bennis M, Bhagoji AN, et al. Advances and open problems in federated learning. *Foundations and Trends® in Machine Learning*. 2021;14(1–2):1-210. doi:10.1561/2200000083.

[53] Torrey L, Shavlik J. Transfer learning. In: Olivas ES, Guerrero JDM, Sober M, Romero FP, Magdalena-Benedito JR, editors. *Handbook of Research on Machine Learning Applications and Trends: Algorithms, Methods, and Techniques*. Hershey: IGI Global; 2010. p. 242–264. doi:10.4018/978-1-60566-766-9.ch011.

[54] Li T, Sahu AK, Talwalkar A, Smith V. Federated learning: Challenges, methods, and future directions. *IEEE Signal Processing Magazine*. 2020;37(3):50–60. doi:10.1109/MSP.2020.2975749.

[55] Gelman A, Hill J. *Data Analysis Using Regression and Multilevel/Hierarchical Models*. Cambridge: Cambridge University Press; 2006.

[56] Vickers AJ, Elkin EB. Decision curve analysis: A novel method for evaluating prediction models. *Medical Decision Making*. 2006;26(6):565–574. doi:10.1177/0272989X06295361.

Apêndice A - Artigo

Neste apêndice é apresentada a cópia do artigo originado deste trabalho. O artigo intitulado *"Deep Learning Models Generalization for Predicting 14-day Mortality in Traumatic Brain Injury Patients"* foi submetido e aceito para apresentação na 47ª Conferência Internacional Anual da *IEEE Engineering in Medicine and Biology Society (EMBC)*, a realizado em Copenhague, Dinamarca.

Deep Learning Models Generalization for Predicting 14-day Mortality in Traumatic Brain Injury Patients *

Fabio Arthur Soares Araújo, Robson L Oliveira de Amorim, Marly Guimarães Fernandes Costa, *Member IEEE*, Henrique Oliveira Martins, Cicero Ferreira Fernandes Costa Filho, *Member IEEE*

***Abstract*— One of the leading causes of morbidity and mortality in the world is Traumatic Brain Injury (TBI). Different outcomes are influenced bay regional access and health infrastructure. In this study, using 17 predictor variables, we evaluate machine learning models performance and generalizability with two different datasets of Brazilian regions. The first region is Manaus, an isolated urban center with differentiated logistical challenges. The second, is São Paulo, an urban center. To the best of our knowledge, this study is the first one that evaluate predictive models in two distinct datasets in the same country. In the results obtained with 1-D convolutional neural network (CNN) models, the area under the ROC curve (AUC) in São Paulo and Manaus were 0.90 and 0.93, respectively. The model trained in São Paulo does not perform well in Manaus. The incorporation of context-specific features, such as time between trauma and admission, and pandemic-related variable significantly increased the model's accuracy in Manaus model, achieving a remarkable AUC of 0.98.**

***Clinical Relevance*— We highlighted the necessity of integrating local variables to improve TBI prediction in different healthcare environments.**

I. Introduction

Traumatic Brain Injury (TBI) contributes to death and disability across all age groups, being a significant global health concern. An estimated 6469 million individuals worldwide suffer from TBIs each year, with falls, violence and road traffic being the most common causes [1]. In low-and middle-income countries, where limited resources often hinder optimal management and rehabilitation of patients [2], the burden of TBI is particularly high. Brazil, with varying access to healthcare in its vast territory, faces huge challenges in addressing the epidemiological consequences of TBI, particularly in remote areas such as the Amazon region [3].

Aiming at guiding clinical decisions and improving patient care, several predictive models have been developed in recently research. Nevertheless, developing reliable and generalizable models is challenging, especially in populations with differing healthcare infrastructures [4]. The delay in treatment access, the variability in available clinical data and diversity of trauma contribute to the low performance of models trained in different contexts.

In Cerasa et al. [5], the authors aimed to evaluate a comparison of the classical line regression (LR) with machine learning (ML) models in predicting the outcome after TBI. The ML models compared were Support Vector Machine (SVM), K-Nearest-Neighbor (KNN), naïve Bayes (NB), decision tree (DT) and an ensemble of ML models. The accuracy for two classes, positive outcome (Glasgow recovery and moderate disability) and negative outcome (severe disability, persistent vegetative state, and death) with 10 cross validations with classical linear model was 84.69%, and with an ensemble of models was 83.67%. The conclusion of the authors was that ML algorithms do not perform better than more traditional regression models in predicting the outcome after TBI. The dataset was acquired in Italy.

In a systematic review and meta-analysis study [6], the authors retrieved information from 47 selected papers. These papers include 122 newly developed ML models and 34 clinically recommended tools. There were 24 ML models predicting out-of-hospital mortality. For these models, the mean values of the sensitivity and specificity were 0.74 and 0.75, respectively. There were 98 ML models predicting in-hospital mortality of TBI patients. For these models, the mean values of the sensitivity and specificity were 0.79 and 0.89, respectively. According to the authors, ML models are relatively accurate in predicting the mortality of TBI. A single model often outperforms traditional scoring tools, but the pooled accuracy of models is close to that of traditional scoring tools.

In Courville et al. [7], the authors presented another systematic review and meta-analysis study. They retrieved information from 15 papers with ML and LR predicting tools. In thirteen studies, the ML tools significantly improved performance when compared with LR method. With both tools, the accuracy was over 80%. The mean AUC values of 0.96, 0.91, 0.89 and 0.83 were obtained for SVM, artificial neural network (ANN), DT and LR, respectively.

In Mekkodathil et al. [8], the authors aimed to identify predictors of in-hospital mortality in TBI patients using ML methods. The dataset was comprised of 922 hospitalized TBI patients. The feature importance analysis indicates that lactic acid, prothrombin time, international normalized ratio (INR), activated partial thromboplastin time (aPTT) and ISS are the most important features in prediction. The AUC scores for LR, SVM, RF and XGBoost models were 0.84, 0.86, 0.86 and 0.85, respectively. The dataset was acquired in Qatar.

*Research supported by SAMSUNG da Amazônia - Brazil.

F. A. S Araújo and M. G. F. Costa are with R&D Center in Electronic and Information Technology, Federal University of Amazonas, Manaus 69067-005, Brazil

C. F. F. Costa Filho is with R&D Center in Electronic and Information Technology, Federal University of Amazonas, Manaus 69067-005, Brazil (corresponding author to provide phone: (55) 92 991464954; e-mail: ecosta@ufam.edu.br).

R. L O. Amorim and H. O Martins are with Faculty of Medicine, Federal University of Amazonas, Manaus, 69020-160, Brazil

Manuscript 892 submitted to 47th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC) . Received February 6, 2025.

Almost all of the papers published in the literature on predicting the mortality of patients with TBI use ML techniques. In this study, we propose new architectures using deep convolutional neural networks for mortality prediction.

This study fills an important gap in the literature: the evaluation of predictive models across two distinct datasets from different geographic regions. By comparing data from Manaus, a remote area with unique logistical challenges, with São Paulo, a densely populated urban center, we aim to evaluate the generalizability of these models. Integrating region-specific variables, such as time to hospital admission, offers a unique opportunity to evaluate how these factors influence TBI outcomes in diverse settings. Finally, we aim to evaluate whether models trained in one region can effectively predict outcomes in another and to identify the most critical variables that impact mortality prediction in these distinct healthcare environments. This research not only contributes to the understanding of TBI outcomes in Brazil but also seeks to inform clinical practices and improve patient management strategies in varying healthcare contexts.

II. Methodology

A. Materials

Two different datasets are used in this study to predict 14-day mortality in patients with traumatic brain injury (TBI).

The first dataset was generated from data collected from patients who were referred to Hospital das Clínicas (São Paulo, Brazil). The data collection period was from March 2012 to January 2015. The final follow-up of patients ended in June 2015. The Institutional Review Board of the University of São Paulo (São Paulo, Brazil) approved this study (CAAE 46831315.3.0000.0068).

The second dataset was collected in Manaus (Amazonas, Brazil), with 469 samples. Data collection was carried out from May 2020 to July 2021, in a tertiary center in the city of Manaus. This study was approved by the Ethics Committee of the Federal University of Amazonas (UFAM) (CAAE: 25366619.1.0000.5020).

The first dataset comprises 517 records and 15 predictor variables, categorized in four classes: 1. Demographic: gender (categorical: male or female) and age (numerical: years); 2. Clinical: level of pupil reactivity at admission (categorical: bilateral reagent, one or two fixed pupils), GCS at trauma site (categorical: mild, moderate, and severe), GCS at admission (categorical: mild, moderate, and severe), the motor score component of the GCS (categorical, 1-6), presence of hypoxia (categorical: yes or no) and hypotension at admission (categorical: yes or no). Hypotension corresponds to a systolic blood pressure < 90 mmHg and hypoxia corresponds to oxygen saturation < 90%, as recommended by the Brain Trauma Foundation, at any time before hospital admission [12]; 3. Tomographic: midline shift (MLS) on CT greater than 5 mm (categorical: yes or no), subarachnoid hemorrhage – TSAH (categorical: yes or no), epidural hematoma (categorical: yes or no), subdural hemorrhage, intracerebral hemorrhage (categorical: yes or no). These variables refer to hospital admission tomography or the first CT scan findings for patients who come from other hospitals; 4. Laboratory: prothrombin time (numerical: seconds) and partial thromboplastin time ratio - rAPTT (numerical: seconds).

The same 15 variables of São Paulo dataset, were also collected in Manaus dataset. Additionally, the Manaus dataset included two extra variables: whether the data was collected during the COVID-19 pandemic or not (pandemic – 1 or 0).and time from trauma to admission (time trauma admission - hours).

The reason variable x is only present in the Manaus database is as follows: Manaus, the capital of Amazonas state, is the only city in the state that has a medical center that treats patients in need of acute neurosurgical care. Patients from the countryside are transported by river and aerial transfer [21]. Even through aerial transfer, the mean time for the patients to reach the hospital for a neurological emergency consultation was 67.1 hours [3]. This is not the case for patients of São Paulo, which has an excellent road network and several medical centers for patients in need of acute neurosurgical care in the countryside.

Concerning the pandemic variable, its inclusion in Manaus dataset was only possible because part of the data collection took place during the pandemic. Its inclusion aimed to assess whether the overload of the hospital network during the pandemic influenced the 14-day mortality of TBI patients.

B. Methods

Fig. 1 shows a block diagram of the methodology proposed in this study.

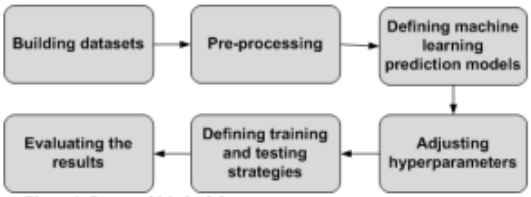


Figure 1. Proposed Methodology

The São Paulo dataset preprocessing was already described by Guimarães et al. [14]. The preprocessing steps reported in this study included filling in missing values using different methods, depending on the variable type. For numerical values, decision tree, random forest, and linear regression were applied. For categorical variables, techniques such as decision tree, random forest, and k-nearest-neighbor were used to fill in the missing values. Additionally, data normalization was performed to ensure that variables such as age and motor score, which had different ranges, were standardized, using techniques like min-max normalization. The Manaus dataset was preprocessed using the same steps previously described: data normalization and filling in missing values.

In this study we employed two CNN architectures. These architectures were designed to extract and refine features from the input data progressively. The first architecture, CNN1, shown in Fig. 2(a), has a parallel structure and, like the inception block of the Google Net [9], use filters with different kernels size: 1x2, 1x3, …1xk. With a smaller filter, less predicting inputs are used in 1D convolution. With a large

Manuscript 892 submitted to 47th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC) . Received February 6, 2025.

filter, more predicting inputs are used in 1D convolution. The variation in kernel size makes it possible to integrate contributions from different predicting variable sets into the final prediction. The best results were achieved with a kernel of size 4. The parallel outputs are concatenated and passed through a dense layer, with 50 neurons and ReLU activation. To mitigate the overfitting, improving the CNN generalization, we used a dropout layer with a rate of 0.2. The output layer uses a sigmoid activation function for binary classification.

The second architecture, CNN2, shown in Fig. 2(b), has a sequential structure comprising 1D convolutional blocks of characteristic extraction. Each extraction block is comprised of a 1D convolutional layer, followed by batch normalization and ReLU activation. The convolution operations maintain the size of the input representation. The best results were achieved with two characteristic extraction blocks. After characteristic extraction blocks, we have a dense layer, with 50 neurons and ReLU activation, followed by a dropout layer with a rate of 0.2. To mitigate the overfitting, improving the CNN generalization, we used a dropout layer with a rate of 0.2. The output layer uses a sigmoid activation function for binary classification.

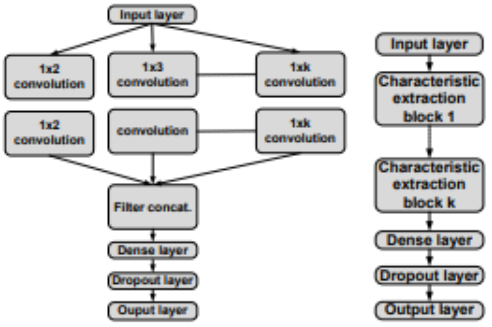


Figure (2). CNN architectures used for predicting 14-day mortality in patients with TBI. (a) CNN1 with parallel architecture; (b) CNN2 with sequential architecture.

After some experiments, the CNNs training were configured with a set of the following hyperparameters: learning rate: 1e-3; optimizers: optimization algorithm: root mean square propagation (RMSProp); number of training epochs: 300; callbacks for saving the best result in training, using the validation accuracy. During the training, the learning rate was progressively reduced, until reaches a value of 1e-6.

To assess the predictive power of deep learning models on 14-day mortality, 5 strategies were employed in this study, using several data combinations (Fig. 3). In experiments 1 and 2, the models are trained and tested separately with each dataset's training and test sets. The aim is to verify the predictive power on each dataset, separately. In experiments 3 and 4, the models are trained with the training set from one dataset and tested with the test set from the other dataset. The aim is to assess the generalization power of the models trained on different datasets. In experiment 5, the models are trained and tested with both the training and test sets from both datasets. The aim is to assess whether a model can capture the peculiarities of each dataset and performs well. Strategies 1, 3, 4 and 5 use only the 15 input variables common to both datasets and shown in Table 2. Strategy 2 uses 15, 16 and 17 variables present in Manaus dataset.

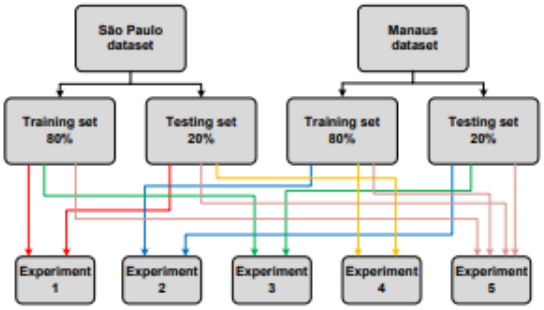


Figure 3. Experiments planned for training and testing the deep learning models.

The models' performances were evaluated using the metrics accuracy, F1-score, and the area under the ROC curve (AUC). The accuracy of a classification system is the degree of closeness of the classification to its actual value. The F1-score is the harmonic mean of the precision and recall, symmetrically representing both metrics. The AUC expresses the trade-off between sensitivity and specificity for different cutoff points in the estimated probability. The higher the AUC value, the greater the discriminatory power of a model [10][11].

In this work, a negative classification indicates the patient's survival in 14 days and a positive classification indicates the patient's death within 14 days. The accuracy and F1-score are described in (1) and (2).

$$accuracy = (TP + TN)/(TP + TN + FP + FN) \quad (1)$$

$$F1\text{-}score = 2*(prec.*recall)/(prec.+recall) \quad (2)$$

Where: TP – True Positive; FP – False Positive; TN – True Negative and FN – False Negative.

III. Results

This section presents results for experiments 1 to 5, using the metrics presented in the last section, and confusion matrix tables. To explain the results, we will show values obtained by the SHapley Additive exPlanations (SHAP) technique [12]. (SHAP) is a game theory-based method for explaining the output of classifying models. SHAP uses Shapley values to assign credit to each feature or feature value for a model's prediction.

Table I shows the results of experiments 1 and 2, while Fig. 4 shows the confusion table for both experiments, with 15 input variables.

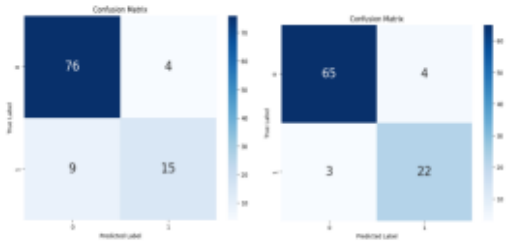
Manuscript 892 submitted to 47th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC) . Received February 6, 2025.

All the metric values for expeiment 2 are better than those obtained for experiment 1. Particularly, the best AUC value for the Manaus dataset is 0.93, while for the São Paulo dataset is 0.91. These values were obtained with the RMSProp optimizer. From the confusion matrix shown in Fig. 4, the sensitivity and specificity for experiment 1 are 0.79 and 0.89, respectively. For experiment 2, 0.84 and 0.96, respectively.

Table II shows the results for experiment 2, with 15, 16 and 17 input variables, while Fig.5 shows the confusion table for experiment II with 17 variables. The pandemic variable provides important context regarding whether the data was collected during the COVID-19 pandemic. The time trauma admission variable captures the time between the trauma and hospital admission. As shown, when using the 17 variables, the AUC increases to 0.98. From the confusion matrix shown in Fig. 5 shows, for 17 variables, the sensitivity and specificity are 0.92 and 0.99, respectively.

TABLE I. RESULTS FOR EXPERIMENTS 1 AND 2 WITH 15 INPUT VARIABLES.

CNN	Optimizer	Experiment 1			Experiment 2		
		Accuracy	F1-Score	AUC	Accuracy	F1-Score	AUC
CNN1	Adam	0.81	0.8	0.83	0.9	0.88	0.91
CNN1	RMSprop	0.82	0.81	0.84	0.92	0.9	0.93
CNN1	SGDM	0.8	0.78	0.81	0.88	0.86	0.89
CNN2	Adam	0.86	0.83	0.89	0.89	0.87	0.9
CNN2	RMSprop	0.87	0.85	0.9	0.9	0.89	0.91
CNN2	SGDM	0.85	0.82	0.87	0.87	0.85	0.89



(a) (b)

Figure 4. Confusion matrices for: (a) experiment 1, with the CNN2 and RMSProp optimizer; (b) experiment 2, with CNN1 and RMSProp optimizer.

TABLE II. RESULTS FOR EPXERIMENT 2, WITH 15, 16 AND 17 INPUT VARIABLES, WITH RMSPROP OPTIMIZER.

Experiment/ N. of variables	Accuracy	F1-Score	AUC
Experiment 2/ 15 variables	0.92	0.9	0.93
Experiment 2/ 15 variables + pandemic	0.95	0.94	0.97
Experiment 2/ 15 variables+ time trauma admission	0.95	0.96	0.97
Experiment/ 15 variables + pandemic + time trauma admission	0.97	0.96	0.98

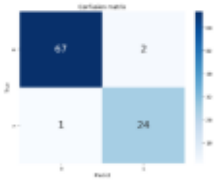


Figure 5. Confusion matrix for experiment 2 with 17 variables and RMSProp optimizer.

Table III shows the results for experiments 3 and 4, with 15 input variables. Compared with the results of experiments 1 and 2 shown in Table I, the models performances decreased. For example, the best AUC for experiments 3 and 4, obtained with CNN1 and RMSProp optimizer, are 0.77 and 0.70, respectively. These values are lower than those obtained with experiments 1 and 2, 0.90 and 0.93, respectively. We also note that training with the Manaus dataset and testing with data from the São Paulo dataset (experiment 3) results in better metric values than training with the São Paulo dataset and testing with the Manaus dataset (experiment 4).

TABLE III. RESULTS FOR EPXERIMENT 3 AND 4, WITH 15 INPUT VARIABLES AND RMSPROP OPTIMIZER

CNN	Optimizer	Experiment 3			Experiment 4		
		Accuracy	F1-Score	AUC	Accuracy	F1-Score	AUC
CNN1	Adam	0.38	0.53	0.52	0.73	0.5	0.69
CNN1	RMSprop	0.27	0.42	0.52	0.77	0.52	0.7
CNN1	SGDM	0.35	0.51	0.51	0.71	0.49	0.68
CNN2	Adam	0.37	0.53	0.52	0.75	0.51	0.7
CNN2	RMSprop	0.36	0.52	0.52	0.74	0.51	0.69
CNN2	SGDM	0.35	0.51	0.5	0.73	0.5	0.69

Table IV shows the results for experiment 5, with 15 input variables, while Fig. 6 shows the respectively confusion matrix. Compared with the results of experiments 1 and 2, shown in Table I, the models performances decreased. For example, the best AUC for experiment 5, obtained with CNN1 and RMSProp optimizer, is 0.83. This value is lower than those obtained with expriments 1 and 2, 0.90 and 0.93, respectively, but better than the value obtained with strategies 3 and 4.

Fig. 7 shows the SHAP plots for both datasets. In each plot, the Y-axis indicates the feature names in order of importance from top to bottom. The X-axis represents the SHAP values, which means the degree of change in log odds. The color of each point on the graph represents the value of the corresponding feature, with red indicating high values and blue indicating low values. Each point represents a row of data from the original dataset.

As shown in the SHAP plot for the São Paulo dataset (Fig. 7(a)), the 5 most essential variables in model prediction include *motor score*, *pupil reactivity*, *midline shift*, *GCS at admission* and *GCS at trauma site*. As shown in the SHAP plot for the Manaus dataset (Fig. 7(b)), the 5 most essential variables in model prediction include *motor score*, *pupil reactivity*, *hypoxia*, *midline shift* and *hypotension at admission*.

Manuscript 892 submitted to 47th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC) . Received February 6, 2025.

TABLE IV. RESULTS FOR EPXERIMENT 5, WITH 15 INPUT VARIABLES AND RMSPROP OPTIMIZER

Model	Optimizer	Accuracy	F1-Score	AUC
CNN1	Adam	0.81	0.63	0.73
CNN1	RMSprop	0.83	0.69	0.77
CNN1	SGDM	0.70	0.09	0.52
CNN2	Adam	0.81	0.62	0.72
CNN2	RMSprop	0.80	0.61	0.72

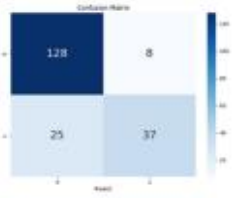
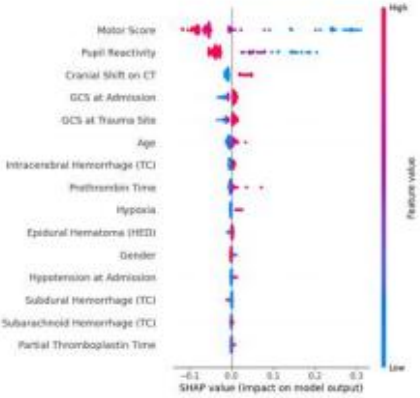
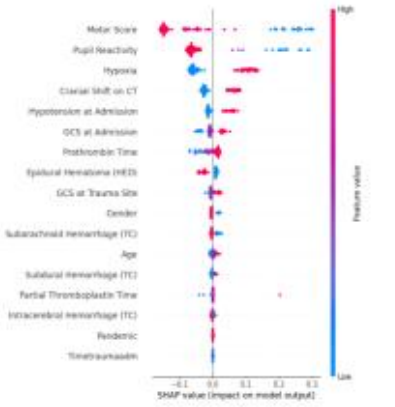


Figure 6. Confusion matrix for experiment 5 with 15 variables and RMSProp optimizer.



(a)



(b)

Figure 7. Shap values for predictions of model CNN1 with RMSprop optimizer. (a) São Paulo dataset, with 15 input variables(b) Manaus dataset with 17 input variables.

IV. DISCUSSION AND CONCLUSION

This study aimed to assess the performance and transferability of deep learning models for predicting 14-day mortality in TBI patients across two distinct Brazilian regions. The results shown that models trained in one region faced challenges in maintaining their predictive power when applied to data from the other. This suggests that regional factors significantly shape TBI patient outcomes, a conclusion supported by previous research that highlights the variability in healthcare access and infrastructure across different geographical settings [13][2]. The deep models performed well in each separated dataset, with São Paulo achieving an area under the curve (AUC) of 0.90 and Manaus showing an AUC of 0.93. This performance was anticipated and aligns with previous findings from our group when studying the São Paulo dataset [14][2].

In Manaus dataset, incorporation of context-specific features, particularly those related to the COVID-19 pandemic and the time from trauma to admission, significantly enhanced model accuracy, achieving an impressive AUC of 0.98 upon their inclusion. Manaus, as one of the epicenters of the COVID-19 pandemic, faced unique challenges that influenced healthcare delivery and patient outcomes during this period [15]. Therefore, the pandemic may exacerbated issues related to healthcare access, particularly for individuals residing in rural areas where road infrastructure is limited [3]. The addition of these features not only improved the models' performance but also underscored the necessity of tailoring predictors to enhance inter-regional applicability.

In terms of generalizability, our findings offer insights into the broader applicability of mortality prediction models for TBI patients across differing healthcare environments. Although performance varied by region, the incorporation of locally relevant variables led to more accurate predictions, particularly for the Manaus dataset.

ACKNOWLEDGMENT

This research, carried out within the scope of the Samsung-UFAM Project for Education and Research (SUPER), according to Article 39 of Decree n°10.521/2020, and was funded by Samsung Electronics of Amazonia Ltda., under the terms of Federal Law n°8.387/1991 through agreement 001/2020, signed with UFAM and FAEPI, Brazil.

REFERENCES

[1] M. C. Dewan *et al.*, "Estimating the global incidence of traumatic brain injury.," *J. Neurosurg.*, vol. 130, no. 4, pp. 1080–1097, Apr. 2019, doi: 10.3171/2017.10.JNS17352.

[2] R. L. Amorim *et al.*, "Prediction of Early TBI Mortality Using a Machine Learning Approach in a LMIC Population.," *Front. Neurol.*, vol. 10, p. 1366, 2019, doi: 10.3389/fneur.2019.01366.

[3] P. C. Nôvo, S. A. B. de Farias, V. do V. Guttemberg, V. R. Félix Dos Santos, J. P. Moreira Guilherme, and R. L. O. de Amorim, "Neurosurgical Emergencies in the Amazon: An Epidemiologic Study of Patients Referred by Air Transport for Neurosurgical Evaluation at a Referral Center in Amazonas.," *World Neurosurg.*, vol. 173, pp. e359–e363, May 2023, doi: 10.1016/j.wneu.2023.02.056.

[4] R. Raj *et al.*, "Dynamic prediction of mortality after traumatic brain injury using a machine learning algorithm," *npj Digit. Med.*, vol. 5, no. 1, p. 96, 2022, doi: 10.1038/s41746-022-00652-3.

[5] A. Cerasa *et al.*, "Predicting Outcome in Patients with Brain Injury: Differences between Machine Learning versus Conventional Statistics," *Biomedicines*, vol. 10, no. 9. 2022, doi: 10.3390/biomedicines10092267.

[6] J. Wang, M. J. Yin, and H. C. Wen, "Prediction performance of the machine learning model in predicting mortality risk in patients with traumatic brain injuries: a systematic review and meta-analysis," *BMC Medical Informatics and Decision Making*, vol. 23, no. 1. 2023, doi: 10.1186/s12911-023-02247-8.

[7] E. Courville *et al.*, "Machine learning algorithms for predicting outcomes of traumatic brain injury: A systematic review and meta-analysis," *Surgical Neurology International*, vol. 14. 2023, doi: 10.25259/SNI_312_2023.

[8] A. Mekkodathil, A. El-Menyar, M. Naduvilekandy, S. Rizoli, and H. Al-Thani, "Machine Learning Approach for the Prediction of In-Hospital Mortality in Traumatic Brain Injury Using Bio-Clinical Markers at Presentation to the Emergency Department," *Diagnostics*, vol. 13, no. 15. 2023, doi: 10.3390/diagnostics13152605.

[9] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the Inception Architecture for Computer Vision," *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, vol. 2016-Decem, pp. 2818–2826, 2016, doi: 10.1109/CVPR.2016.308.

[10] T. Fawcett, "An introduction to ROC analysis," *Pattern Recognit. Lett.*, vol. 27, no. 8, pp. 861–874, 2006, doi: 10.1016/j.patrec.2005.10.010.

[11] D. M. W. Powers, "Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation," no. May, 2020, [Online]. Available: http://arxiv.org/abs/2010.16061.

[12] S. M. Lundberg and S. I. Lee, "A unified approach to interpreting model predictions," *Adv. Neural Inf. Process. Syst.*, vol. 2017-Decem, no. Section 2, pp. 4766–4775, 2017.

[13] P. I. Warman *et al.*, "Machine Learning for Predicting In-Hospital Mortality After Traumatic Brain Injury in Both High-Income and Low- and Middle-Income Countries.," *Neurosurgery*, vol. 90, no. 5, pp. 605–612, May 2022, doi: 10.1227/neu.0000000000001898.

[14] K. A. A. Guimarães, R. L. O. de Amorim, M. G. F. Costa, and C. F. F. Costa Filho, "Predicting early traumatic brain injury mortality with 1D convolutional neural networks and conventional machine learning techniques," *Informatics Med. Unlocked*, vol. 31, pp. 1–23, 2022, doi: 10.1016/j.imu.2022.100984.

[15] E. Fernandes *et al.*, "Exploring Prehospital Data for Pandemic Preparedness: A Western Brazilian Amazon Case Study on COVID-19.," *Int. J. Environ. Res. Public Health*, vol. 21, no. 9, Sep. 2024, doi: 10.3390/ijerph21091229.

Manuscript 892 submitted to 47th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC) . Received February 6, 2025.

Apêndice B - Artigo

Neste apêndice é apresentada a cópia do artigo originado deste trabalho. O artigo intitulado *"Evaluating the Generalization of Machine Learning Models for Predicting 14-day Mortality in Traumatic Brain Injury Patients"* foi submetido e aceito para publicação no periódico *Biocybernetics and Biomedical Engineering*.

Evaluating the Generalization of Machine Learning Models for Predicting 14-day Mortality in Traumatic Brain Injury Patients

Abstract

Traumatic Brain Injury (TBI) remains a leading cause of morbidity and mortality worldwide, with significant disparities in outcomes influenced by regional healthcare access and infrastructure. This study evaluates the performance and generalizability of machine learning models for predicting 14-day mortality in TBI patients using datasets from two distinct Brazilian regions: São Paulo, an urban center, and Manaus, an isolated urban center with unique logistical challenges. To our knowledge, this research represents the first cross-validation of predictive models across two datasets within the same country, underscoring the critical need for localized approaches in TBI research. Our findings indicate that while convolutional neural network (CNN)-based models achieved high performance, with an area under the curve (AUC) of 0.90 in São Paulo and 0.93 in Manaus, the best model from São Paulo exhibited a strikingly low AUC when applied to the Manaus dataset. The incorporation of context-specific features, such as pandemic-related variables and time from trauma to admission, significantly enhanced model accuracy, with the Manaus model reaching an impressive AUC of 0.98. Notably, the study highlights key regional differences in predictors of mortality, with hypoxia and hypotension being more critical in Manaus, emphasizing the importance of tailoring predictive models to local contexts. Our results indicate that CNN-based models have the potential to enhance mortality predictions for patients with traumatic brain injury (TBI). Additionally, we highlighted the necessity of conducting cross-regional validation and integrating local variables to improve patient outcomes across different healthcare environments.

Keywords: Traumatic brain injury, Machine Learning, Mortality, LMIC, Convolutional Neural Networks

1. Introduction

Traumatic Brain Injury (TBI) is a significant global health concern, contributing to death and disability across all age groups. Each year, an estimated 64 to 69 million individuals worldwide suffer from TBIs, with road traffic accidents, falls, and violence being the most common causes [1]. The burden of TBI is particularly high in low- and middle-income countries, where limited resources often hinder optimal management and rehabilitation of patients [2]. Brazil, with its vast territory and varying access to healthcare, faces considerable challenges in addressing the epidemiological consequences of TBI, particularly in remote areas such as the Amazon region [3].

The development of predictive models for TBI outcomes has been a focus of recent research aimed at guiding clinical decisions and improving patient care. However, building reliable and generalizable models is challenging, particularly in heterogeneous populations with differing healthcare infrastructures [4]. Factors such as variability in available clinical data, diversity of trauma

mechanisms, and delays in treatment access complicate the ability of models to accurately predict outcomes like mortality [5][6]. Machine learning techniques have shown promise in enhancing predictive accuracy; for instance, studies have demonstrated that machine learning models can outperform traditional logistic regression in predicting TBI mortality [2][7]. Nevertheless, the lack of external validation across diverse datasets remains a key obstacle [8].

Table 1 shows a summary of the literature review on papers published in the last years about mortality prediction of traumatic brain injury (TBI) patients.

None of the reviewed studies applied deep neural models to predict mortality in patients with TBI. In Tu et al. [5], to predict in-hospital mortality of TBI patients, the authors evaluated the performance of 6 machine learning (ML) models: linear regression (LR), random forest (RF), support vector machine (SVM), LigthGBM, XGBoost, and multilayer perceptron (MLP). In Cerasa et al. [9], the authors aimed to evaluate a comparison of the classical LR with machine learning models in predicting the outcome after TBI. The ML models compared were SVM, K-Nearest-Neighbor (KNN), naïve Bayes (NB), decision tree (DT) and an ensemble of ML models. In Wang et al. [10], in a systematic review and meta-analysis study, the authors retrieved information from 47 selected papers. These papers include 122 newly developed ML models.

None of the studies addressed the generalization power of models trained on different datasets. In Hsu et al. [11], to predict patient in-hospital mortality using clinical and demographic data, the only dataset used, with 3,331 TBI patients, was acquired in Taiwan Triage and Acuity Scale from January 2008 to June 2018. In Ding et al. [12], the authors evaluated TBI-related death with a dataset was acquired in Cameroon. In Fonseca et al. [13], the authors examine the mortality of pediatric TBI patients using only a dataset acquired in Denver, Colorado, USA. In Rodrigues de Souza et al. [14], the author's objective was to evaluate, in the prediction of 14-day in-hospital mortality, the increase in variance explained when adding each of three computer tomography (CT) classification systems: Marshall computerized tomography (CT) classification and Rotterdam and Helsinki CT scores. The only dataset used was acquired in São Paulo, Brazil.

In systematic literature reviews [15] and [10], there is a divergence in evaluating the performance of machine learning methods against classic linear regression methods. In Courville et al. [15], the authors retrieved information from 15 papers with ML and classical linear regression predicting tools. In thirteen studies, the ML learning tools significantly improved performance when compared with classical linear regression method. With both tools, the accuracy was over 80%. The mean AUC values of 0.96, 0.91, 0.89 and 0.83 were obtained for SVM, artificial neural network (ANN), DT and classical LR, respectively. In another systematic review, Wang et al. [10], the authors retrieved information from 47 selected papers. There were 98 ML models predicting in-hospital mortality of TBI patients. According to the authors, ML models are relatively accurate in predicting the mortality of TBI. A single model often outperforms traditional scoring tools, but the pooled accuracy of models is close to that of traditional scoring tools.

This study addresses a significant gap in the literature: the validation of predictive models across two distinct datasets. Thus, the primary objective of this paper is to evaluate the performance and generalizability of classical and deep machine learning models for predicting 14-day mortality in TBI patients using datasets from two centers in Brazil: São Paulo, a densely populated urban center, and Manaus, in the Amazon region, a more remote area with unique logistical challenges. We aim to explore whether models trained in one region can effectively predict outcomes in another and to identify the most critical variables that impact mortality prediction in these distinct healthcare environments. Using techniques that assess the importance of variables in the performance of prediction tools, we identify a reduced set of the most important variables and investigate the performance of the best predictor when it is trained with this reduced set. This research not only contributes to the understanding of TBI outcomes in Brazil, but also seeks to inform clinical practices and improve patient management strategies in varying healthcare contexts.

Table 1.
Summary of literature review

Reference	Dataset	Predicting Variables	Predicted Variables/Predictor Models	Results
Kashkoush et al., [16]	3,279 TBI admissions at 45 US trauma centers from 2017 to 2019. Analysis of 695 patients aged 80 and over.	GCS, Pupillary reactivity, ISS, Use of anticoagulants/antiplatelets, Comorbidities (e.g., CHF, diabetes), Intraventricular hemorrhage, Neurosurgical intervention	Hospital mortality, hospital discharge with palliative care, withdrawal of life support measures/multivariate logistic regression	Prediction of mortality: AUC = 0.885; prediction of withdrawal of life support measures: AUC = 0.894
Hsu et al., [11]	3,331 TBI patients who visited the emergency department of a high-complexity hospital in northern Taiwan from January 2008 to June 2018.	GCS, ISS, Systolic blood pressure (SBP), Heart rate (HR), Pulse pressure difference (PP), Age, Gender	Hospital mortality/ J48 decision tree (DT), RF, random tree, K-nearest neighbors KNN), naïve Bayes (NB) and SVM.	Best performance: J48 decision tree - AUC > 0.80; Accuracy = 93.2%
Ding et al., 2022 [12]	Patients who were hospitalized for TBI in four hospitals in Cameroon, that were prospectively followed after being discharged, from July 2019 to March 2021.	Age, Gender, Educational level, Injury Severity Score (ISS), Type of fracture (open/closed), Neurological deficit, Mechanism of injury (e.g. fall, stab wound, etc.).	Post-discharge mortality, functional disability measured by the Glasgow Outcome Scale-Extended (GOSE)/ Statistic models	Mortality: 71 deaths recorded, 90% occurred by 2 weeks post discharge; severe functional disability at 6 months: 22.1%; Good recovery: 70.3% at 6 months.
Rodrigues de Souza et al., [14]	447 patients with TBI treated at a tertiary hospital associated with the University of São Paulo, Brazil, from January 2012 to December 2015.	Marshall, Rotterdam and Helsinki CT classification; age, GCS, pupillary response, hypoxia, hypotension, hemoglobin values	14-day mortality, hospital mortality/multi-variate regression	Marshall: AUC = 0.610 (14 days), 0.575 (hospital); Rotterdam: AUC = 0.762 (14 days), 0.712 (hospital); Helsinki: AUC = 0.752 (14 days), 0.716 (hospital).
Tu et al., [5]	18,249 TBI patients admitted to the emergency room of three Chi Mei Medical Group hospitals in Taiwan from January 2010 to December 2019.	Age, Gender, BMI, TTAS, Heart rate, Body temperature, Respiratory rate, GCS, Pupil size (L and R), Pupillary reflex (L and R).	Hospital mortality/LR, RF, SVM, LigthGBM, XGBoost, MLP	Logistic Regression - AUC = 0.925; SVM (AUC = 0.920), MLP (AUC = 0.893), XGBoost (AUC = 0.871),
Fonseca et al., [13]	300 hospitalized pediatric TBI patients from the Pediatric Traumatic Brain Injury (HPTBI) Hackathon dataset	Age, Gender, GCS, CT scan results (e.g. brain swelling, midline shift), Enteral nutrition, Cardiac arrest, Fixed pupils	Mortality at hospital discharge/ XGBoost, KNN, artificial neural network (ANN), DT.	XGBoost: AUC = 0.91; KNN: AUC = 0.90 (with feature selection); RF: AUC = 0.85; ANN: AUC = 0.84
Cerasa et al., [9]	A multi-center dataset with 278 TBI patients. The data was acquired in Italy. The acquisition period is not cited.	Age, GCS, Pupillary Response, Subarachnoid Hemorrhage, Level of Education, Hypotension, Hyperglycemia, Coagulopathy	Hospital mortality, Functional recovery/Classical linear regression, SVM, KNN, NB, DT, Ensemble of models.	Accuracy of 2 classes with 10 folder cross-validation: classical linear model: 84.69%, Ensemble of models: 83.67%.

Reference	Dataset	Predicting Variables	Predicted Variables/Predictor Models	Results
Wang et al., [10]	Meta-analysis of 47 studies, including 156 ML models, with a total of 2,080,819 patients from various parts of the world, mainly in Asia, Europe and North America.	GCS, Age, CT classification, Pupil size, Pupillary reflex, Glucose, Systolic blood pressure	In-hospital mortality, Out-of-hospital mortality/ ML models:	ML models: In-hospital: mean C-Index = 0.86 (95% CI: 0.84-0.87), mean Sensitivity = 0.79, mean Specificity = 0.89. Linear Regression:
Courville et al., [15]	Meta-analysis of 15 studies involving 32,721 patients with TBI, covering various institutions and diverse populations.	Age, GCS, Serum acid level, Abnormal glucose, Pupils, Radiological findings, Arrival methods, Time of day	Hospital mortality, 14-day mortality/ ML models:	ML models: SVM: AUC ≈ 0.96 for hospital mortality; ANN: AUC ≈ 0.91; Decision tree: AUC ≈ 0.89; linear regression: AUC ≈ 0.83
Mekkodathil et al., [17]	922 TBI patients hospitalized at the Hamad Trauma Center in Qatar from June 2016 to May 2021.	GCS, ISS, aPTT, PT, INR, Hemoglobin, Lactic Acid, Sodium, Potassium, Calcium, Magnesium, Phosphate, Bicarbonate Levels	Hospital mortality/ SVM, XGBoost, RF, LR.	SVM: AUC = 0.86; RF: AUC = 0.86; XGBoost: AUC = 0.85; LR: AUC = 0.84
Cao et al., [18]	545,388 patients with severe TBI isolated from the TQIP database (2013-2021).	Age, GCS on admission, Head AIS, Hypotension, Cirrhosis, Epidural hematoma, Shock index, Oxygen saturation, Body temperature, Number of PRBC units transfused	Hospital mortality/ XGBoost-powered Cox model.	XGBoost-powered Cox model: C-index = 0.897 (training), 0.896 (test); Time-dependent AUC ≤ 5 days: 0.917, ≤ 20 days: 0.813

4. Materials and Methods

4.1. Materials

This study utilized two distinct datasets to predict 14-day mortality in patients with traumatic brain injury (TBI) using machine learning models. The first database was generated from data collected from patients who were referred to Hospital das Clínicas (São Paulo, Brazil). Data collection was carried out from March 2012 to January 2015. The final follow-up of patients ended in June 2015. This database contains 517 records with 15 predictor variables, categorized in four classes: 1. Demographic: *gender* and *age* in years; 2. Clinical: *level of pupil reactivity at admission* (bilateral reagent, one or two fixed pupils), *GCS at trauma site* (mild, moderate, and severe), *GCS at admission* (mild, moderate, and severe), the *motor score* component of the GCS, presence of *hypoxia* (yes or no) and *hypotension* at admission (yes or no). *Hypotension* corresponds to a systolic blood pressure < 90 mmHg and *hypoxia* corresponds to oxygen saturation < 90%, as recommended by the Brain Trauma Foundation, at any time before hospital admission [11]; 3. Tomographic: *midline shift* (MLS) *on CT* greater than 5 mm (yes or no), *subarachnoid hemorrhage* – TSAH (yes or no), *epidural hematoma* (yes or no), *subdural hemorrhage*, *intracerebral hemorrhage* (yes or no). These variables refer to hospital admission tomography or the first CT scan findings for patients who come from other hospitals; 4. Laboratory: *prothrombin time* (seconds) and *partial thromboplastin time ratio* - rAPTT (seconds). The Institutional Review Board of the University of São Paulo (São Paulo, Brazil) approved this study (CAAE 46831315.3.0000.0068).

The second dataset was collected in Manaus (Amazonas, Brazil), comprising 469 samples. Data collection was carried out from May 2020 to July 2021, in a tertiary center in the city of Manaus. This study was approved by the Ethics Committee of the Federal University of Amazonas (UFAM) (CAAE: 25366619.1.0000.5020).

Both datasets included crucial demographic, clinical, and laboratory information relevant to the prognosis of TBI patients, ensuring a robust foundation for model development. The same 15 variables collected in São Paulo dataset, were also collected in Manaus dataset. Additionally, the Manaus dataset included two extra variables: time from trauma to admission (*time trauma admission - hours*), whether the data was collected during the COVID-19 pandemic or not (*pandemic – 1 or 0*).

The following arguments justify the inclusion of the variable *time trauma admission* only in the Manaus dataset: Manaus, the capital of Amazonas state, is the only city in the state that has a medical center that treats patients in need of acute neurosurgical care. Patients from the countryside are transported by river and aerial transfer [19]. In [3], the authors started that, even though aerial transfer, the mean time for the patients to reach the hospital for a neurological emergency consultation was 67.1 hours. In this work, nevertheless, the maximum transport time was 12 hours. The reason for such high times, even using air transport, is that aircraft is not available in the city where the TBI occurred. The waiting time for an aircraft is longer than the flight time. This is not the case for patients of São Paulo, which has an excellent road network and several medical centers for patients in need of acute neurosurgical care in the countryside. Including the pandemic variable in the Manaus dataset was only possible because part of the data collection took place during the pandemic. Its inclusion aimed to account for contextual changes during the pandemic period, which may have encompassed factors such as healthcare system strain, shifts in trauma epidemiology, delays in referral and treatment, and broader logistical challenges, all of which could influence 14-day mortality.

Table 2 shows a list of all variables used in both databases, showing the class, type, range, frequency of values and evaluating the statistical significance of the frequency differences. Both datasets were stored in csv files. Patients were recruited consecutively following the inclusion criteria: patients or their legal guardians who signed the informed consent; patients' victims of TBI with brain CT scan abnormalities; and patients with GCS less than or equal to 14 after stabilization at the emergency room and older than 14 years. In Table 2, the GCS variable takes values in the range 1-3. The value 1 corresponds to severe TBI, 2 to moderate TBI, and 3 to severe TBI. The exclusion criteria adopted were the following: patients transferred from a different Intensive Care Unit (ICU), patients with chronic subdural hematoma, and patients with medium-fixed pupils with a GCS of three without recovery after cardiopulmonary resuscitation.

Figure 1 shows, for each dataset, the number of patients who died within 14 days and of patients who survived in 14 days. In São Paulo, 22.82% of the patients died in 14 days, while in Manaus, 27% of patients died in 14 days. Applying the chi-square test for evaluating the differences between different mortality in the two datasets, we obtained $\chi^2 = 2.38$. This value is not significant at 5% ($\chi^2_{critic} = 3.84, dof = 1$).

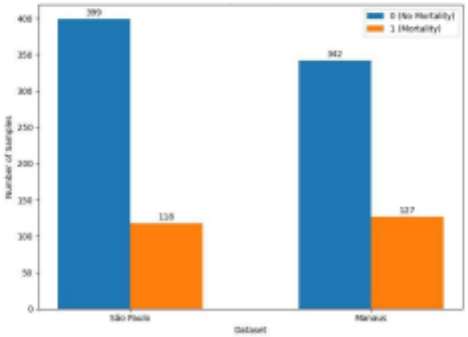


Fig. 1. For each dataset, the number of patients who died within 14 days and of patients that survived in 14 days

4.2 Methods

Figure 2 shows a block diagram of the proposed methodology. Next, we'll evaluate each step of this methodology.

Table 2.
Variables used in the study for 14-day mortality prediction

Class	Variable	Type/ Range	Values	São Paulo Database	Manaus Database	Chi-Square Test/ Student-t test	Statistical significant? (1%)
Demographic	Gender	Categorical/ 0-1	1 0	440 77	414 55	$\chi^2 = 2.12$ $p = 0.14$	No
	Age (Mean ± SD)	Numerical 16-99	-	41.46 ± 17.96	37.56 ± 15.51	$t = 3.6627, p > 0.01$	No
Clinical	Pandemic (Manaus exclusive)	Categorical 0-1	1 0	- -	332 137	- -	- -
	Pupil reactivity / Ca	Categorical 0-2	2 1 0	395 63 59	394 12 63	$\chi^2 = 32.55$ $p < 0.00001$	Yes
	GCS at trauma site	Categorical 1-3	3 2 1	353 93 71	140 250 79	$\chi^2 = 162.36$ $p < 0.00001$	Yes
	GCS at admission	Categorical 1-3	3 2 1	348 97 72	141 249 79	$\chi^2 = 152.75$ $p < 0.00001$	Yes
	Motor score	Categorical 1-6	6 5 4 3 2 1	158 187 61 22 18 71	244 84 16 3 4 118	$\chi^2 = 116.82$ $p < 0.00001$	Yes
	Hypoxia	Categorical 0-1	1 0	56 461	121 346	$\chi^2 = 37.81$ $p < 0.00001$	Yes
	Hypotension at admission	Categorical 0-1	1 0	62 455	99 368	$\chi^2 = 15.19$ $p < 0.000097$	Yes
	Time trauma admission (Manaus exclusive) Mean ± SD	Numerical 0-12h	-	-	0.44 ± 1.17	-	-
Tomographic	Midline shift	Categorical 0-1	1 0	122 395	128 341	$\chi^2 = 1.77$ $p = 0.18$	No
	Subarachnoid hemorrhage (CT)	Categorical 0-1	1 0	222 295	345 124	$\chi^2 = 94.36$ $p < 0.00001$	Yes
	Epidural hematoma (CT)	Categorical 0-1	1 0	407 110	187 282	$\chi^2 = 154.98$ $p < 0.00001$	Yes
	Subdural hemorrhage (CT)	Categorical 0-1	1 0	37 480	172 297	$\chi^2 = 128.27$ $p < 0.00001$	Yes
	Intracerebral hemorrhage (CT)	Categorical 0-1	1 0	256 261	253 216	$\chi^2 = 1.93$ $p < 0.16$	No

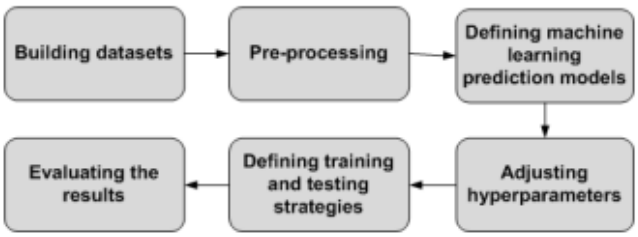


Fig. 2. Proposed Methodology

4.2.1. Preprocessing

Given the importance of data completeness, evaluating and addressing missing values in both datasets was necessary. In the São Paulo dataset, approximately 18% of the samples had at least one

missing value, with some variables, such as "Hypoxia" and "GCS at trauma site," showing higher levels of missing data. In contrast, the Manaus dataset presented a lower percentage of missing data, with around 2% of the samples containing incomplete information.

As shown in Table 2, most of the variables used in this study are categorical. Only two variables, "age" and "time of trauma admission", are numerical. Missing values were more prevalent in the categorical variables. The techniques used for filling missing values in both type of variables were already described by Guimarães et al. [20]. The best performances were obtained with random forest and decision tree methods. For the categorical variables hypoxia, hypotension at admission and level of pupil reactivity at admission, the best results for filling in missing values were obtained with the random forest method (accuracy = 0.809, 0.871, and 0.777, respectively). For the categorical variables GCS at trauma site, GCS at admission and motor score, the best results for filling in the missing values were obtained with the decision tree method (accuracy = 0.508, 0.454 and 0.901, respectively).

Additionally, data normalization was performed to ensure that variable "age" and "time trauma admission", which have different ranges, were standardized using techniques like min-max normalization, cubic root transformation, and z-score normalization, as described in [20].

4.2.2 Defining Machine Learning Predicting Models

In this study, the following models were used for predicting 14-day mortality of TBI patients: logistic regression LR), random forest (RF), multilayer perceptron (MLP), and convolutional neural networks (CNN) [21][22][23]. Each model has a binary output of 1, to indicate the patient's 14-day mortality, or 0, to indicate patient survival in 14 days.

The LR model is the most straightforward neural architecture available, with only one neuro layer, and can be used for both classification and regression tasks. For multiclass classification, the softmax function is used in the output layer. Both the softmax and the sigmoid function can be used for binary classification. This implements a binary classification, using the sigmoid activation function in the output layer. To prevent overfitting, it was implemented with L_2 regularization. The stopping criterion used for training was 100 epochs.

The RF model is an ensemble learning technique that utilizes multiple decision trees to improve predictive accuracy and control overfitting. In this method, several datasets are built using the bagging technique, and the predictors are randomly chosen for each decision tree. For this study, the random forest model was configured with the following hyperparameters: number of trees: 100; node impurity function: Gini index; maximum tree depth: none (nodes are expanded until all leaves are pure or contain fewer than the minimum samples required to split); minimum samples split: 2; minimum samples per leaf: 1.

The MLP model used in this study, designed to capture complex relationships in the data, used two hidden layers. According to Ismayilova et al. [24], two hidden layer neural networks can precisely represent continuous, discontinuous bounded and all unbounded multivariate functions. The MLP architecture shown in Figure 3 consists of an input layer, followed by two hidden neuron layers, with 128 and 64 neurons, respectively, all utilizing the ReLU activation function. To prevent overfitting, dropout layers with a dropout rate 0.2 were introduced after the first and second hidden layers. The final layer is a classification layer with a single neuron using a sigmoid activation function for binary classification.

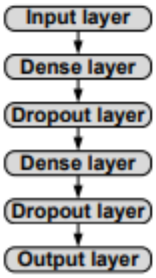


Fig. 3. Multilayer perceptron architecture used for predicting 14-day mortality in patients with TBI

This study employed two CNN architectures to capture patterns within the data. These architectures were designed to extract and refine features from the input data progressively. The first architecture, CNN1, shown in Figure 4(a), features a parallel structure and, like the inception block of the Google Net [25], use filters with different kernels size: 1x2, 1x3, …1x *k*. With a smaller filter, less predicting inputs are used in 1D convolution. With a large filter, more predicting inputs are used in 1D convolution. This variation in kernel size makes it possible to integrate contributions from different predicting variable sets into the final prediction. We evaluated several *k* values, but the best results were achieved with *k=4*. The parallel outputs are concatenated and passed through a dense layer, with 50 neurons and ReLU activation, followed by a dropout layer with a rate of 0.2 to mitigate overfitting, improving the CNN generalization. Finally, the output layer uses a sigmoid activation function for binary classification.

The second architecture, CNN2, shown in Figure 4(b), features a sequential structure comprising 1D convolutional blocks of characteristic extraction. Each extraction block is comprised of a 1D convolutional layer, followed by batch normalization and ReLU activation. The convolution operations maintain the size of the input representation. We evaluated several *k* values, but the best results were achieved with *k=2*. After the last characteristic extraction block, the signal passes through a dense layer, with 50 neurons and ReLU activation, followed by a dropout layer with a rate of 0.2, to mitigate overfitting, improving the CNN generalization. Finally, the output layer uses a sigmoid activation function for binary classification.

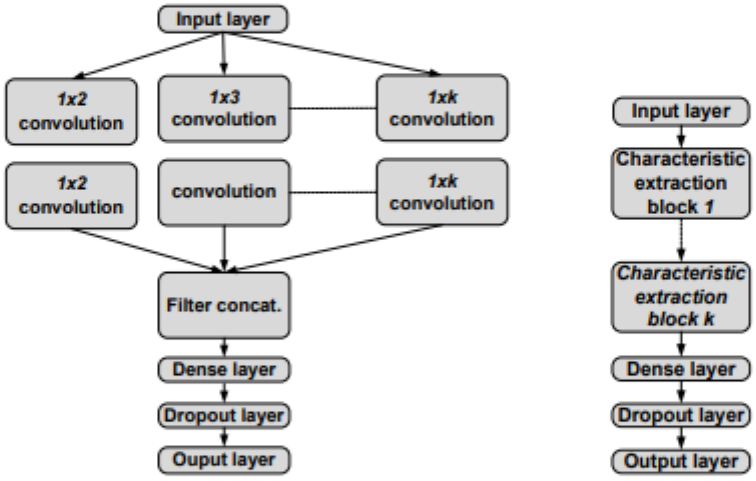


Fig. 4. CNN architectures used for predicting 14-day mortality in patients with TBI. (a) CNN1 with parallel architecture; (b) CNN2 with sequential architecture.

4.2.3 Adjusting hyperparameters

The MLP and CNNs training were configured with a set of hyperparameters, obtained through experiments, that optimized the prediction performance: learning rate: 1e-2 for MLP and 1e-3 for CNNs; optimizers: root mean square propagation (RMSProp), adaptive moment (Adam) and stochastic gradient descent with moment (SGDM); number of training epochs: 300; callbacks for saving the best result in training, using the validation accuracy. During the training, the learning rate was progressively reduced, until reaches a value of 1e-6.

4.2.4 Defining training and testing strategies

As shown in Figure 5, the goal of the strategies defined in this study was to assess the predictive power of classical and deep machine learning models on 14-day mortality, across various

configurations and data combinations. In strategies 1 and 2, the models are trained and tested separately with each dataset's training and test sets. The aim is to verify the predictive power on each dataset, separately. In strategies 3 and 4, the models are trained with the training set from one dataset and tested with the test set from the other dataset. The aim is to assess the generalization power of the models trained on different datasets. In strategy 5, the models are trained and tested with both the training and test sets from both datasets. The aim is to assess whether a model can capture the peculiarities of each dataset and performs well. Strategies 1, 3, 4 and 5 use only the 15 input variables common to both datasets and shown in Table 2. Strategy 2 uses 15, 16 and 17 variables present in Manaus dataset.

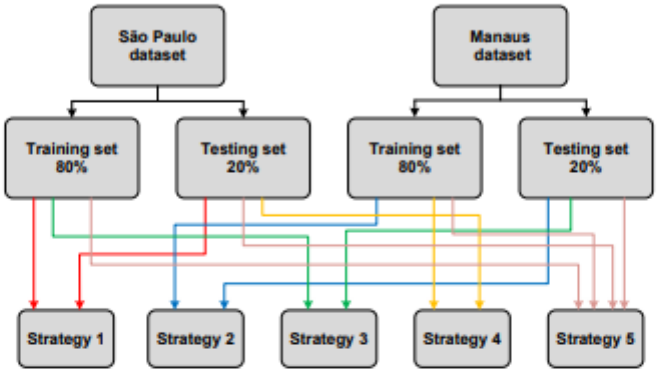


Fig. 5. Strategies adopted for training and testing

4.2.5 Metrics for evaluating the results

The performances of the models were evaluated using the following metrics: accuracy, F1-score, and the area under the ROC curve (AUC). The accuracy of a classification system is the degree of closeness of the classification to its actual value. The F1-score is the harmonic mean of the precision and recall, symmetrically representing both metrics. The AUC demonstrates the trade-off between sensitivity and specificity for different cutoff points in the estimated probability. The higher the AUC value, the greater the discriminatory power of a model [26][27].

In this work, a positive classification indicates the patient's death within 14 days, and a negative classification indicates the patient's survival in 14 days. The accuracy and F1-score are described as follows:

$$accuracy = (TP + TN)/(TP + TN + FP + FN) \quad (1)$$

$$F1\text{-}score=2*(Precision*Recall)/(Precision+Recall) \quad (2)$$

Where: TP – True Positive; FP – False Positive; TN – True Negative and FN – False Negative.

5. Results

In this section, we will present the results for strategies 1 to 5 using the metrics presented in the last section, and confusion matrix tables. To explain the results, we will show tables with Pearson's coefficient of the correlation between each variable and the 14-day mortality, and values obtained by the SHapley Additive exPlanations (SHAP) technique [28]. (SHAP) is a game theory-based method for explaining the output of machine learning models. SHAP uses Shapley values to assign credit to each feature or feature value for a model's prediction. Finally, from the SHAP values, we obtain a set with 5 variables with the most significant positive impact on predicting mortality in 14 days, and evaluate the performance of CNN1 predictor in strategies 1, 2 and 5.

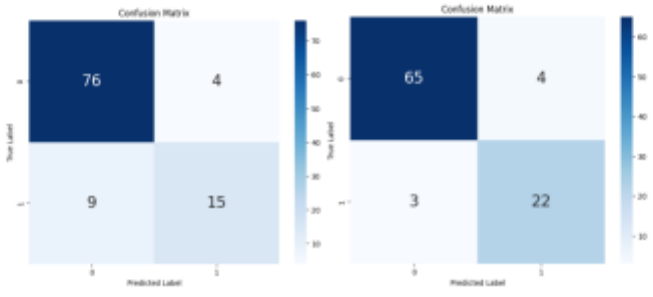
5.1 Results for strategies 1, with 15 input variables, and for strategy 2 with, 15, 16 and 17 input variables

The performance metrics accuracy, F1-score, and AUC, calculated for strategies 1 (São Paulo dataset) and 2 (Manaus dataset), with 15 input variables, are shown in Table 3. As shown, the CNN models achieved the best performances on both datasets. In contrast, the Logistic Regression model, being a linear model, performed worse. All the metric values for strategy 2 are better than those obtained for strategy 1. In particular, the AUC for the Manaus dataset is 0.93, while for the São Paulo dataset is 0.91. These values were obtained with CNN1 and CNN2, respectively, using the RMSProp optimizer shows the best performance in both strategies. The confusion matrices for both strategies, obtained with the best predictors are shown in Figure 6. The sensitivity and specificity for strategy 1 are 0.79 and 0.89, respectively. For strategy 2, 0.84 and 0.96, respectively.

To explore the impact of the additional variables available in the Manaus dataset, we conducted a series of experiments by incrementally adding the extra variables (pandemic and time trauma admission) to the 15 shared variables common to both datasets. The pandemic variable provides important context regarding whether the data was collected during the COVID-19 pandemic. The time trauma admission variable captures the time between the trauma and hospital admission. Table 4 shows the results obtained for strategy 2 with these two new extra variables, using the best predictor in Manaus dataset. The goal was to determine whether these extra variables could improve the model's predictive power. As shown, when using the 17 variables, the AUC increases to 0.98. Figure 7 shows, for strategy 2, with 17 variables, the confusion matrix obtained with the best model. The sensitivity and specificity are 0.92 and 0.99, respectively.

Table 3 Performance metrics for strategies 1 and 2 with 15 input variables

Machine learning model	Optimizer	Strategy 1			Strategy 2		
		Accuracy	F1-Score	AUC	Accuracy	F1-Score	AUC
Logistic Regression	-	0.81	0.79	0.83	0.81	0.77	0.79
Random Forest	-	0.81	0.80	0.83	0.82	0.79	0.81
MLP	Adam	0.79	0.76	0.80	0.89	0.88	0.90
MLP	RMSprop	0.80	0.78	0.81	0.88	0.86	0.89
MLP	SGDM	0.80	0.78	0.82	0.87	0.85	0.89
CNN1	Adam	0.81	0.80	0.83	0.90	0.88	0.91
CNN1	RMSprop	0.82	0.81	0.84	0.92	0.90	0.93
CNN1	SGDM	0.80	0.78	0.81	0.88	0.86	0.89
CNN2	Adam	0.86	0.83	0.89	0.89	0.87	0.90
CNN2	RMSprop	0.87	0.85	0.90	0.90	0.89	0.91
CNN2	SGDM	0.85	0.82	0.87	0.87	0.85	0.89



(a) (b)

Fig. 6. Confusion matrices for both strategies: (a) using the CNN2 with RMSProp optimizer for strategy 1; (b) using CNN1 with RMSProp optimizer for strategy 2.

Table 4 Results for strategy 2, with 15, 16 and 17 variables, with the best model

Variables	Accuracy	F1-Score	AUC
15 variables	0.92	0.90	0.93
15 variables + pandemic variable	0.95	0.94	0.97
15 variables+ time trauma admission variable	0.95	0.96	0.97
15 variables + pandemic and time trauma admission variables	0.97	0.96	0.98

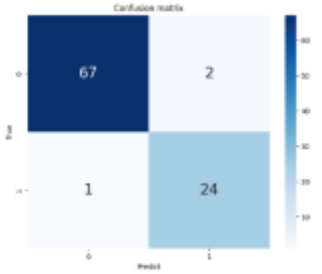


Fig. 7. Confusion matrices for strategy 2 with 17 input variables, using the model CNN1 with RMSProp optimizer.

5.2 Results for Strategies 3 and 4 with 15 input variables

The results for strategies 3 and 4, are shown in Table 5. Compared with the results of strategies 1 and 2 shown in Table 3, the models performances decreased. For example, the best AUC for strategies 3 and 4 are also obtained with CNN1 and RMSProp optimizer, 0.77 and 0.70, respectively. These values are lower than those obtained with strategies 1 and 2, 0.90 and 0.93, respectively. We also note that training with the Manaus dataset and testing with data from the São Paulo dataset (strategy 3) results in better metric values than training with the São Paulo dataset and testing with the Manaus dataset.

5.3 Results for Strategy 5 with 15 input variables

The performance metrics calculated for strategy 5 are shown in Table 5. Compared with the results of strategies 1 and 2, shown in Table 3, the models performances decreased. For example, the best AUC for strategy 5 is also obtained with CNN1 and RMSProp optimizer, 0.83. This value is lower than those obtained with strategies 1 and 2, 0.90 and 0.93, respectively, but better than the value obtained with strategies 3 and 4. Figure 8 shows the confusion matrix obtained with the best model for strategy 5.

5.4 Explaining Results

5.4.1 Pearson's coefficient analysis

To understand the correlation between predictor variables and 14-day mortality in both datasets, we used Pearson's correlation coefficient. Pearson's correlation measures the strength and direction of the relationship between predictor variables and 14-day mortality, helping to identify critical factors that may contribute to patient outcomes. Table 7 shows Pearson's correlation coefficients to the São Paulo dataset, while Table 8 shows Pearson's correlation coefficients to the Manaus dataset.

For the São Paulo dataset, variables such as *pupil reactivity*, *motor score*, *midline shift*, *age* and p-*prothrombin time* showed the highest absolute correlation values with 14-day mortality. *Pupil reactivity* and *motor score* showed a negative correlation with survival outcomes, while *midline shift age* and *prothrombin time* showed a positive correlation.

For the Manaus dataset, *motor score, pupil reactivity, hypoxia, midline shift and hypotension* were among the most strongly correlated variables. As in the São Paulo dataset, m*otor score* and *pupil reactivity* emerged as critical predictors, showing a robust negative correlation with 14-day mortality, -0.654 and -0.588, respectively. Additionally, the variable *pandemic* and *time trauma admission* showed lower values of Person's coefficients.

Comparing the absolute values of Pearson's correlation coefficient between the Manaus dataset and the São Paulo dataset, we verified that, in the Manaus dataset, the five variables cited before have absolute values above 0.3. In contrast, in the São Paulo dataset, only one variable has an absolute value above 0.3, *pupil reactivity*.

5.4.2 SHAP values analysis

To gain insight into the importance of each feature and its impact on model predictions, we used SHAP (SHapley Additive exPlanations) values for the best model. SHAP analysis helps explain the contribution of each variable to the prediction, providing transparency to the model's decision-making process. Figure 8 shows the SHAP plots for both datasets. In each plot, the Y-axis indicates the feature names in order of importance from top to bottom. The X-axis represents the SHAP values, which means the degree of change in log odds. The color of each point on the graph represents the value of the corresponding feature, with red indicating high values and blue indicating low values. Each point represents a row of data from the original dataset.

Table 5 Results for strategies 3 and 4, with 15 input variables

Machine learning model	Optimizer	Strategy 3			Strategy 4		
		Accuracy	F1-Score	AUC	Accuracy	F1-Score	AUC
LR	-	0.30	0.49	0.51	0.60	0.45	0.61
RF	-	0.33	0.51	0.50	0.64	0.47	0.67
MLP	Adam	0.36	0.52	0.50	0.71	0.50	0.68
MLP	RMSprop	0.36	0.52	0.51	0.70	0.49	0.68
MLP	SGDM	0.35	0.51	0.53	0.69	0.48	0.67
CNN1	Adam	0.38	0.53	0.52	0.73	0.50	0.69
CNN1	RMSprop	0.27	0.42	0.52	0.77	0.52	0.70
CNN1	SGDM	0.35	0.51	0.51	0.71	0.49	0.68
CNN2	Adam	0.37	0.53	0.52	0.75	0.51	0.70
CNN2	RMSprop	0.36	0.52	0.52	0.74	0.51	0.69
CNN2	SGDM	0.35	0.51	0.50	0.73	0.50	0.69

Table 6 Results for strategy 5 with 15 input variables

Model	Optimizer	Accuracy	F1-Score	AUC
LR	-	0.90	0.61	0.72
RF	-	0.80	0.57	0.70
MLP	Adam	0.81	0.62	0.72
MLP	RMSprop	0.81	0.62	0.73
MLP	SGDM	0.71	0.12	0.53
CNN1	Adam	0.81	0.63	0.73
CNN1	RMSprop	0.83	0.69	0.77
CNN1	SGDM	0.70	0.09	0.52
CNN2	Adam	0.81	0.62	0.72
CNN2	RMSprop	0.80	0.61	0.72

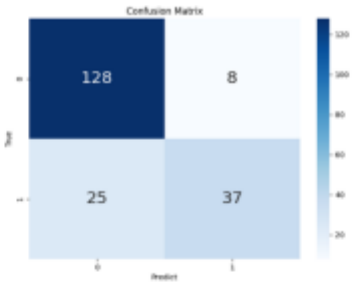


Fig. 8. Confusion matrix for strategy 5 using the best model, CNN1 with RMSProp optimizer, and 15 variables.

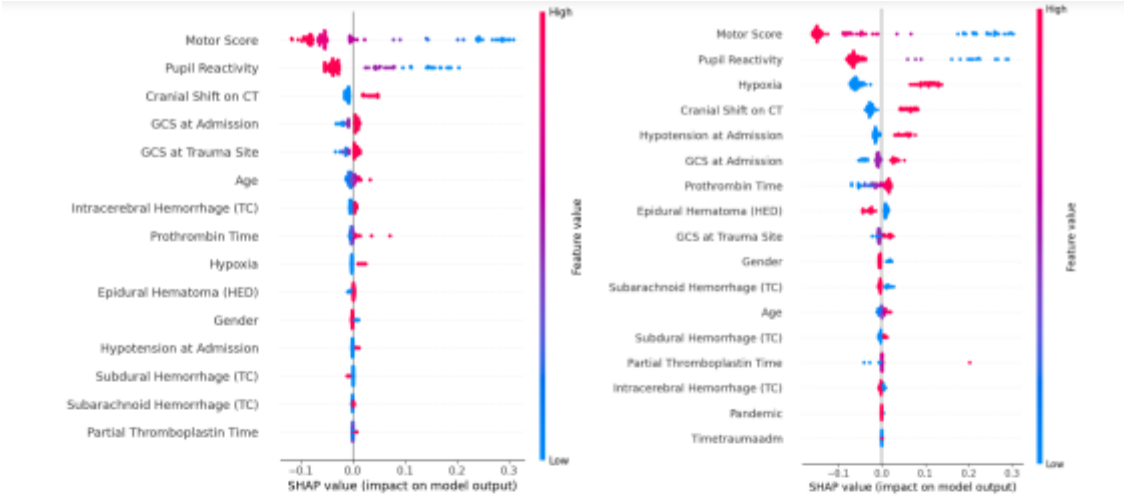
As shown in the SHAP plot for the São Paulo dataset, Figures 9(a), the 5 most essential variables in model prediction include *motor score*, *pupil reactivity*, *midline shift*, *GCS at admission* and *GCS at trauma site*. As shown in the SHAP plot for the Manaus dataset, Figure 9(b), the 5 most essential variables in model prediction include *motor score*, *pupil reactivity, hypoxia, midline shift and hypotension at admission*.

Table 7 Pearson's correlation coefficients to the São Paulo dataset

Predictor Variable	Pearson Coefficient
Sex	-0.122
Age	0.190
Pupil reactivity	-0.373
GCS at trauma site	0.119
GCS at admission	0.121
Motor score	-0.281
Hypoxia	0.107
Hypotension	0.140
Midline shift	0.219
Subarachnoid hemorrhage	0.059
Epidural hematoma	0.080
Subdural hemorrhage	-0.044
Intracerebral hemorrhage	0.051
Prothrombin time	0.165
Partial thromboplastin time	0.159

Table 8 Pearson's correlation coefficients to the Manaus dataset

Predictor Variable	Pearson Coefficient
Sex	-0.043
Age	-0.088
Pupil reactivity	-0.588
GCS at trauma site	0.268
GCS at admission	0.580
Motor Score	-0.654
Hypoxia	0.458
Hypotension	0.375
Midline shift	0.402
Subarachnoid hemorrhage	0.050
Epidural hematoma	-0.094
Subdural hemorrhage	0.263
Intracerebral hemorrhage	-0.024
Prothrombin time	-0.271
Partial thromboplastin time	-0.064
Pandemic	0.107
Time trauma admission	0.026



(a) (b)

Fig. 9. Shap values for predictions of model CNN1 with RMSprop optimizer. (a) São Paulo dataset, with 15 input variables(b) Manaus dataset with 17 input variables.

For the São Paulo dataset, *motor score* had the most significant positive impact on predicting mortality, while *pupil reactivity* and *GCS at the trauma site* have less importance. In contrast, *midline shift on CT* had a relatively lower impact. In the Manaus dataset, *motor score* remained the most critical variable, but *pupil reactivity* and *hypoxia* also played significant roles in predictions. Additionally, *hypotension* and *gender* had more influence on predictions in the Manaus dataset than in the São Paulo dataset.

Comparing Pearson's analysis with the SHAP analysis, we notice a good agreement for the Manaus dataset. The 5 variables with the highest absolute values for the Pearson coefficient correspond to the 5 variables with the highest SHAP values. For the São Paulo dataset, this coincidence of highest values is only for 3 variables. Table 8 shows the results of training the best model with strategies 1, 2 and 5, using the sets of 5 most significant impact variables on SHAP analysis, for both datasets.

Table 9 shows the results for the São Paulo and Manaus databases, using the 5 best variables mentioned before and the best predictor for each of them

Table 9 Results for São Paulo and Manaus datasets, with best predictors and 5 best variables for each of them

Dataset/ Variables	Accuracy	F1-Score	AUC
São Paulo / motor score, pupil reactivity, midline shift, GCS at admission and GCS at trauma site	0.86	0.61	0.87
Manaus/ motor score, pupil reactivity, hypoxia, midline shift and hypotension at admission	0.92	0.83	0.97

As shown in Table 9, for São Paulo dataset, the AUC obtained using only 5 best variables for the São Paulo, 0.87, is a little smaller than the AUC obtained with 15 variables, 0.91 (Table 3). For Manaus dataset, Tabel 9 shows that the AUC obtained only with 5 best variables 0.97, is better than the AUC obtained with 15 variables, 0.93 (Table 3), but worse than the AUC obtained with 17 variables, 0.98 (Table 4).

6. Discussion

This study aimed to assess the performance and transferability of machine learning models for predicting 14-day mortality in TBI patients across two distinct Brazilian regions. Our findings indicate

that models trained in one region faced challenges in maintaining their predictive power when applied to data from the other. This suggests that regional factors significantly shape TBI patient outcomes, a conclusion supported by previous research that highlights the variability in healthcare access and infrastructure across different geographical settings Warman et al. [29], Amorim et al. [2] Notably, deep models, using CNN-based models—specifically CNN1 and CNN2 optimized via the RMSprop method—demonstrated high performance across both datasets, when compared to classic machine learning methods. The model performed well in each separated dataset, with São Paulo achieving an area under the curve (AUC) of 0.90 and Manaus showing an AUC of 0.93, both for the best model (CNN1 with RMSProp). This performance was anticipated and aligns with previous findings from our group when studying the São Paulo dataset [20][2].

The incorporation of context-specific features, particularly those related to the COVID-19 pandemic and the time from trauma to admission, significantly enhanced model accuracy, with the Manaus model achieving an impressive AUC of 0.98 upon their inclusion. Manaus, as one of the epicenters of the COVID-19 pandemic, faced unique challenges that influenced healthcare delivery and patient outcomes during this period [30]. Therefore, the pandemic may exacerbated issues related to healthcare access, particularly for individuals residing in rural areas where road infrastructure is limited [3]. This finding aligns with Zimmerman study [6], that emphasizes the role of localized variables in improving predictive models, especially in low-resource environments. The addition of these features not only improved the models' performance but also underscored the necessity of tailoring predictors to enhance inter-regional applicability. Such adaptations are crucial, particularly when considering the differences in healthcare access and infrastructure between urban centers like São Paulo and more isolated cities as Manaus [2][31].

The inclusion of the COVID-19 pandemic variable in the Manaus dataset provided valuable contextual information, reflecting both systemic strain and changes in trauma patterns during that period. Manaus faced a severe collapse in its healthcare infrastructure, including overcrowded ICUs and delayed access to neurosurgical care. These conditions likely contributed to higher early mortality among TBI patients. At the same time, lockdown measures may have reduced certain trauma mechanisms, such as road traffic accidents. The significant improvement in model performance with the addition of the pandemic variable suggests that context-specific variables can meaningfully enhance predictive accuracy, especially in regions experiencing extreme healthcare fluctuations.

When comparing the most important variables associated with outcomes, we found that motor score, pupil reactivity, and midline shift were significant predictors in both models. Motor score and pupil reactivity, which consistently emerged as top predictors, reflect core aspects of neurological function and injury severity—namely, brainstem integrity and intracranial pressure dynamics. Midline shift on CT is a radiological indicator of mass effect and potential herniation, directly associated with increased mortality risk. While age was a critical factor in São Paulo, hypoxia and hypotension emerged as more important predictors in Manaus. These physiological derangements are known to exacerbate primary brain injury through mechanisms such as reduced cerebral perfusion and impaired autoregulation. Together, these findings align with established TBI pathophysiology and highlight the importance of integrating clinically meaningful variables into predictive modeling, especially in heterogeneous and resource-limited environments. Moreover, this observation is consistent with existing literature, which indicates that these variables are highly correlated with outcomes and are present in both the CRASH and IMPACT models [32][33]. In the IMPACT model, midline shift (MLS) is implicitly represented through the Marshall Classification, while the CRASH model does not explicitly include motor score, as it is incorporated within the Glasgow Coma Scale (GCS). Notably, hypoxia and hypotension are absent from the CRASH model, which may limit its applicability in settings where these factors are prevalent. Several authors argue that higher rates of hypoxia and hypotension correlate with poorer pre-hospital assessments [34] [35]. In the Manaus cohort, 25.9% of patients exhibited hypotension, and 21.1% presented with hypoxia, with these conditions occurring more frequently in severe TBI patients (55.3% and 43.8%, respectively) - the baseline data of the Manaus cohort will be published in another paper. In contrast, in São Paulo, hypotension and hypoxia were observed in 21.2% and 13.2%

of patients, respectively, with rates in severe TBI patients being 25.3% and 14.8% [2]. These findings depict the differences in pre-hospital management between the two settings and highlight the importance of addressing these variables in the Manaus cohort.

However, the interpretation of these findings must be approached with caution due to the inherent variability within the datasets. Merchant et al. [36] consistently underscored the complexity of creating effective prediction through trauma scoring systems in trauma patients in a LMIC. While our models and comparisons provided a broad evaluation, it also complicated the identification of a single "best" model even in a unique country due to its inherent inequalities, especially in LMIC. However, the use of ML models shed a light to science to potentially create faster personalized predictions, considering diverse factors that is hidden such as socioeconomic factors and previous mental status, for instance. Collecting data is crucial and is more challenging in LMIC. Recently, Tritt et al. [37] demonstrated that 19 clusters of TBI outcomes can be predicted from intake data, a ~ 6× improvement in precision over clinical standards. They used supervised and unsupervised approaches to identify the clusters. In our study, deep models, with CNN models showed better performance than simpler logistic regression models, likely due to their inability to capture the intricate relationships within the data. This pattern is consistent with previous research indicating that advanced algorithms generally outperform traditional methods, particularly in complex datasets [29] [34].

Table 4 shows that the experiment carried out with only the 5 best variables in each dataset showed a very relevant result. For the São Paulo dataset, the AUC was only slightly lower than that obtained with 15 variables. For the Manaus database, the AUC was better than that obtained with 15 variables. Increasing the complexity to 15 and 17 variables, implied a small decrease in performance and a small gain in performance, respectively. The main reason for this is the small size of both databases. The higher the ratio between the number of training patterns and the number of classifier parameters, the better the generalization properties of the neural network [23]. Many input variables are directly translated into a large number of predictor parameters. Thus, for a finite and usually limited number of N training patterns, keeping the number of variables as small as possible is in line with our desire to design predictors with generalization capabilities.

From a medical point of view, good prediction values obtained with a reduced set of variables are essential because it is very difficult to record many variables for patients with TBI in emergency medical settings.

The findings from this study represent a significant contribution to the broader discourse on TBI research, as it marks, to the best of our knowledge, the first tentative cross-validation of predictive models across two datasets from the same country, Brazil. The stark differences in healthcare access and infrastructure between São Paulo and Manaus likely account for the observed variations in model performance. These disparities highlight the necessity for models trained in better-equipped settings to undergo substantial adjustments to function effectively in regions with limited resources [2]. By illustrating these regional disparities, our study underscores the importance of localized approaches in the development and validation of predictive models for TBI outcomes in continental countries.

In terms of generalizability, our findings offer insights into the broader applicability of mortality prediction models for TBI patients across differing healthcare environments. Although performance varied by region, the incorporation of locally relevant variables led to more accurate predictions, particularly for the Manaus dataset. This indicates that similar adaptations may benefit models intended for other low-resource environments, although broader application could still be constrained by region-specific factors, such as healthcare access, transportation logistics, and available resources [29][2].

This study has several notable limitations. First, although both datasets were prospectively collected, reliance on retrospective analysis can introduce bias, particularly in handling missing data, where imputation methods were employed but may not fully capture certain unmeasured variables. Furthermore, the focus on only two regions, while beneficial for direct comparison, limits the study's relevance to other regions with distinct healthcare systems. Additionally, potential confounding factors, such as transfer times and pandemic-related variables, may have influenced mortality predictions, particularly in the Manaus dataset, thereby reducing external validity.

7. Conclusion

This study aimed to assess the performance and transferability of machine learning models for predicting 14-day mortality in TBI patients across two distinct Brazilian regions, São Paulo and Manaus. Deep models, using CNN-based models demonstrated high performance across both datasets, when compared to classic machine learning methods. The top-performing model demonstrated strong results across both datasets, achieving an AUC of 0.90 for São Paulo and an impressive AUC of 0.98 for Manaus. However, our findings indicate that models trained in one region faced challenges in maintaining their predictive power when applied to data from the other.

Author Contributions: Conceptualization, M.G.F.C.; methodology, F.A.S.A., C.F.F.C.F., M.G.F.C. and R.L.O.A.; software, F.A.S.A.; formal analysis, F.A.S.A., C.F.F.C.F., and M.G.F.C.; investigation, R.L.O.A., F.A.S.A., H.O.M., M.G.F.C. and C.F.F.C.F.; data curation, M.G.F.C, H.O.M.; writing—original draft preparation, C.F.F.C.F., M.G.F.C., F.A.S.A. and R.L.O.A.; writing—review and editing, C.F.F.C.F.; visualization, M.G.F.C.; supervision, C.F.F.C.F., R.L.O.A. and M.G.F.C.; project administration, R.L.O.A. and C.F.F.C.F.; funding acquisition, M.G.F.C. All authors have read and agreed to the published version of the manuscript.

References

[1] M. C. Dewan *et al.*, "Estimating the global incidence of traumatic brain injury.," *J. Neurosurg.*, vol. 130, no. 4, pp. 1080–1097, Apr. 2019, doi: 10.3171/2017.10.JNS17352.

[2] R. L. Amorim *et al.*, "Prediction of Early TBI Mortality Using a Machine Learning Approach in a LMIC Population.," *Front. Neurol.*, vol. 10, p. 1366, 2019, doi: 10.3389/fneur.2019.01366.

[3] P. C. Nôvo, S. A. B. de Farias, V. do V. Guttemberg, V. R. Félix Dos Santos, J. P. Moreira Guilherme, and R. L. O. de Amorim, "Neurosurgical Emergencies in the Amazon: An Epidemiologic Study of Patients Referred by Air Transport for Neurosurgical Evaluation at a Referral Center in Amazonas.," *World Neurosurg.*, vol. 173, pp. e359–e363, May 2023, doi: 10.1016/j.wneu.2023.02.056.

[4] R. Raj *et al.*, "Dynamic prediction of mortality after traumatic brain injury using a machine learning algorithm," *npj Digit. Med.*, vol. 5, no. 1, p. 96, 2022, doi: 10.1038/s41746-022-00652-3.

[5] K. C. Tu *et al.*, "A Computer-Assisted System for Early Mortality Risk Prediction in Patients with Traumatic Brain Injury Using Artificial Intelligence Algorithms in Emergency Room Triage," *Brain Sciences*, vol. 12, no. 5. 2022, doi: 10.3390/brainsci12050612.

[6] A. Zimmerman *et al.*, "Machine learning models to predict traumatic brain injury outcomes in Tanzania: Using delays to emergency care as predictors.," *PLOS Glob. public Heal.*, vol. 3, no. 10, p. e0002156, 2023, doi: 10.1371/journal.pgph.0002156.

[7] J. T. Senders *et al.*, "Machine Learning and Neurosurgical Outcome Prediction: A Systematic Review.," *World Neurosurg.*, vol. 109, pp. 476-486.e1, Jan. 2018, doi: 10.1016/j.wneu.2017.09.149.

[8] E. Courville *et al.*, "Machine learning algorithms for predicting outcomes of traumatic brain injury: A systematic review and meta-analysis.," *Surg. Neurol. Int.*, vol. 14, p. 262, 2023, doi: 10.25259/SNI_312_2023.

[9] A. Cerasa *et al.*, "Predicting Outcome in Patients with Brain Injury: Differences between Machine Learning versus Conventional Statistics," *Biomedicines*, vol. 10, no. 9. 2022, doi: 10.3390/biomedicines10092267.

[10] J. Wang, M. J. Yin, and H. C. Wen, "Prediction performance of the machine learning model in predicting mortality risk in patients with traumatic brain injuries: a systematic review and meta-analysis," *BMC Medical Informatics and Decision Making*, vol. 23, no. 1. 2023, doi: 10.1186/s12911-023-02247-8.

[11] S. Der Hsu, E. Chao, S. J. Chen, D. Y. Hueng, H. Y. Lan, and H. H. Chiang, "Machine learning algorithms to predict in-hospital mortality in patients with traumatic brain injury," *J. Pers. Med.*, vol. 11, no. 11, 2021, doi: 10.3390/jpm11111144.
[12] K. Ding *et al.*, "Mobile telephone follow-up assessment of postdischarge death and disability due to trauma in Cameroon: a prospective cohort study," *BMJ open*, vol. 12, no. 4. p. e056433, 2022, doi: 10.1136/bmjopen-2021-056433.
[13] J. Fonseca, X. Liu, H. P. Oliveira, and T. Pereira, "Learning Models for Traumatic Brain Injury Mortality Prediction on Pediatric Electronic Health Records," *Front. Neurol.*, vol. 13, no. June, pp. 1–11, 2022, doi: 10.3389/fneur.2022.859068.
[14] M. Rodrigues de Souza *et al.*, "Evaluation of Computed Tomography Scoring Systems in the Prediction of Short-Term Mortality in Traumatic Brain Injury Patients from a Low- to Middle-Income Country.," *Neurotrauma reports*, vol. 3, no. 1, pp. 168–177, 2022, doi: 10.1089/neur.2021.0067.
[15] E. Courville *et al.*, "Machine learning algorithms for predicting outcomes of traumatic brain injury: A systematic review and meta-analysis," *Surgical Neurology International*, vol. 14. 2023, doi: 10.25259/SNI_312_2023.
[16] A. Kashkoush, J. C. Petit, H. Ladhani, V. P. Ho, and M. L. Kelly, "Predictors of Mortality, Withdrawal of Life-Sustaining Measures, and Discharge Disposition in Octogenarians with Subdural Hematomas.pdf," *World Neurosurgerysurg.*, vol. 157, no. January, pp. e179–e187, 2022, doi: doi:10.1016/j.wneu.2021.09.121.
[17] A. Mekkodathil, A. El-Menyar, M. Naduvilekandy, S. Rizoli, and H. Al-Thani, "Machine Learning Approach for the Prediction of In-Hospital Mortality in Traumatic Brain Injury Using Bio-Clinical Markers at Presentation to the Emergency Department," *Diagnostics*, vol. 13, no. 15. 2023, doi: 10.3390/diagnostics13152605.
[18] Y. Cao, M. P. Forssten, B. Sarani, S. Montgomery, and S. Mohseni, "Development and Validation of an XGBoost-Algorithm-Powered Survival Model for Predicting In-Hospital Mortality Based on 545,388 Isolated Severe Traumatic Brain Injury Patients from the TQIP Database," *J. Pers. Med.*, vol. 13, no. 9, 2023, doi: 10.3390/jpm13091401.
[19] M. C. P. Campos, R. Venzel, L. P. de Oliveira, F. Reis, and R. L. O. de Amorim, "Management of Traumatic Brain Injury at a Medium Complexity Hospital in a Remote Area of Amazonas, 2017–2019," *World Neurosurg.*, vol. 148, pp. 151–154, 2021, doi: 10.1016/j.wneu.2020.12.088.
[20] K. A. A. Guimarães, R. L. O. de Amorim, M. G. F. Costa, and C. F. F. Costa Filho, "Predicting early traumatic brain injury mortality with 1D convolutional neural networks and conventional machine learning techniques," *Informatics Med. Unlocked*, vol. 31, pp. 1–23, 2022, doi: 10.1016/j.imu.2022.100984.
[21] Charu C. Aggarwal, *Neural Networks and Deep Learning*. 2023.
[22] F. Chollet, *Deep Learning with Python*. 2021.
[23] S. Theodoridis and K. Koutroumbas, *Pattern Regconition*. 2009.
[24] A. Ismayilova and V. E. Ismailov, "On the Kolmogorov neural networks," *Neural Networks*, vol. 176, pp. 1–14, 2024, doi: 10.1016/j.neunet.2024.106333.
[25] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the Inception Architecture for Computer Vision," *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, vol. 2016-Decem, pp. 2818–2826, 2016, doi: 10.1109/CVPR.2016.308.
[26] T. Fawcett, "An introduction to ROC analysis," *Pattern Recognit. Lett.*, vol. 27, no. 8, pp. 861–874, 2006, doi: 10.1016/j.patrec.2005.10.010.
[27] D. M. W. Powers, "Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation," no. May, 2020, [Online]. Available: http://arxiv.org/abs/2010.16061.
[28] S. M. Lundberg and S. I. Lee, "A unified approach to interpreting model predictions," *Adv. Neural Inf. Process. Syst.*, vol. 2017-Decem, no. Section 2, pp. 4766–4775, 2017.
[29] P. I. Warman *et al.*, "Machine Learning for Predicting In-Hospital Mortality After Traumatic

Brain Injury in Both High-Income and Low- and Middle-Income Countries.," *Neurosurgery*, vol. 90, no. 5, pp. 605–612, May 2022, doi: 10.1227/neu.0000000000001898.

[30] E. Fernandes *et al.*, "Exploring Prehospital Data for Pandemic Preparedness: A Western Brazilian Amazon Case Study on COVID-19.," *Int. J. Environ. Res. Public Health*, vol. 21, no. 9, Sep. 2024, doi: 10.3390/ijerph21091229.

[31] M. Rodrigues De Souza *et al.*, "Evaluation of Computed Tomography Scoring Systems in the Prediction of Short-Term Mortality in Traumatic Brain Injury Patients from a Low- to Middle-Income Country," *Neurotrauma Reports*, vol. 3, no. 1. pp. 168–177, 2022, doi: 10.1089/neur.2021.0067.

[32] A. Faried, F. C. Satriawan, and M. Z. Arifin, "Feasibility of Online Traumatic Brain Injury Prognostic Corticosteroids Randomisation After Significant Head Injury (CRASH) Model as a Predictor of Mortality.," *World Neurosurg.*, vol. 116, pp. e239–e245, Aug. 2018, doi: 10.1016/j.wneu.2018.04.180.

[33] E. W. Steyerberg *et al.*, "Predicting outcome after traumatic brain injury: Development and international validation of prognostic scores based on admission characteristics," *PLoS Med.*, vol. 5, no. 8, pp. 1251–1261, 2008, doi: 10.1371/journal.pmed.0050165.

[34] A. Abujaber, A. Fadlalla, D. Gammoh, H. Abdelrahman, M. Mollazehi, and A. El-Menyar, "Prediction of in-hospital mortality in patients with post traumatic brain injury using National Trauma Registry and Machine Learning Approach.," *Scand. J. Trauma. Resusc. Emerg. Med.*, vol. 28, no. 1, p. 44, May 2020, doi: 10.1186/s13049-020-00738-5.

[35] J. R. Huie, C. A. Almeida, and A. R. Ferguson, "Neurotrauma as a big-data problem," *Curr. Opin. Neurol.*, vol. 31, no. 6, 2018, [Online]. Available: https://journals.lww.com/co-neurology/fulltext/2018/12000/neurotrauma_as_a_big_data_problem.7.aspx.

[36] A. A. H. Merchant *et al.*, "Which curve is better? A comparative analysis of trauma scoring systems in a South Asian country.," *Trauma Surg. acute care open*, vol. 8, no. 1, p. e001171, 2023, doi: 10.1136/tsaco-2023-001171.

[37] A. Tritt *et al.*, "Data-driven distillation and precision prognosis in traumatic brain injury with interpretable machine learning.," *Sci. Rep.*, vol. 13, no. 1, p. 21200, Dec. 2023, doi: 10.1038/s41598-023-48054-z.