An Approach to HLA Allele Imputation in Bone Marrow Donor Registries

Felipe S. C. Eduardo[1], **Nathalia de Azevedo**[1], **Luís Cristóvão M. S. Pôrto**[2], **Karla Figueiredo**[1], **Alexandre C. Sena**[1]

[1]Instituto de Matemática e Estatística – Universidade do Estado do Rio de Janeiro (UERJ)
Rio de Janeiro – RJ – Brazil

[2]Laboratório de Histocompatibilidade e Criopreservação
Universidade do Estado do Rio de Janeiro (UERJ)
Rio de Janeiro – RJ – Brazil

{asena,karlafigueiredo}@ime.uerj.br, lcporto@uerj.br

Abstract. *The main information in bone marrow donor records is the alleles of the HLA genes. Due to the costs and types of tests required to obtain this information, many of these alleles are not found in the databases. Thus, the objective of this study is to evaluate, in an unprecedented way, the possibility of imputing the alleles of genes not reported in these databases. For this purpose, a Recurrent Neural Network of the Long-Short Time Memory (LSTM) type was used. The accuracy of 76% achieved shows the feasibility of imputing the missing alleles, despite the strong imbalance of the classes and because it is one of the most polymorphic regions of human DNA (i.e. many options of distinct alleles).*

Resumo. *As principais informações dos registros de doadores de medula óssea são os alelos dos genes HLA. Em função dos custos e tipos dos exames necessários para se obter essas informações, muitos desses alelos não se encontram nos banco de dados. Assim, o objetivo deste trabalho é, de forma inédita, avaliar a possibilidade de imputar os alelos dos genes não informados nesses bancos de dados. Para mitigar essas lacunas, foram investigados algoritmos baseados em Rede Neural Recorrente do tipo Long-Short Time Memory (LSTM). A acurácia de 76% mostra a viabilidade de imputar os alelos faltantes, apesar do forte desbalanceamento das classes e por se tratar de uma das regiões mais polimórificas do DNA humano (i.e. muitas opções de alelos distintos).*

1. Introdução

O transplante de células-tronco hematopoéticas é, muitas vezes, a única opção de tratamento para pacientes com câncer ou neoplasias hematológicas, onde a compatibilidade entre os alelos HLA do paciente e doador é o principal fator para o sucesso do transplante [Tiercy 2016]. Embora alguns pacientes encontrem doadores compatíveis na família, a maioria depende de doadores voluntários cadastrados em registros como Registro Nacional de Doadores Voluntários de Medula Óssea (REDOME), no Brasil, com mais de 5 milhões de inscritos. A compatibilidade é determinada pela quantidade de alelos compartilhados nos genes HLA-A, -B, -C, -DRB1, -DQB1 e -DPB1 [Geffard et al. 2019], sendo que quanto maior o número de alelos coincidentes, maior a chance de sucesso.

A identificação desses alelos é realizada através da tipagem HLA, um exame laboratorial para determinar os alelos dos genes HLA-A, -B, -C, -DR e -DQ. O custo do exame varia conforme o número de genes analisados e a resolução, que pode ser baixa, média ou alta. Inicialmente, a compatibilidade era determinada apenas pelos genes HLA-A, -B e -DRB1, em baixa resolução. Mais recentemente, com o avanço das técnicas, a análise foi ampliada para incluir também os genes HLA-C, -DQB1 e -DPB1, todos em alta resolução, proporcionando uma compatibilidade mais precisa. Em razão disso, o REDOME possui a maioria dos dados apenas dos genes HLA-A, -B e -DRB1 em baixa resolução. Nesse contexto, seria importante complementar as informações ausentes dos indivíduos, considerando a relevância dos outros alelos e a inviabilidade de refazer os testes em todos os indivíduos. No entanto, até o momento deste estudo, não foi encontrado nenhum trabalho que realize a imputação de alelos em baixa resolução no formato padrão adotado pela Organização Mundial da Saúde (OMS). De modo geral, os métodos de imputação de HLA se dividem entre aqueles que convertem dados de polimorfismos de nucleotídeo único (SNPs, *Single Nucleotide Polymorphisms*) para tipagem padrão e os que convertem alelos de baixa resolução para alta resolução.

Assim, o objetivo deste trabalho é imputar alelos do gene HLA-C, baseado nos alelos dos genes HLA-A, -B e -DRB1, em baixa resolução, utilizando uma Rede Neural Recorrente do tipo Long-Short Time Memory (LSTM). A rede implementada foi capaz de prever com acurácia, 76% dos alelos faltantes da base de teste, o que pode ser considerado um resultado significativo, por ser um problema multiclasses e, também, porque os genes HLA são considerados a região mais polimórfica do DNA humano.

O restante do trabalho está dividido em mais quatro seções. A Seção 2 apresenta brevemente a fundamentação teórica. Por sua vez, a metodologia adotada é descrita na Seção 3. Em seguida, na quarta seção, são apresentados e discutidos os resultados obtidos. Por fim, as conclusões e perspectivas de continuidade do trabalho são descritas na Seção 5.

2. Referencial teórico

Esta seção apresenta a fundamentação teórica deste trabalho. Inicialmente, o Sistema de Antígenos Leucocitários Humanos (HLA) é descrito na Seção 2.1. Em seguida, aspectos teóricos sobre as redes neurais recorrentes LSTM são descritos na Seção 2.2. Por fim, os trabalhos relacionados são apresentados na Seção 2.3.

2.1. Sistema de antígenos leucocitário humano (HLA)

O Complexo Principal de Histocompatibilidade (MHC, sigla em ingles para *Major Histocompatibility Complex*) é uma família multigênica encontrada em um longo trecho de DNA em todos os organismos vertebrados, desempenhando um papel central no sistema imunológico. Em humanos, essa região é tradicionalmente chamada de Antígeno Leucocitário Humano (HLA). Proteínas codificadas por esses genes são expressas na superfície de células nucleadas (ou seja, todas as células, exceto hemácias) e contribuem significativamente para o reconhecimento imunológico [Shaz et al. 2013].

A identificação dos genes de um indivíduo é realizada por meio da tipagem HLA, um processo complexo que utiliza diversas técnicas para determinar os alelos presentes. A tipagem HLA evoluiu de métodos sorológicos para análises moleculares mais complexas, como sequenciamento de nova geração (NGS, do inglês Next-Generation

Sequencing), que permite a leitura detalhada de regiões específicas dos genes HLA, identificando diretamente a sequência de nucleotídeos (A, T, C, G) no material genético [Jeanmougin et al. 2017].

No contexto do REDOME, os dados precisam seguir uma nomenclatura padronizada pela Organização Mundial da Saúde (OMS) para assegurar que as informações sobre os alelos HLA sejam compreendidas e compartilhadas de maneira clara e eficiente entre diferentes bancos de dados de doadores, A nomenclatura dos alelos HLA é organizada em níveis de resolução que refletem o grau de detalhe na identificação molecular. Na baixa resolução, a tipagem limita-se a indicar o grupo alélico, geralmente associado a especificidades sorológicas amplas, como no exemplo HLA-A*02 [Torres and Moraes 2011]. A resolução intermediária acrescenta informações adicionais sobre a sequência, permitindo diferenciar subgrupos dentro do mesmo grupo alélico, um exemplo seria HLA-B*14:HUJ, que identifica um subconjunto específico, mas não define com precisão todas as diferenças de nucleotídeos. Já a alta resolução oferece o nível máximo de detalhe, possibilitando a distinção exata entre alelos com base em variações pontuais na sequência de DNA ou nas diferenças correspondentes na proteína codificada. Nesse nível, alelos como HLA-A*02:01:01 e HLA-A*02:01 são discriminados, refletindo diferenças mínimas, porém relevantes, para aplicações clínicas e imunogenéticas [Kishore and Petrek 2018].

2.2. Redes Neurais Recorrentes Long Short-Term Memory LSTM

O aprendizado profundo (*Deep Learning* - DL) revolucionou o campo da inteligência artificial nos últimos anos, proporcionando um desempenho de ponta em diversas tarefas [Al-lQubaydhi et al. 2024], desde a classificação de imagens até o processamento de linguagem natural. O processamento de dados sequenciais, como textos ou séries temporais, exige arquiteturas especializadas capazes de capturar as relações dentro da sequência. As redes neurais tradicionais *MultiLayer Perceptron (MLP)* apresentam dificuldades nesse tipo de processamento, enquanto as Redes Neurais Recorrentes (RNNs) são mais adequadas para capturar dependências temporais [Yu et al. 2019].

O modelo de Rede Neural Recorrente foi introduzido na década de 1980 para modelagem de dados de séries temporais. Este modelo preserva informações históricas por meio de conexões entre unidades ocultas juntamente com o atraso de tempo. Esse recurso permite detectar correlações temporais entre eventos distantes nos dados. Embora o principal objetivo das RNNs seja capturar e modelar dependências de longo prazo, estudos teóricos e experimentais demonstram que o processamento eficiente dessas informações ao longo do tempo é notoriamente desafiador. Uma solução para este desafio envolve adicionar memória explícita à rede. Assim, Hochreiter e Schmidhuber, em 1997, introduziram o primeiro modelo deste tipo, denominado, em português, de Redes de Memória de Longo e Curto Prazo (LSTM, sigla em inglês para *Long-Short Time Memory*), que incorpora unidades ocultas específicas, que podem aprender, se necessário, a reter informações de entrada por um período prolongado [Hochreiter and Schmidhuber 1997].

2.3. Trabalhos relacionados

Os estudos voltados à imputação de alelos HLA dividem-se em duas vertentes metodológicas. A primeira concentra-se na inferência de genótipos HLA de alta resolução a partir de dados genotipados por SNPs, enquanto a segunda foca na conversão de tipagens de baixa ou intermediária resolução em alelos completos. Nesta última, é comum que

se inclua também a estimativa de informações faltantes em determinados locos, que correspondem a posições específicas no cromossomo onde os genes HLA estão localizados. Em ambas as abordagens, a imputação baseia-se em painéis de referência que reúnem informações haplotípicas, ou seja, conjuntos de alelos em diferentes locos que tendem a ser herdados juntos no mesmo cromossomo. A dependência desses painéis pode, entretanto, comprometer a aplicabilidade dos métodos em populações sub-representadas, uma vez que a distribuição dos alelos HLA varia entre grupos étnicos e regiões geográficas.

Na primeira vertente, a imputação baseada em SNPs tornou-se uma alternativa eficaz frente à tipagem molecular de alta resolução, por ser mais acessível e menos onerosa. Essa abordagem teve início com o trabalho de [Stephen Leslie 2008], que propôs um modelo probabilístico bayesiano para imputar alelos HLA a partir de haplótipos SNP. Tal metodologia deu origem ao HLA*IMP [Alexander T Dilthey 2011] e à sua versão expandida HLA*IMP:02 [Alexander Dilthey 2013], esta última incorporando grafos para acomodar a diversidade haplotípica de populações multiétnicas. O SNP2HLA, por sua vez, codifica alelos como variáveis binárias e utiliza o algoritmo BEAGLE para imputação com base em estruturas de haplótipos SNP [Xiaoming Jia 2013]. Existem outras abordagens de imputação que utiliza aprendizagem profunda (*Deep Learning*). O trabalho de [Junjie Chen 2019] introduziu um *autoencoder* convolucional esparso para imputação de genótipos a partir de dados incompletos. Por sua vez, o trabalho apresentado em [Song et al. 2022] implementou um laço de treinamento personalizado, resultando em uma acurácia de imputação superior em relação ao trabalho de [Junjie Chen 2019].

A segunda vertente, que converte dados de baixa resolução para alta, também tem grande relevância em contextos clínicos e populacionais. O *Easy-HLA* emprega um painel fixo de mais de 600 mil haplótipos dos registros NMDP (Programa Nacional de Doadores de Medula Óssea dos Estados Unidos) e RFGM (Registro Francês de Doadores de Medula Óssea), aplicando um modelo baseado em equilíbrio de *Hardy–Weinberg* para selecionar os diplótipos mais prováveis compatíveis com os dados de entrada [Geffard et al. 2020]. O *GRIMM*, por outro lado, estrutura cada haplótipo conhecido como um caminho em grafo direcionado, armazenando frequências nas arestas e permitindo a imputação probabilística de locos ausentes com alta eficiência computacional [Maiers et al. 2019]. Já o *HaploSFHI* implementa um algoritmo EM iterativo, treinado com dados de NGS de alta qualidade de coortes francesas, e realiza imputações baseadas em blocos trígenicos de alto LD, destacando-se por sua capacidade de estimar locos pouco frequentes como `DQA1` e `DRB3/4/5` com elevada precisão [Lhotte et al. 2024].

É importante ressaltar que, até o presente momento, não há na literatura registros de métodos que realizem a imputação direta de alelos no formato de adotado pela Organização Mundial da Saúde (OMS), conforme proposto neste trabalho.

3. Metodologia

Esta seção descreve a metodologia adotada neste trabalho. Na Seção 3.1, correspondente à etapa (A) da Figura 1, apresenta-se a base de dados do REDOME. A Seção 3.2, apresenta o pré-processamento, etapa (B), e a seleção dos dados em baixa resolução, etapa (C). Por fim, a Seção 3.3, referente às etapas (D) e (E), detalha o método LSTM utilizado na imputação dos dados ausentes. Ressalta-se que outros modelos de redes neurais estão sendo avaliados e serão discutidos em trabalhos futuros.

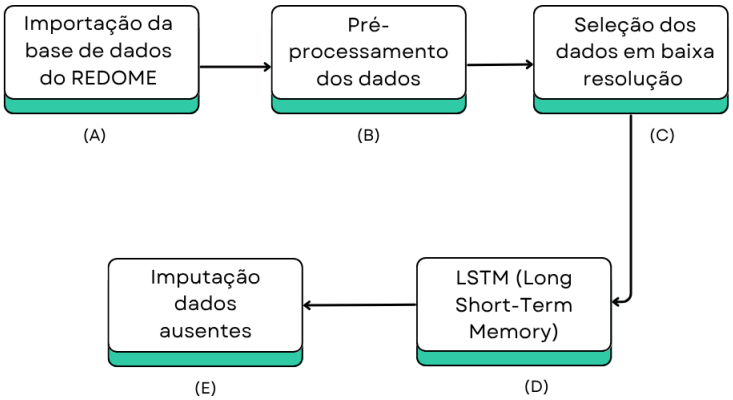

Figura 1. Fluxograma da metodologia adotada neste trabalho

3.1. Base de Dados - REDOME

Este trabalho foi realizado utilizando os dados do Registro Brasileiro de Doadores Voluntários de Medula Óssea (REDOME), que foi criado em 1993, em São Paulo, com intuito de reunir informações de pessoas dispostas a doar medula óssea para pessoas que necessitam de transplante. Seu banco de dados é composto atualmente por mais de 5,5 milhões de doadores cadastrados, sendo o terceiro maior banco de doadores de medula óssea do mundo. O REDOME é maior banco com financiamento exclusivamente público e o Ministério da Saúde é o responsável pela sua administração [Instituto Nacional de Câncer (INCA) 2023]. Quando não há um doador relacionado (como um irmão ou outro parente próximo), a alternativa para realizar o transplante é buscar um doador compatível entre indivíduos não relacionados, ou seja, na população, seja em nível regional ou global.

Esses bancos de dados apresentam uma nomenclatura complexa para o HLA, além de conterem registros dos doadores com dados heterogêneos. Isso se deve ao fato das genotipagens serem registradas durante longos períodos de tempo por meio de diversas técnicas de resolução de tipagem e ambiguidades genotípicas.

É importante destacar que todos os dados utilizados neste estudo foram recebidos com uma codificação anonimizada de registros, em conformidade com a Lei Geral de Proteção de Dados Pessoais (LGPD), de acordo com os princípios éticos descritos na Declaração de Helsinque. Além disso, o estudo foi aprovado pelo Comitê de Ética do Hospital Universitário Pedro Ernesto (CAAE: 56628116.3.0000.5259).

3.2. Análise exploratória e pré-processamento dos dados

A base de dados utilizada é composta por 5.672.482 registros, que corresponde à quantidade de doadores. As informações a respeito dos locos (genes) dos indivíduos correspondem aos pares de alelos dos genes HLA-A, HLA-B, HLA-C, HLA-DRB1, HLA-DQB1, HLA-DPB1, HLA-DQA1. Cada um desses alelos é armazenado em uma coluna. Por exemplo, os dois alelos do loco HLA-A são armazenados nas colunas `dna_a_1n` e `dna_a_2n`, conforme pode ser visto na Tabela 1. Outras informações como sexo, data de nascimento, etnia e unidade federativa também estão presentes, porém, não foram utilizadas neste trabalho. Cabe ressaltar que a etnia tem influencia no alelo que pode ser encontrado e, por isso, será explorado em trabalhos futuros.

Como já descrito na subseção 3.1, é importante destacar que a base de dados contém muitos registros com dados ausentes. Essa falta de dados, em grande parte, está atrelada ao fato de que inicialmente alguns locos não eram tipados. A Tabela 1 exibe a quantidade de valores faltantes para cada alelo. Verifica-se que os locos em A, B e DRB1 são os que possuem as menores quantidades de informações ausentes. Por sua vez, os locos C, DQB1, DPB1 e DQA1 possuem mais de 5 milhões de valores faltantes. Outra observação importante é que a ausência de alelos sempre ocorre em pares, logo para um alelo faltante de `dna_c_1` por exemplo, `dna_c_2` também estará ausente.

Tabela 1. Quantidade de valores faltantes para cada alelo.

Alelos	Valores Faltantes	Alelos	Valores Faltantes
dna_a_1n	3.129	dna_a_2n	3.129
dna_b_1n	1.700	dna_b_2n	1.700
dna_c_1n	5.289.288	dna_c_2n	5.289.288
dna_drb1_1n	13.437	dna_drb1_2n	13.437
dna_dqb1_1n	5.141.111	dna_dqb1_2n	5.141.111
dna_dpb1_1n	5.451.942	dna_dpb1_2n	5.451.942
dna_dqa1_1n	5.422.419	dna_dqa1_2n	5.422.419
sex	0	ethnicity	0
dt_nasc	0	uf_res	5

A Tabela 2 exibe a quantidade de registros documentados em baixa, intermediária e alta resolução para algumas combinações de genes (locos).

Tabela 2. Resumo das resoluções de HLA por diferentes combinações de locos.

Combinações de locos	Baixa	Intermediária	Alta
A, B, DRB1	5.658.832	4.111.958	195.653
A, B, C, DRB1	380.435	332.897	184.514
A, B, C, DRB1, DQB1	342.554	296.027	80.292
A, B, C, DRB1, DQB1, DPB1	220.304	220.226	38.146

Em razão do loco **C** possuir muitos dados ausentes (Tabela 1) e da combinação de locos **A**, **B**, **C** e **DRB1** possuir a maior quantidade de dados dentre os locos ausentes (Tabela 2), esta combinação foi considerada para o treinamento da rede. Após restringir a base de dados a estas 8 colunas e eliminar todos valores faltantes, o número de registros completos foi reduzido para 380.435.

Inicialmente, o treinamento foi realizado considerando apenas baixa resolução, visto a complexidade dos dados genéticos quando expandidos para resoluções mais altas. Logo, registros contendo médias e altas resoluções foram truncados para baixa resolução. Por exemplo, o alelo A*02:01:01 foi truncado para A*02, que corresponde a informação em baixa resolução. Assim, nesse nível de resolução o alelo é identificado pelo seu gene (loco), seguido do valor do alelo. Por exemplo, A*01 indica que se trata do alelo `01` do loco A, enquanto DRB1*02 indica que se trata do alelo `02` do loco DRB1.

Considerando todas as colunas dessa nova base de dados foram identificadas 64 categorias (i.e. 64 alelos distintos em baixa resolução). Desses 64 alelos distintos, 21 pertecem ao loco A, 35 ao B, 13 ao DRB1 e 14 ao C. Vale ressaltar, que alguns alelos são comuns a mais de um loco. Por exemplo, os alelos 01 e 03 aparecem nos locos A, B, DRB1 e C, enquanto o alelo 02 aparece apenas nos locos A e C. Uma nova coluna foi criada, que corresponde à concatenação dos pares de alelos para cada um dos 380.435 registros da base de dados utilizada separados por uma *string* vazia.

Os dados foram separados em dados de entrada (características) e dados de saída (variável alvo). Levando em consideração o número de classes (64 classes) únicas do vocabulário de palavras, os dados de saída foram convertidos em uma matriz one-hot-encoding (nesse caso, uma palavra binária com 64 bits). A utilização do tamanho correto do vocabulário é crucial para definir as dimensões de saída no modelo, especialmente em problemas de classificação ou geração de textos. Além disso, muitos modelos de aprendizado profundo trabalham melhor com saídas categóricas codificadas como one-hot [Hancock 2020]. Os dados foram divididos em 80% para treinamento, 10% para validação e 10% para teste.

3.3. Arquitetura da rede neural

A rede neural investigada neste trabalho foi do tipo LSTM (Long Short-Term Memory), brevemente descrita na subseção 2.2. Nas redes LSTM, cada camada recebe entradas sequenciais e mantém uma conexão recorrente entre os estados internos ao longo do tempo. A saída de cada camada pode ser passada para a próxima camada na arquitetura empilhada. A primeira camada foi a de incorporação (*Embedding Layer*), responsável por transformar os tokens de entrada em representações vetoriais contínuas de números reais. As camadas de incorporação transformam palavras ou outros dados categóricos de entrada em vetores de números reais. Elas são amplamente utilizadas em redes neurais profundas para processamento de linguagem natural (PLN) e outras aplicações que envolvem dados discretos [Hrinchuk et al. 2020]. Elas mapeiam palavras de entrada em representações contínuas e geralmente são implementadas como tabelas de pesquisa (*lookup tables*), armazenadas em matrizes de *embeddings* [Hrinchuk et al. 2020]. Neste trabalho, o número de índices a serem mapeados foi definido com base no tamanho do vocabulário. O tamanho da entrada da rede foi definido como 7, correspondendo aos pares de alelos nos locos HLA-A, HLA-B, HLA-DRB1, além de um alelo individual contido na coluna `dna_c_1n`.

A função de perda empregada foi a entropia cruzada categórica (categorical cross-entropy), adequada para problemas de classificação multiclasse com saída representada como one-hot encoding.

O otimizador *Adaptive Moment Estimation* (Adam) foi escolhido para atualizar os pesos da rede neural. O Adam combina as vantagens do (*stochastic gradient descent - SGD* com *momentum* e do *RMSProp*, ajustando a taxa de aprendizado de forma adaptativa com base no gradiente da função de perda definida. Embora o gradiente descendente estocástico (SGD) seja amplamente utilizado na otimização de redes neurais, variantes como Adam e RMSProp são frequentemente preferidas em redes profundas devido à sua capacidade de ajustar dinamicamente a taxa de aprendizado durante o treinamento. Em contrapartida, o uso do SGD exige a definição cuidadosa da taxa de aprendizado, que muitas vezes precisa ser ajustada manualmente ao longo do processo de treinamento para

garantir uma boa convergência. A validação foi utilizada para monitorar o desempenho do modelo e sinalizar a ocorrência de *overfitting*, identificada quando a perda na validação começa a aumentar, enquanto a perda no treinamento continua diminuindo. O Early Stopping monitora uma métrica de desempenho específica, geralmente a perda de validação ou a acurácia de validação, interrompendo o treinamento quando essa métrica deixa de melhorar após um número pré-definido de épocas. No modelo empregado, foi definido, de forma empírica, durante o processo de modelagem da arquitetura, o valor de *patience* = 5, o que significa que o treinamento é interrompido se a perda de validação não apresentar melhora por 5 épocas consecutivas. A métrica utilizada para monitorar o desempenho do modelo durante o treinamento e validação foi a acurácia.

4. Resultados e Discussão

Foram investigadas arquiteturas com uma ou duas camadas LSTM. Para os modelos com apenas uma camada, foram avaliados diferentes números de unidades LSTM: 32, 64, 96, 128, 130, 140, 150, 160, 170, 180, 200, 220, 240, 256 e 512. Nas arquiteturas com duas camadas, foram testadas as seguintes combinações de unidades: 32×32, 64×64, 128×32, 128×64 e 128×128 (onde o primeiro número corresponde ao número de unidades da primeira camada, e o segundo ao número de unidades da segunda camada). Para essas arquiteturas, o atributo $return\_sequences = True$ foi mantido apenas na primeira camada. Dessa forma, a saída da primeira camada será uma sequência com o mesmo comprimento da entrada, permitindo que a segunda camada LSTM processe esses dados. Com $return\_sequences = False$ na segunda camada, a saída final do modelo corresponde ao último timestep da sequência. Também foi adicionada uma camada totalmente conectada com 128 unidades e função de ativação ReLU, escolhida por favorecer a esparsidade, rápida convergência, simplicidade computacional e por mitigar o problema do desvanecimento do gradiente.

Além disso, foi incluída uma camada de saída totalmente conectada, responsável por calcular a probabilidade de cada token do vocabulário. O número de classes foi definido com base no tamanho do vocabulário, com cada unidade da camada representando um token único, utilizando a função de ativação softmax. Esta função aplica uma transformação exponencial e reescalonamento das saídas, de modo a garantir que a soma das probabilidades seja igual a 1. Dessa forma, a saída do modelo é um vetor de probabilidades que indica a distribuição de probabilidade sobre todas as palavras do vocabulário para uma determinada entrada.

O melhor modelo identificado, a partir dos resultados obtidos com o conjunto de validação, possui duas camadas, a primeira com 128 unidades e a segunda com 32 unidades. O treinamento foi interrompido ao completar 48 épocas, uma vez que o *Early Stopping* foi utilizado. Assim, o treinamento é interrompido de maneira a aumentar a capacidade de generalização do modelo, conforme pode-se observar nas Figuras 2 e 3, que ilustram o comportamento das perdas e acurácias durante as épocas de treinamento.

Após a última época do treinamento, o modelo obteve os seguintes índices. Perda no treinamento: 0,6428, acurácia no treinamento: 0,7809, perda na validação: 0,7306, acurácia na validação: 0,7631. O modelo foi avaliado com os dados de teste, alcançando uma acurácia de 0,7596.

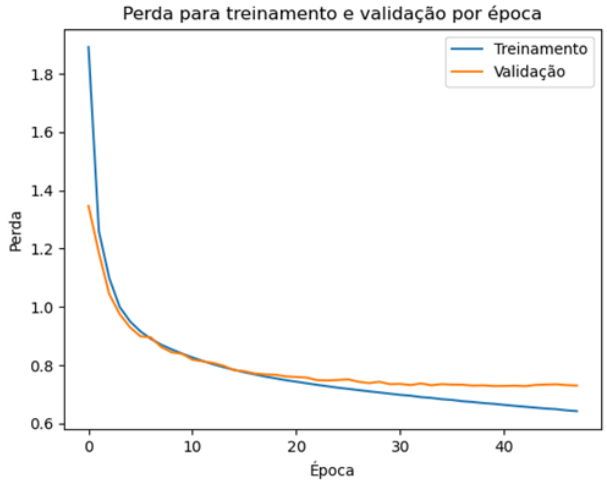

Figura 2. Comparativo das perdas para treinamento e validação durante as épocas

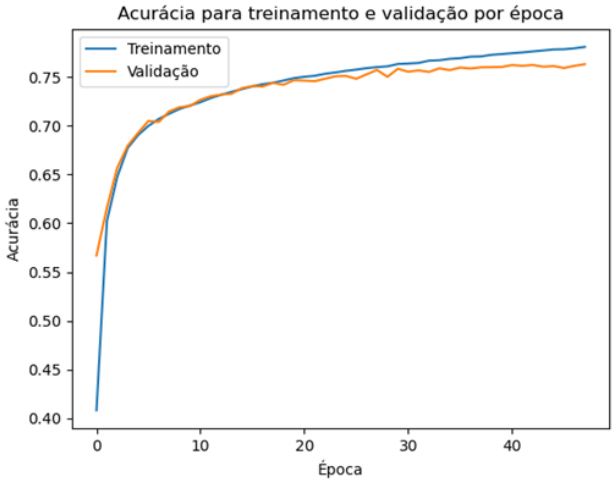

Figura 3. Comparativo das acurácias de treinamento e validação durante as épocas

A acurácia no treinamento (78,09%) é maior do que a acurácia na validação (76,31%), o que indica que o modelo conseguiu se ajustar aos padrões dos dados de treinamento. Essa diferença pode sugerir um leve sobreajuste (overfitting), mas dentro de um nível aceitável, dado que a diferença entre as métricas não é grande e é esperado que a acurácia da validação seja um pouco menor.

Além disso, a acurácia nos dados de teste (75,96%) está próxima da acurácia na validação (76,31%), sugerindo que o modelo generaliza bem para novos dados sem sinais evidentes de sobreajuste, reforçando a consistência do modelo entre diferentes conjuntos de dados. Entretanto, uma análise mais aprofundada, como a validação cruzada, poderia fornecer uma avaliação mais robusta sobre a estabilidade e generalização do modelo.

A Tabela 3 fornece avaliações individuais para todos os alelos presentes no loco HLA-C. Um total de 14 alelos diferentes, em baixa resolução, estão presentes neste loco. Dentre as métricas, individuais e agregadas, contidas nesta tabela podem ser destacadas: precisão, que mede a proporção de previsões positivas corretas em relação ao total de previsões positivas feitas pelo modelo; *recall* (sensibilidade), que mede a capacidade do modelo de identificar corretamente todas as instâncias positivas entre as amostras de uma classe; o *F1-score*, representa a média harmônica entre precisão e *recall*, sendo útil quando há um desequilíbrio entre as classes. Ainda, **suporte** refere-se ao número real de ocorrências de cada classe no conjunto de dados. Dentre as métricas, a **acurácia** representa a proporção de previsões corretas em relação ao total de amostras; *macro average*, que representa a média das métricas (precisão, recall e F1-score) calculadas separadamente para cada classe, atribuindo o mesmo peso a todas elas, independentemente do número de instâncias de cada classe (suporte); e o *weighted average* (média ponderada) calcula a média das métricas, atribuindo pesos proporcionais ao suporte de cada classe, garantindo que classes mais frequentes tenham maior influência na métrica calculada.

Tabela 3. Avaliação dos alelos HLA

Alelo	precisão	Recall	F1-score	suporte
07	0,81	0,84	0,83	10199
02	0,42	0,13	0,20	251
04	0,78	0,85	0,82	3686
03	0,59	0,68	0,63	1162
15	0,56	0,71	0,62	3583
01	0,38	0,31	0,34	26
08	0,91	0,87	0,89	3032
14	0,41	0,30	0,35	1622
16	0,79	0,68	0,73	4360
06	0,78	0,79	0,78	2708
12	0,81	0,69	0,75	3227
05	0,64	0,65	0,65	1607
18	0,80	0,77	0,78	654
17	0,93	0,93	0,93	1927
acurácia			0,76	38044
macro avg	0,69	0,66	0,66	38044
weighted avg	0,76	0,76	0,76	38044

Alelos como `07` (precisão: 0,81, *recall*: 0,84, F1: 0,83) e `08` (precisão: 0,91, *recall*: 0,87, F1: 0,89) apresentaram resultados satisfatórios, com uma relação equilibrada entre precisão e *recall*, indicando que o modelo consegue classificar essas classes de forma consistente. Esse resultado era esperado para classes com maior suporte, pois uma maior quantidade de exemplos disponíveis durante o treinamento, tende a melhorar a capacidade do modelo de aprender e generalizar para essas classes. O alelo `17` obteve precisão (0,93), *recall* (0,93) e *F1-score* (0,93), indicando um desempenho altamente confiável nessa classe. Esse resultado sugere que o modelo consegue identificar corretamente essa classe, devido à presença de um número adequado de exemplos no conjunto de treinamento. Os alelos `02` (*F1-score*: 0,20, suporte: 251) e `01` (*F1-score*: 0,34, suporte: 26) apresentaram desempenho abaixo do esperado, possivelmente devido ao desequilíbrio de classes. O número reduzido de exemplos pode ter limitado a capacidade do modelo de aprender padrões representativos para essas classes. O recall de 0,13 para o alelo `02` indica que o modelo raramente detecta corretamente essa classe, resultando em uma alta taxa de falsos negativos. Os alelos `15` (*F1-score*: 0,62) e `14` (*F1-score*: 0,35) mostram um desempenho intermediário. Embora possuam um suporte maior que algumas classes

minoritárias, o modelo ainda apresenta dificuldades em classificá-los com confiabilidade.

A acurácia global do modelo é 76%, resultado razoável para problemas multi-classes (64 classes considerando todos os locos e 14 classes apenas para o loco C) com desequilíbrio de classes. O *macro average F1-score* de 0,66 reflete o desempenho médio do modelo em todas as classes, atribuindo peso igual a cada categoria, independentemente do suporte. Esse valor confirma que o modelo tem dificuldades em manter um equilíbrio entre as classes minoritárias e majoritárias. O *weighted average F1-score* de 0,76 está alinhado com a acurácia geral, reforçando que o modelo tem um viés para classes majoritárias, pois essa métrica pondera o desempenho de cada classe pelo número de amostras.

5. Conclusões

Este estudo apresentou uma abordagem baseada em Redes Neurais Recorrentes do tipo LSTM para a imputação de alelos HLA em registros de doadores de medula óssea, e que até o momento não foram identificados trabalhos correlatos. A proposta busca mitigar a limitação de informações nos bancos de dados, suprindo a ausência de alelos tipificados por meio de aprendizado profundo. A rede neural implementada alcançou uma acurácia de 76% na imputação dos alelos ausentes do loco HLA-C, evidenciando a viabilidade do método para preenchimento dessas lacunas em registros de baixa resolução.

Além disso, a análise das métricas individuais demonstrou que o modelo possui um desempenho satisfatório para classes com maior suporte, mas encontra dificuldades em prever alelos de baixa frequência devido ao desequilíbrio de classes. Apesar disso, os resultados obtidos são promissores e podem contribuir para o aprimoramento dos bancos de doadores, facilitando a identificação de compatibilidade em transplantes hematopoiéticos. Trabalhos futuros podem explorar estratégias para balanceamento de classes, validação cruzada e incorporação de informações adicionais para melhorar a acurácia da imputação e ampliar a aplicabilidade do modelo no contexto clínico e genético.

Agradecimentos

Os autores agradecem o apoio do CNPq através dos projetos Universal 404087/2021-3 e CNPq/AWS 421828/2022-6.

Referências

Al-lQubaydhi, N., Alenezi, A., Alanazi, T., Senyor, A., Alanezi, N., Alotaibi, B., Alotaibi, M., Razaque, A., and Hariri, S. (2024). Deep learning for unmanned aerial vehicles detection: A review. *Computer Science Review*, 51:100614.

Alexander Dilthey, Stephen Leslie, L. M.-J. S.-C. C. M. R. N. G. M. (2013). Multi-population classical hla type imputation. *PLoS Comput. Bio.*, 9(2):e1002877.

Alexander T Dilthey, Loukas Moutsianas, S. L. G. M. (2011). Hla*imp—an integrated framework for imputing classical hla alleles from snp genotypes. *Bioinformatics*, 27(7):968–972.

Geffard, E. et al. (2019). Easy-HLA: a validated web application suite to reveal the full details of HLA typing. *Bioinformatics*, 36(7):2157–2164.

Geffard, E., Limou, S., Walencik, A., Daya, M., Watson, H., Torgerson, D., Barnes, K. C., CAAPA, Cesbron Gautier, A., Gourraud, P.-A., et al. (2020). Easy-hla: a validated web

application suite to reveal the full details of hla typing. *Bioinformatics*, 36(7):2157–2164.

Hancock, J.T., K. T. (2020). Survey on categorical data for neural networks. *J Big Data*, 7:28.

Hochreiter, S. and Schmidhuber, J. (1997). Long short-term memory. *Neural Computation*, 9(8):1735–1780.

Hrinchuk, O., Khrulkov, V., Mirvakhabova, L., Orlova, E., and Oseledets, I. (2020). Tensorized embedding layers. In Cohn, T., He, Y., and Liu, Y., editors, *Findings of the Association for Computational Linguistics: EMNLP 2020*, Online.

Instituto Nacional de Câncer (INCA) (2023). Quem somos. Acesso em: 03 jan. 2025.

Jeanmougin, M., Noirel, J., Coulonges, C., and Zagury, J.-F. (2017). Hla-check: evaluating hla data from snp information. *BMC bioinformatics*, 18:1–8.

Junjie Chen, X. S. (2019). Sparse convolutional denoising autoencoders for genotype imputation. *Genes*, 10(9):652.

Kishore, A. and Petrek, M. (2018). Next-generation sequencing based hla typing: deciphering immunogenetic aspects of sarcoidosis. *Frontiers in genetics*, 9:503.

Lhotte, R., Letort, V., Usureau, C., Jorge-Cordeiro, D., Consortium, P. A., Siemowski, J., Gabet, L., Cournede, P.-H., Taupin, J.-L., Guillaume, N., et al. (2024). Improving hla typing imputation accuracy and eplet identification with local next-generation sequencing training data. *HLA*, 103(1):e15222.

Maiers, M., Halagan, M., Gragert, L., Bashyal, P., Brelsford, J., Schneider, J., Lutsker, P., and Louzoun, Y. (2019). Grimm: Graph imputation and matching for hla genotypes. *Bioinformatics*, 35(18):3520–3523.

Shaz, B. H., Hillyer, C. D., and Gil, M. R. (2013). *Blood Banking and Transfusion Medicine - History, Industry, and Discipline*.

Song, M., Greenbaum, J., Luttrell, J., Zhou, W., Wu, C., Luo, Z., et al. (2022). An autoencoder-based deep learning method for genotype imputation. *Frontiers in Artificial Intelligence*, 5.

Stephen Leslie, Peter Donnelly, G. M. (2008). A statistical method for predicting classical hla alleles from snp data. *American Journal of Human Genetics*, 82(1):48–56.

Tiercy, J.-M. (2016). How to select the best available related or unrelated donor of hematopoietic stem cells? *Haematologica*, 101(6):680–687.

Torres, M. A. and Moraes, M. E. H. (2011). Nomenclatura dos fatores do sistema hla. *einstein (São Paulo)*, 9:249–251.

Xiaoming Jia, Buhm Han, S. O.-G. W.-M. C. P. J. C. S. S. R. S. R. P. I. W. d. B. (2013). Imputing amino acid polymorphisms in human leukocyte antigens. *PLoS One*.

Yu, Y., Si, X., Hu, C., and Zhang, J. (2019). A review of recurrent neural networks: Lstm cells and network architectures. *Neural Computation*, 31(7):1235–1270.