UNIVERSIDADE DE SÃO PAULO
Instituto de Ciências Matemáticas e de Computação

Inteligência Artificial na predição da Doença de Parkinson

Pedro Quinaud Minchilo

Trabalho de Conclusão de Curso
MBA em Inteligência Artificial e Big Data

ICMC USP SÃO CARLOS

UNIVERSIDADE DE SÃO PAULO
Instituto de Ciências Matemáticas e de Computação

Inteligência Artificial na predição da Doença de Parkinson

Pedro Quinaud Minchilo

USP - São Carlos
2023

Ficha catalográfica elaborada pela Biblioteca Prof. Achille Bassi
e Seção Técnica de Informática, ICMC/USP,
com os dados inseridos pelo(a) autor(a)

Q7i

Quinaud Minchilo, Pedro
Inteligência Artificial na predição da Doença de Parkinson / Pedro Quinaud Minchilo; orientador Diego Raphael Amancio. -- São Carlos, 2023.
26 p.

Trabalho de conclusão de curso (MBA em Inteligência Artificial e Big Data) -- Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, 2023.

1. Inteligência Artificial. 2. Doença de Parkinson. 3. LSTM. 4. Redes Recorrentes. I. Raphael Amancio, Diego, orient. II. Título.

Bibliotecários responsáveis pela estrutura de catalogação da publicação de acordo com a AACR2:
Gláucia Maria Saia Cristianini - CRB - 8/4938
Juliana de Souza Moraes - CRB - 8/6176

RESUMO

QUINAUD, P. **Inteligência Artificial na predição da Doença de Parkinson**. 2023. 26p. Monografia (Trabalho de Conclusão de Curso) - Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos, 2021.

A doença de Parkinson é uma doença degenerativa do sistema nervoso central que afeta de diversas formas o ser humano, desde sua cognição até funções motoras básicas. As causas da doença ainda não são completamente conhecidas, apesar dos avanços na ciência e da grande mobilização acerca do tema. Estudos recentes apontam que anormalidades de proteínas nos pacientes possuem uma influência fundamental na evolução da enfermidade, e um melhor entendimento desta relação pode ajudar no combate e cura da doença. O presente estudo investiga a relação entre a evolução da doença de Parkinson e os níveis de proteína no líquido cefalorraquidiano dos pacientes. Inicialmente, se buscou prever a evolução da doença na escala *Unified Parkinson's Disease Rating Scale (MDS-UPDRS)* através de sua própria série temporal, usando uma rede recorrente *Long-Short Term Memory (LSTM)*. Em seguida, esta previsão foi refinada enriquecendo os dados de entrada com informações de abundância de proteínas no líquido cefalorraquidiano dos pacientes obtidos em exames de espectrometria de massa. A expectativa seria de que o algoritmo adaptado fosse capaz de usar os dados enriquecidos para melhorar a previsão, e assim possivelmente aprender as relações complexas existentes. O resultado não corroborou com a expectativa, e o desempenho do modelo adaptado não se mostrou melhor considerando as métricas adotadas de erro percentual médio e *Root-Mean Squared Error (RMSE)*. Porém, é precipitado afirmar que o modelo não seja capaz de aprender tais relações, e outras ações visando atingir o objetivo foram discutidas e propostas para trabalhos futuros.

Palavras-chave: Inteligência Artificial. Doença de Parkinson. LSTM. Redes Recorrentes.

ABSTRACT

QUINAUD, P. **Artificial Intelligence in prediction of Parkinson's Disease**. 2023. 26p. Monografia (Trabalho de Conclusão de Curso) - Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos, 2021.

Parkinson's disease is a degenerative disease of the central nervous system that affects humans in many ways, from cognition to basic motor functions. The causes of the disease are still not completely known, despite advances in science and the great mobilization on the subject. Recent studies indicate that patient’s protein abnormalities have a fundamental influence on the evolution of the disease, and a better understanding of this relationship can help in combating and curing the disease. The present study investigates the relationship between the evolution of Parkinson's disease and protein levels in the patients' cerebrospinal fluid. Initially, the sutdy sought to predict the evolution of the disease on the Unified Parkinson's Disease Rating Scale (MDS-UPDRS) through its historical time series, using an Long-Short Term Memory (LSTM) recurrent network. This prediction was then refined by enriching the input data with information on the abundance of proteins in the patients' cerebrospinal fluid obtained from mass spectrometry scans. The expectation was that the adapted algorithm would be able to use the enriched data to improve the prediction, and thus possibly learn the existing complex relationships. The result did not corroborate expectations, and the performance of the adapted model was not better considering the adopted metrics of average percentage error and Root-Mean Squared Error (RMSE). However, it is hasty to say that the model is not capable of learning such relationships, and other actions aimed at achieving the objective were discussed and proposed for future researchs.

Keywords: Artificial Intelligence. Parkinson's disease. LSTM. Recurrent Networks.

SUMÁRIO

1. INTRODUÇÃO

1.1. Contextualização do problema

A doença de Parkinson (em inglês *Parkinson's disease, PD*) é uma doença degenerativa do sistema nervoso central que afeta a cognição, a coordenação motora, o sono e outras funções normais. Descrita pela primeira vez em 1817 por James Parkinson em um artigo intitulado *"An Essay on The Shaking Palsy"* (em tradução livre "Um ensaio sobre a paralisia trêmula"), a doença ocorre devido à degeneração dos neurônios produtores de dopamina na região do cérebro conhecida como substância negra, o que leva à uma deficiência deste neurotransmissor que conduz os impulsos nervosos ao corpo (Poewe, 2017).

A Organização Mundial da Saúde (OMS) estima que a Doença de Parkinson afete aproximadamente 1% da população acima de 65 anos, sendo a segunda causa mais comum de desordem neurodegenerativa (Ministério da Saúde, 2018). No Brasil, a quantidade de pessoas que sofrem atualmente com a doença é estimada em mais de 200 mil, e este número tende a aumentar nos próximos anos. Já nos Estados Unidos, 1.6 milhões de pessoas terão a doença até 2037, representando um custo na saúde de US$ 80 Bilhões (Willis et al., 2022).

1.2. Motivação e justificativa

As causas da degeneração não são totalmente compreendidas, mas acredita-se que se trata de combinações de fatores genéticos e do ambiente. Pesquisas recentes indicam que anormalidades de proteínas ou peptídeos possuem um papel fundamental no aparecimento e agravamento da doença (Ganguly et al, 2022). Infelizmente, atualmente não há cura para a doença e a mesma evolui com o tempo.

Algumas organizações como a *Accelerating Medicines Partnership® Parkinson's Disease* (AMP®PD) vêm conduzindo diversos estudos visando entender melhor a doença, suas causas e sua evolução. Conforme a organização citada salienta, os esforços resultaram em dados clínicos e neurobiológicos complexos de milhares de indivíduos para amplo compartilhamento com a comunidade de pesquisa. Apesar de importantes descobertas terem sido publicadas usando esses dados, ainda faltam biomarcadores claros ou curas. A ciência de dados pode ajudar a ter mais compreensão da doença através dos dados, contribuindo para caminhos que possam desacelerar a progressão da doença ou até mesmo curá-la.

1.3 Objetivos

Recentemente, têm sido crescente o interesse e o uso de Inteligência Artificial (IA) em estudos e diagnósticos da doença de Parkinson (Rastegar, 2019). Exemplos de uso envolvem a aplicação de Deep Learning para analisar imagens de exames de ressonância e assim identificar biomarcadores para a doença.

O presente trabalho tem como objetivo usar inteligência artificial para prever a evolução da doença de Parkinson através de uma escala padrão, e em seguida usará dados de níveis de proteínas dos pacientes para avaliar se é possível melhorar esta previsão. Com isso, o trabalho pode ajudar a fornecer pistas importantes sobre quais proteínas e peptídeos influenciam na doença à medida que esta progride.

2. REVISÃO BIBLIOGRÁFICA

O presente estudo se fundamenta em dois campos do conhecimento, sendo eles: a área de estudos sobre distúrbios no sistema nervoso central, reunindo conhecimentos em medicina, fisiologia e biologia (Vatansever, 2021); e a área de Inteligência Artificial, um segmento da computação que estuda sistemas capazes de aprender a executar tarefas sem a necessidade de uma instrução específica prévia (Haykin, 2009).

2.1 Doença de Parkinson – Sintomas e progressão da doença

Apesar dos sintomas clínicos mais perceptíveis da doença estarem ligados às questões motoras como rigidez, tremor, lentidão (bradicinesia) e outras implicações; diversos sintomas não-motores estão relacionados ao quadro geral de inaptidão. Sua patogênese molecular envolve múltiplos mecanismos incluindo função mitocondrial, proteostase da α-synuclein (mecanismo de degradação da proteína), stress oxidativo, entre outros (Poewe, 2017).

Sintomas não-motores se tornaram prevalecentes e óbvios ao longo do curso da doença e são os maiores determinantes da qualidade de vida, progressão da doença, inaptidão geral e necessidade de suporte médico e de enfermeiros. No longo-prazo, fatores como demência acometem 83% dos casos, alucinose em 74% e constipação 40% (Poewe, 2017).

2.2 Doença de Parkinson – Biomarcadores e Diagnóstico

De acordo com Evans e Rosenberg no livro "*Biomarkers: The 10 Keys to Prolonging Vitality*", um biomarcador é uma característica que é medida objetivamente e avaliada como

um indicador de um processo normal ou patológico ou como uma resposta à intervenção terapêutica. Eles podem ajudar clínicos a monitorar a evolução da patologia, bem como contribuir com estudos para entender melhor os mecanismos da doença e possíveis desenvolvimentos de drogas para combatê-la. Além disso, biomarcadores são importantes em quadros onde a apresentação de sintomas não é clássica ou facilmente identificada em quadros clínicos (Ganguly et. al, 2021).

Biomarcadores para a Doença de Parkinson são normalmente categorizados como genéticos, baseados em imagens ou bioquímicos. No caso de marcadores bioquímicos, algumas proteínas presentes em fluídos corporais demonstram relativo potencial (Emamzadeh and Surguchov, 2018; Ganguly et. al, 2021). Porém, apesar de estudos constatarem uma variedade de proteínas no CSF de pacientes com Parkinson comparado à pacientes controle, os resultados não são conclusivos e atualmente não há um biomarcador definido para diagnóstico via CSF útil clinicamente. (Poewe, 2017)

Nos casos em que as implicações motoras da Doença de Parkinson estão presentes de forma clássica, a avaliação clínica pode ser suficiente para o diagnóstico. Porém, devido às manifestações não usuais e similaridade com outras patologias, erros de diagnóstico nos estágios iniciais podem atingir o patamar de até 24% mesmo em centros especializados. A melhoria de acurácia e até mesmo a realização de um diagnóstico definitivo passam pela avaliação dos biomarcadores de imagem, genéticos ou bioquímicos

2.3 Doença de Parkinson – Escala unificada de avalição da doença (UPDRS)

Testes clínicos e estudos observacionais até então têm focado muito na progressão de implicações motoras, padronizadas pela escala unificada de classificação da Doença de Parkinson (em inglês *Unified Parkinson's Disease Rating Scale*, UPDRS), sendo essa a escala mais comum e utilizada para monitorar a evolução da inaptidão motora da doença (Poewe, 2017).

A escala foi publicada pela primeira vez por Fahn, S.; Elton, R.; e membros do Comitê de Desenvolvimento da UPDRS em 1987. Atualmente, a escala é revisada pela instituição *Movement Disorder Society*, e sua utilização se dá através de questionários-padrão aplicados em exames clínicos. Ao todo, a MDS-UPDRS possui quatro partes principais:

I - Experiências não-motoras da vida diária

II - Experiências motoras da vida diária

III - Exame motor

IV - Complicações motoras

2.4 Inteligência artificial – conceitos fundamentais

Segundo Russel e Norvig (2020), a Inteligência Artificial trata da concepção de sistemas que possuem inteligência para aprender, perceber e tomar decisões em um ambiente. Já *Machine Learning* é um subcampo da inteligência artificial focado na criação de programas computacionais que aprendem a partir de exemplos. O mecanismo de aprendizado destes sistemas pode ser classificado como aprendizado supervisionado, aprendizado não supervisionado e aprendizado por reforço, dependendo da maneira como os dados para o 'aprendizado' do sistema são fornecidos.

Diversas são as formas de implementação de sistemas de Inteligência Artificial, dependendo do problema proposto. Tais implementações usualmente remetem ao conceito de Redes Neurais Artificiais, que utilizam nós de processamento e interconexões análogos às estruturas do cérebro biológico (Haykin, 2009). Para o problema no presente estudo, a solução passa por implementações voltadas às séries temporais. Neste caso, as estruturas mais comuns são:

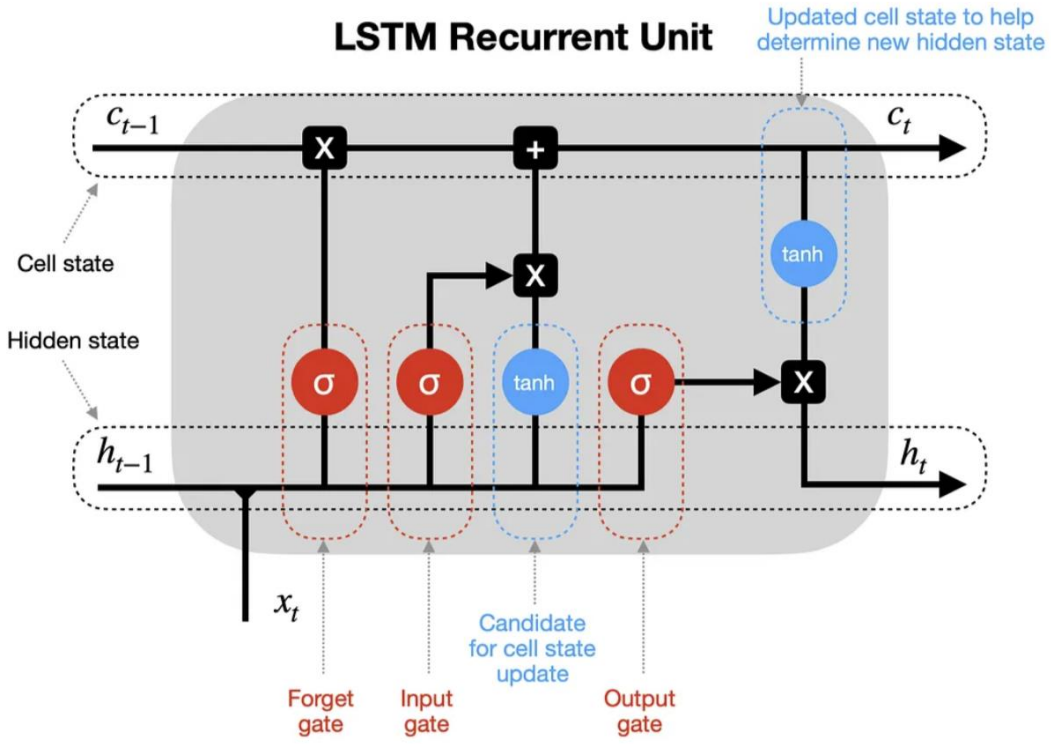
- Redes Neurais Recorrentes (em inglês, *Recurrent Neural Network - RNN*): tipo de rede neural que possui uma memória interna, permitindo que a rede se lembre de informações do passado para informar suas previsões futuras. São especialmente adequadas para dados sequenciais, como series temporais.
- Redes Temporais Convolucionais (em inglês, *Temporal Convolutional Network - TCN*): adaptação de redes neurais convolucionais, utilizando conceitos de dilatação e relação causal para que seja aplicado em séries temporais.

2.5 Inteligência Artificial – Redes LSTM

Redes Neurais Recorrentes são um tipo de rede neural artificial projetadas para lidar com sequências de dados, onde a ordem e a dependência temporal dos elementos são importantes. Dentre estas, um modelo específico é a Rede *Long-Short Term Memory (LSTM)*, projetada para lidar com as limitações das redes recorrentes tradicionais como os problemas de desaparecimento ou explosão dos gradientes. Tal modelo foi proposto em 1997 pelos cientistas da computação Sepp Hochreiter e Jürgen Schmidhuber. Em redes LSTM, a unidade de processamento principal é chamada de célula de memória (*memory cell*), e as camadas ocultas são encapsuladas dentro desta célula. Esta arquitetura possui três componentes principais: uma porta de entrada (*input gate*), uma porta de esquecimento (*forget gate*) e uma porta de saída (*output gate*). Estas portas controlam o fluxo de informações, permitindo que o neurônio se

lembre ou esqueça informações em diferentes pontos temporais (Hochreiter e Schmidhuber, 1997). A ilustração abaixo representa tais componentes de forma ilustrativa:

Figura 1 – Esquema de representação da célula LSTM



Fonte: Saul Dobilas, website Towards Data Science, 2022

2.6 Inteligência Artificial – aplicações envolvendo Parkinson

Pesquisadores da Faculdade de Medicina e Saúde da Universidade de Sidney, Australia, utilizaram modelos de *Machine Learning* para predizer a progressão da doença de Parkinson, tendo como alvo as escalas *Hoehn and Yahr* e *UPDRS*. As predições foram feitas através de amostras dos grupos de proteínas citocinas e quimiocinas dos pacientes, bem como dados clínicos. Foram usados algoritmos de *Random Forest* e *Elastic-net* para a predição das escalas, e a medida de desempenho utilizada foi a Raiz do Erro Quadratico Médio normalizada.

Como resultado, foi constatada a contribuição de proteínas como MIP1α e MCP1 na predição, bem como uma aderência maior na escala UPDRS citada, dando indícios da contribuição de citocinas para a predição da progressão da doença com o auxílio de *Machine Learning*. Além disso, quando os dados clínicos e dados das proteínas foram combinados para

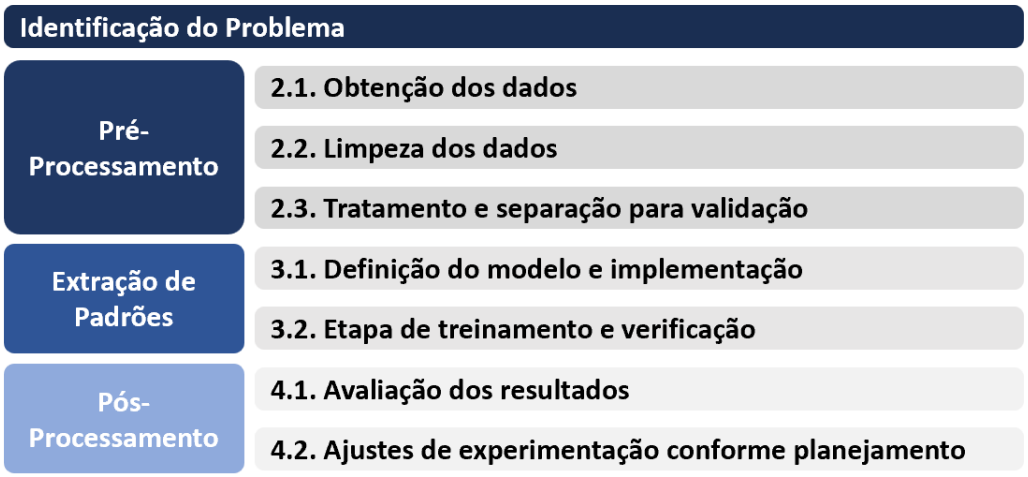
análise, o desempenho do modelo *Random Forest* se mostrou pior em relação ao uso dos dados das proteínas separadamente. (Ahmadi Rastegar et al. 2019)

Outro estudo que traz contribuições para a aplicação de Inteligência Artificial em pesquisas relacionadas à Doença de Parkinson foi feito por pesquisadores da Universidade de Bonn e Schloss Birlinghoven (Alemanha), intitulado "Artifcial intelligence-based clustering and characterization of Parkinson's disease trajectories". Na ocasião, foi constatado que existem diferentes padrões de progressão da Doença de Parkinson, e estabelecer clusters baseados nesta progressão traz impactos relevantes durante a pesquisa e conduzem a resultados mais precisos. Para o estudo, foram feitos agrupamentos LASSO utilizando seis diferentes variáveis considerando sintomas tanto motores quanto não-motores (Birkenbihl et al, 2023).

3. DESENVOLVIMENTO

Tratando-se de uma abordagem voltada à ciência de dados, o desenvolvimento deste problema se deu em um ambiente computacional comumente utilizado pela área, através do uso de Python e das bibliotecas *Pandas* para leitura e manipulação dos dados, *Sci-Kit Learn* para recursos em *machine learning*, e *PyTorch* para implementações de redes neurais. O trabalho seguiu uma adaptação do processo de mineração de dados (Rezende, Solange ; 2003), sendo representado pela seguinte metodologia:

Figura 2 - Diagrama ilustrando metodologia adotada e suas etapas



Fonte: próprio autor

O detalhamento de cada uma das etapas constantes na metodologia será exposto nos capítulos subsequentes.

3.1. Identificação do Problema

Como exposto no capítulo inicial, o objetivo do trabalho é buscar prever a evolução da Doença de Parkinson utilizando a escala 'UPDR'. Foram usados três dos quatro *scores* devido à disponibilidade dos dados, sendo eles:

I. Experiências não-motoras da vida diária ;

II. Experiências motoras da vida diária ;

III. Exame motor

Em seguida, esta previsão foi refinada enriquecendo os dados de input do modelo com informações de abundância de proteínas no líquido cefalorraquidiano dos pacientes obtidos em exames de espectrometria de massa. Caso o algoritmo seja capaz de aprender a usar os dados enriquecidos para melhorar a previsão, isso poderá representar um avanço no entendimento de quais proteínas têm mais influência na progressão da doença. Algoritmos auxiliares para avaliar a explicabilidade do modelo poderão ser aplicados, como o algoritmo SHAP (SHapley Additive exPlanations). Visando melhor organização do experimento, sua execução foi dividida em duas partes, que seguiram as mesmas etapas expostas na Figura 2, sendo elas:

- *Parte 01 - Predição dos scores UPDR usando série temporal histórica*
- *Parte 02 - Enriquecimento do modelo usando dados de abundância de proteína*

3.2. Pré-Processamento – Obtenção dos dados

O experimento conta com dados clínicos de pacientes ao longo de 36 meses (três anos) em um intervalo semestral. Ao todo, serão 7 amostras de cada paciente (meses 0, 6, 12, 18, 24, 30 e 36), representando suas respectivas consultas (ou visitas). Tais amostras contém os valores dos três scores UPDRS, bem como medições de abundância de 227 diferentes proteínas.

Os dados foram disponibilizados pela Accelerating Medicines Partnership® Parkinson’s Disease (AMP®PD), para pesquisa acadêmica e educacional através da plataforma Kaggle. Foram usados majoritariamente dois dos arquivos disponibilizados: Dados Clínicos, relativos às informações de consulta de cada paciente como mês de visita e o respectivo score

UPDR ; e Dados de Proteínas, relativo aos dados de abundância de proteína que cada paciente possuía na consulta específica. O detalhamento de cada um destes arquivos segue abaixo.

Tabela 1 - Dados relativos às consultas do paciente

Atributo	Descrição
patient_id	Código de identificação para o paciente.
visit_month	O mês da visita, em relação à primeira visita do paciente.
visit_id	Código de identificação para a visita.
updrs_1 a updrs_4	Score do paciente para cada parte da Escala Unificada de Avaliação da Doença de Parkinson (UPDRS). Números mais altos indicam sintomas mais graves.
clinical_state_on_medication	Se o paciente estava ou não tomando medicação como Levodopa durante a avaliação UPDRS. Espera-se que afete principalmente as pontuações da Parte 3 (função motora). Esses medicamentos desaparecem rapidamente (na ordem de um dia), por isso é comum que os pacientes façam o exame de função motora duas vezes em um único mês, com e sem medicação.

Fonte: AMP®PD - plataforma Kaggle, 2023

Tabela 2 - Frequências de expressão de proteínas nos pacientes

Atributo	Descrição
patient_id	Código de identificação para o paciente.
visit_month	O mês da visita, em relação à primeira visita do paciente.
visit_id	Código de identificação para a visita.
UniProt	O código de ID UniProt para a proteína associada
NPX	Expressão proteica normalizada. A frequência da ocorrência da proteína na amostra.

Fonte: AMP®PD - plataforma Kaggle, 2023

3.3. Pré-Processamento – Limpeza dos dados

Inicialmente, a base contou com registros de 248 pacientes distintos. Cada paciente possui um número diferente de visitas, variando entre 3 e 17. O intervalo entre as visitas também varia, de 3 a 18 meses. Neste problema, cada paciente representa uma série temporal diferente, e por isso os dados de cada paciente deverão ser tratados de forma independente. Além disso, outros ajustes serão necessários para que o modelo possa interpretar corretamente os dados.

Os dados se encontram em formato tabular bidimensional, e precisam ser reorganizados em uma estrutura tridimensional de formato:

batch x length x features

Sendo *batch* o bloco de dados de cada paciente correspondendo à série temporal ; *length* representa o registro das visitas, também referidas como *time-steps* ; *features* são os atributos relacionados à cada visita, como o score UPDRS e os níveis de proteína. Como a implementação se dará através de uma Rede Neural Recorrente, é sugerido que cada paciente possua a mesma quantidade de registros (*length*). Além disso, é importante que o intervalo entre as visitas seja o mesmo.

Dessa forma, foi adotado o intervalo padrão de 6 meses entre visitas, ao longo de 3 anos. Para ajustar a base, foi feita a limpeza retirando registros de visitas e de pacientes que não atendiam a este requisito. Por fim, também foram eliminados pacientes que possuíam dados faltantes para os três scores a serem preditos.

3.4. Pré-Processamento – Tratamento de dados e separação para validação

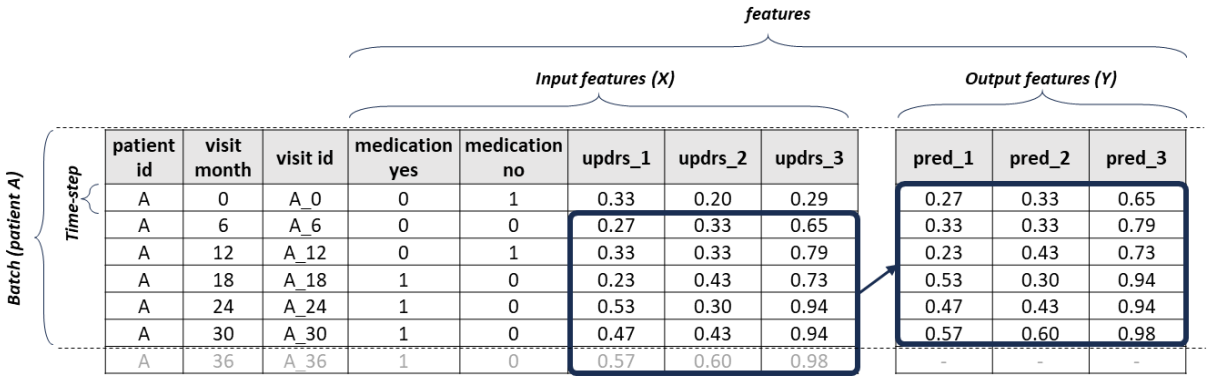
Após a limpeza inicial dos dados, estes foram submetidos ao tratamento de normalização para os atributos de score UPDRS 1 a 3. A implementação se deu através da equação:

$$x_{norm} = \frac{x - min(x)}{min(x) - max(x)} \qquad \textbf{[1]}$$

Também foi realizado o procedimento *one-hot encoding* para o atributo 'em_medicação'. Este procedimento transforma os diferentes valores que o atributo pode assumir em respectivas colunas, e atribui valores 0 ou 1 às colunas dado o estado do atributo em cada registro.

Com os dados já normalizados e tratados, estes foram separados em dois grupos, um para treino e outro para teste. A separação foi feita utilizando a função *train_test_split*, da biblioteca *Scki-Kit learn*, e usou como parâmetro de separação 70% para treino e 30% para testes. Uma vez separados, é necessário ajustar os dados a serem usados como input (denotados por X) e os dados alvo a serem preditos (denotados por Y). Para tal, os *batchs* foram organizados de forma que cada registro de visita (*time-step*) tivesse os seus *scores* UPDR presentes nas input features (X) sendo usados como dado alvo (Y) do *time-step* anterior. A estrutura final dos dados pode ser observada na figura abaixo. O ajuste descrito, visando definir os dados alvos Y, é representado no esquema através das bordas espessas contornando os valores numéricos.

Figura 3 – Estrutura de dados tratados para utilização



patient id	visit month	visit id	medication yes	medication no	updrs_1	updrs_2	updrs_3	pred_1	pred_2	pred_3
A	0	A_0	0	1	0.33	0.20	0.29	0.27	0.33	0.65
A	6	A_6	0	0	0.27	0.33	0.65	0.33	0.33	0.79
A	12	A_12	0	1	0.33	0.33	0.79	0.23	0.43	0.73
A	18	A_18	1	0	0.23	0.43	0.73	0.53	0.30	0.94
A	24	A_24	1	0	0.53	0.30	0.94	0.47	0.43	0.94
A	30	A_30	1	0	0.47	0.43	0.94	0.57	0.60	0.98
A	36	A_36	1	0	0.57	0.60	0.98	-	-	-

Fonte: próprio autor

Os *batchs* resultantes foram então transformados em tensores para serem processados na implementação em PyTorch.

3.5. Extração de Padrões – Definição do modelo e implementação

O modelo a ser usado para realizar a predição é uma rede neural LSTM (*Long-Short Term Memory*), a ser implementada em PyTorch. O uso de um modelo LSTM é interessante por ser capaz de capturar padrões e relações entre diferentes séries temporais (Hochreiter e Schmidhuber, 1997). Para o modelo implementado, foi utilizado apenas um *layer* de camada oculta contendo 30 unidades de memória (*hidden size*), e uma camada de saída *full-connect* linear com 3 *outputs*, um para cada score.

Para a função de perda, foi utilizada a métrica de Erro Quadrático Médio (MSE, 'Mean Squared Error'). O uso da métrica MSE como função de perda foi escolhido devido aos seus benefícios de sensibilidade a grandes erros, a interpretação intuitiva e tendência a treinamento estável. A métrica pode ser calculada conforme a equação:

$$\mathrm{MSE} = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2 \qquad \textbf{[2]}$$

3.6. Extração de Padrões – Etapa de treinamento e verificação

Em relação ao treinamento da rede, foram definidos como critérios 1.500 épocas de treino a uma taxa de aprendizado de 0.001. Após o treinamento, as amostras foram avaliadas sob o critério MSE e também sua raiz quadrada, conhecida por Erro Quadrático Médio da Raiz (RMSE, 'Root Mean Squared Error'). Estes valores serão usados posteriormente para comparação com os valores obtidos pelas amostras de teste, bem como para comparação com o segundo momento do experimento envolvendo a avaliação do modelo utilizando os dados enriquecidos.

Com o modelo já treinado, este é alterado para o modo de avaliação utilizando o método *eval* do PyTorch. Isso afeta o comportamento do modelo durante a inferência e garante que não haja influência de mecanismos como *Dropout* e *Batch Normalization*. No modo de avaliação, é então realizada a predição para os dados de teste e assim obtendo os valores que serão avaliados pelas mesmas métricas anteriormente citadas.

A comparação entre o desempenho das amostras de treino e teste é fundamental para entender a quão boa está a capacidade de resolução do problema e generalização do modelo, possibilitando a detecção de *overfitting* (quando um modelo se ajusta excessivamente aos detalhes dos dados de treinamento e não é capaz de generalizar bem para novos dados). Esta comparação auxilia na avaliação do modelo e permite um melhor ajuste dos hiper parâmetros, caso seja necessário.

3.7. Pós-Processamento – Avaliação dos resultados

Após o treinamento do modelo e a validação com os dados de teste, os resultados das predições serão retornados à escala original, desfazendo o processo de normalização e permitindo um entendimento mais próximo da realidade. Este procedimento é realizado através do método *inverse_transform* presente na mesma classe *MinMaxScaler* usada inicialmente. Em seguida, os dados serão analisados através da métrica erro percentual médio, permitindo

interpretar numericamente a precisão e o erro do modelo. O cálculo se dará apenas para a predição do último time-step, referente ao mês 36. Visando obter mais clareza no comportamento, algumas amostras de pacientes também serão tratadas graficamente, utilizando gráficos de linhas para expressar a série temporal.

3.8. Pós-Processamento – Ajustes de experimentação conforme estudo planejado

Finalizada a primeira parte do experimento, envolvendo a implementação da rede neural LSTM e sua respectiva utilização para predição dos scores UPDRS, o experimento será adaptado para incluir novos dados relativos à abundância de proteínas no líquido cefalorraquidiano dos pacientes em cada *time-step*. Para tal, os passos de pré-processamento expostos no início do capítulo de desenvolvimento serão realizados novamente, desta vez utilizando também a base de proteínas já exposta.

A manipulação consistirá em uma operação de união (comumente denominada *Join*) entre os dados clínicos das visitas e os dados de proteína, através de uma chave composta pelo binômio paciente-visita. A base resultante contará tanto com os dados já utilizados de paciente, visita, scores UPDRS e medicação, como também contará com 227 novos atributos referentes aos níveis de proteína de cada paciente em cada consulta. Uma vez obtida esta base resultante, será necessário avaliar os dados faltantes dos pacientes e eliminar casos de inconsistência.

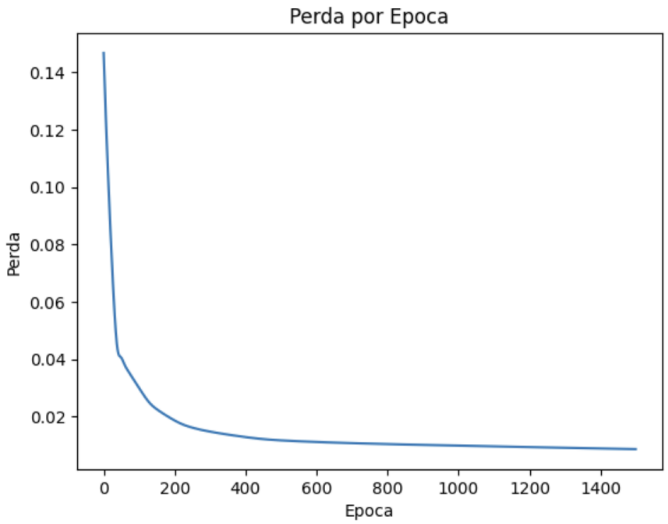
Como o acesso aos dados de proteína é mais limitado e possui uma quantidade relevante de registros faltantes, o intervalo de tempo entre visitas precisará ser ajustado de 6 meses para 01 ano. Este ajuste irá alterar o tamanho (*length*) de cada batch de 07 para 04 *time-steps*. Além disso, os hiper parâmetros usados poderão necessitar de ajustes uma vez que a dimensionalidade das *features* teve um aumento expressivo em ordem de grandeza.

4. RESULTADOS E DISCUSSÃO

4.1. Parte 01 - Predição dos scores UPDR usando série temporal histórica

Após a implementação da rede LSTM exposta anteriormente, a mesma foi treinada e validada conforme os dados sanitizados expostos. Ao todo, os dados sanitizados resultaram em 150 pacientes, sendo divididos em 105 para treino (70%) e 45 para teste (30%). O treinamento ocorreu em 1.500 épocas. O resultado da métrica MSE para cada época pode ser observado no gráfico abaixo:

Gráfico 1 - Perda por Época em Parte 01, métrica MSE



Fonte: Próprio autor

O comportamento mostra uma grande queda até cerca de 300 épocas, com diminuição no ritmo de minimização da perda após este valor. O resultado final de treinamento para as métricas MSE e RMSE, considerando apenas o último período de predição (trigésimo-sexto mês), pode ser observado abaixo:

Mean Squared Error: 0.0065
Root Mean Squared Error: 0.0804

De forma análoga, os dados de teste foram submetidos ao mesmo procedimento, atingindo as métricas MSE e RMSE abaixo:

Mean Squared Error: 0.0161
Root Mean Squared Error: 0.1270

Comparando os resultados de treino e teste, é possível observar que a métrica MSE atingiu um valor três vezes maior no teste em relação ao treino. Para que haja maior interpretabilidade, iremos reverter a normalização imposta aos valores preditos e utilizar uma comparação de erro percentual médio nos valores finais, sendo:

$$erro\ percentual\ = \frac{|\ y_{predito}\ -\ y_{real}\ |}{y_{real}} \qquad \textbf{[3]}$$

Os valores de erro percentual médio para o trigésimo-sexto mês de cada score foram:

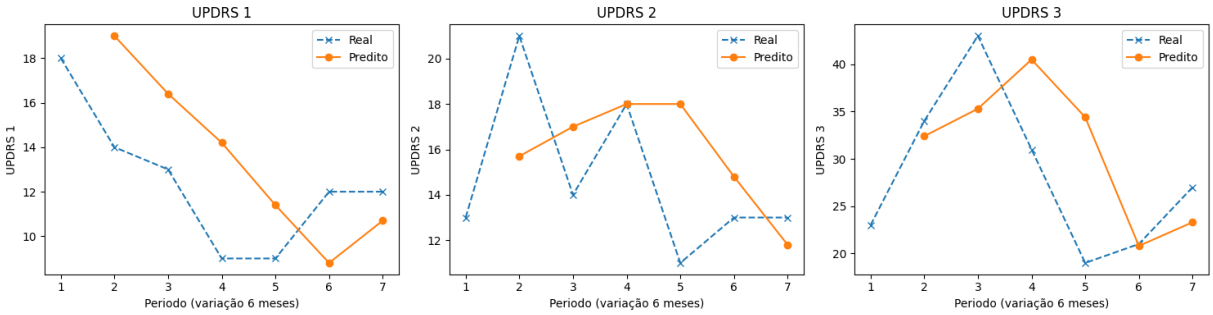
Tabela 3 – Resultados do modelo considerando score original, Parte 01

Score	Erro percentual médio
UPDRS 1	57.8%
UPDRS 2	41.3%
UPDRS 3	29.0%

Fonte: Próprio autor

Além disso, foi selecionada uma amostra aleatória visando entender melhor o comportamento do preditor:

Gráfico 2 – Curva de predição para Batch: 34 | Paciente: 40340



Fonte: Próprio autor

Observa-se que o modelo acompanha de certa forma a curva real, porém com um atraso. Para tal, ela não usa apenas o último valor real na predição, mas sim pondera o passado recente e o período anterior. Isso pode ser observado nos casos de variação mais brusca ou mudança na direção da tendência. O comportamento visto pode ser comparado aos modelos de predição de séries temporais clássicos, que utilizam a técnica de suavização exponencial para balizar o peso

entre a última informação e a informação de períodos anteriores, através de um parâmetro α chamado de constante de suavização.

Antes de avançar para o próximo passo e enriquecer o modelo usando os dados de proteínas, um ponto observado é a relativa diferença entre o desempenho do modelo entre os dados de treino e teste. Visando evitar um possível *overfitting*, o parâmetro de treinamento foi ajustado para 300 épocas. Como resultado, os dados de treinamento e teste foram os seguintes:

Tabela 4 – Resultados do modelo em Parte 01, métricas MSE e RMSE

Fase do experimento	Resultado do modelo
Treinamento	*Mean Squared Error: 0.0110* *Root Mean Squared Error: 0.1049*
Teste	*Mean Squared Error: 0.0138* *Root Mean Squared Error: 0.1173*

Fonte: Próprio autor

Apesar do resultado no período de treinamento ter piorado, o resultado utilizando os dados de teste foi ligeiramente melhor. Além disso, os resultados de treinamento e teste se mantiveram próximos em relação ao resultado, o que é mais condizente com o comportamento esperado para o modelo, dado a sua expectativa de generalização. O experimento foi refeito e, apesar do valor final oscilar, é possível afirmar que não houve prejuízo para a previsão dos dados de teste com a diminuição das épocas.

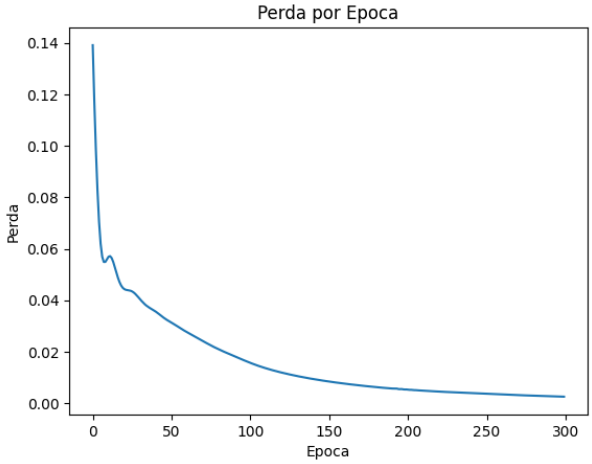
Visando maior compatibilidade com a Parte 2 do experimento, um último ajuste nos dados usados foi feito, alterando o intervalo considerado entre as visitas dos pacientes de 6 meses para 01 ano. Esta mudança trouxe resultados muito próximos dos obtidos com o intervalo de 06 meses, de modo que não se fez necessário expô-los e discuti-los explicitamente.

4.2. Parte 02 - Enriquecimento do modelo usando dados de abundância de proteína

Considerando os ajustes de experimentação expostos no capítulo anterior (seção 3.8), os dados foram novamente sanitizados e tratados para esta etapa. Como resultado, a amostra conta com 83 pacientes divididos em 58 pacientes para treino (70%) e 25 para teste (30%). Além disso, como já exposto, cada paciente (*batch*) possui 4 *time-steps* (meses 0, 12, 24 e 36), havendo um total de 332 amostras únicas. Cada amostra foi enriquecida com as informações de abundância de proteína em um total de 227 *features*. Sendo assim, após tratamento o formato dos dados de input no treinamento é *25 x 3 x 232*.

O modelo foi treinado usando a mesma rede LSTM e um parâmetro de 300 épocas. A métrica de perda MSE em relação às épocas seguiu um comportamento semelhante ao observado na Parte 1, como pode ser observado:

Gráfico 3 - Perda por Época em Parte 02, métrica MSE



Fonte: Próprio autor

Como resultado, considerando apenas o trigésimo-sexto mês, o resultado das métricas MSE e RMSE para o modelo adaptado foi de:

Tabela 5 – Resultados do modelo adaptado em Parte 02, métricas MSE e RMSE

Fase do experimento	Resultado do modelo
Treinamento	*Mean Squared Error: 0.0013* *Root Mean Squared Error: 0.0360*
Teste	*Mean Squared Error: 0.0246* *Root Mean Squared Error: 0.1568*

Fonte: Próprio autor

Observando a grande diferença de resultado entre os dados de treinamento e teste, fica evidente a situação de possível *overfitting* do modelo. Este ocorrido provavelmente é reflexo do cenário para esta parte do experimento, que conta com uma quantidade menor de amostras somado ao aumento expressivo de *features*. Além disso, o valor observado pela métrica MSE neste modelo adaptado não se mostrou superior ao modelo utilizado na parte 1 do experimento. Para aprofundar na comparação, foi calculado também o erro percentual médio considerando o último mês de predição, assim como feito anteriormente. O resultado é apresentado a seguir:

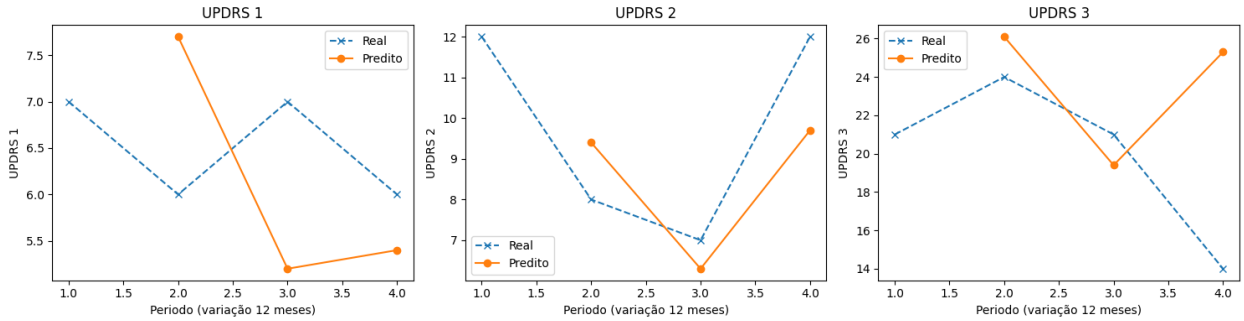
Tabela 6 – Resultados do modelo considerando score original, Parte 02

Score	Erro percentual médio
UPDRS 1	69.2%
UPDRS 2	48.6%
UPDRS 3	68.3%

Fonte: Próprio autor

Apesar de mesma ordem de grandeza, é perceptível a piora de desempenho na predição do modelo adaptado, sobretudo na predição do score UPDRS 3. Da mesma forma, foi selecionada uma amostra aleatória para observar o comportamento do modelo em relação à série histórica

Gráfico 2 – Curva de predição para Batch: 13 | Paciente: 57646



Fonte: Próprio autor

5. CONCLUSÕES

Visando sumarizar a discussão, estão dispostos lado a lado os resultados obtidos na Parte 1, em que foi desenvolvido um modelo para predizer scores usando a série temporal histórica; e na Parte 2, em que o modelo foi adaptado para incluir informações dos níveis de proteína.

Tabela 7 – Resultados sumarizados, Parte 01 e 02

Modelo Inicial - Parte 01	
Fase	**Resultado do modelo**
Treinamento	*MSE: 0.0110*
Teste	*MSE: 0.0138*
Score	**Erro percentual médio**
UPDRS 1	57.80%
UPDRS 2	41.30%
UPDRS 3	29.00%

Modelo Adaptado - Parte 02	
Fase	**Resultado do modelo**
Treinamento	*MSE: 0.0013*
Teste	*MSE: 0.0246*
Score	**Erro percentual médio**
UPDRS 1	69.20%
UPDRS 2	48.60%
UPDRS 3	68.30%

Fonte: Próprio autor

Comparando os resultados entre o modelo inicial e o modelo adaptado, é evidente que não houve uma melhora ao se incluir as informações dos níveis de proteína. Com isso, o modelo pareceu não absorver as relações existentes entre estes níveis e a evolução da doença. Tal fator impediu estudos mais profundos, planejados para o comportamento do modelo e os seus dados.

Embora sejam relações complexas, é precipitado afirmar que o modelo não seja capaz de aprender tais relações. Algumas limitações acima já expostas, como o problema de *overfitting* em um contexto de baixo volume de amostras, acabam por prejudicar o desempenho do experimento. Três ações podem ser tomadas com o objetivo de melhorar a qualidade do modelo, e assim tentar cumprir o objetivo de identificar pistas que ligam a abundância de proteínas à evolução da doença. Estas ficam como recomendação para trabalhos futuros, sendo elas:

- ***Adoção de feature selection*** - A quantidade de amostras frente ao volume de *features* quando se inclui os níveis de proteína se mostra um empecilho ainda maior, considerando a realidade do experimento proposto. Para contornar este problema de dimensionalidade dos dados, uma técnica denominada *feature selection* (em português '*seleção de características*') poderia minimizar o impacto. A técnica consiste em implementar o modelo sucessivas vezes com diferentes subgrupos das *features* disponíveis, e comparar o desempenho entre eles para selecionar as *features* mais relevantes.

- ***Separação das amostras em diferentes clusters*** - Como visto na bibliografia prévia estudada, a doença de Parkinson evolui de forma diferente em cada paciente. Uma forma de evoluir a abordagem seria separar os pacientes em diferentes *clusters* conforme o padrão de evolução visando melhorar a predição do modelo. Além disso, os dados usados no experimento contam grupo controle (isto é, tanto pacientes com doença de Parkinson diagnosticada quanto pacientes saudáveis). Por um lado, conduzir o experimento desta forma permite maior capacidade de generalização, por outro, pode estar afetando a qualidade do modelo na predição.

- ***Experimento envolvendo amostra maior de dados*** - Ficou evidente durante o experimento a necessidade de um volume maior de dados com o objetivo de capturar melhor o aprendizado e evitar o problema de *overfitting*. Porém, a captura destes dados depende do envolvimento de diversos voluntários, de procedimentos específicos, profissionais capacitados e um intervalo de tempo suficiente para que a amostra se torne robusta. Neste contexto complexo de acesso aos dados, a busca por parcerias junto às instituições especializadas na área e o envolvimento de agentes referência na sociedade são de grande importância para contornar o problema.

6. REFERÊNCIAS

POEWE, Werner ; SEPPI, Klaus; TANNER, Caroline M.; HALLIDAY, Glenda M. Halliday, Patrik Brundin, Jens Volkmann, Anette-Eleonore Schrag & Anthony E. Lang ; **Parkinson Disease**. Nature Reviews Disease Primers volume 3, artigo 17013, 2017.

MINISTÉRIO DA SAÚDE. **Dia Mundial de Conscientização da Doença de Parkinson**. 2018. Disponível em: <https://bvsms.saude.gov.br/11-4-dia-mundial-de-conscientizacao-da-doenca-de-parkinson-avancar-melhorar-educar-colaborar/>. Acesso em: 24 ago. 2023.

WILLIS, A.W. ; ROBERTS, E. ; BECK, J.C. et al. ; **Incidence of Parkinson disease in North America**. Nature Partner Journals. Parkinsons Disease 8-170, 2022.

EVANS, W. J.; ROSENBERG, I. H. **Biomarkers: The 10 keys to prolonging vitality**. Pegasus Books, 2017.

GANGULY U.; SINGH, S.; PAL, S.; PRASAD, S.; AGRAWAL, B.K.; SAINI, R.V.; CHAKRABARTI, S. **Alpha-Synuclein as a Biomarker of Parkinson's Disease: Good, but Not Good Enough**. Frontiers in Aging Neuroscience, 2021.

EMAMZADEH, F. N.; SURGUCHOV, A. **Parkinson's disease: biomarkers, treatment, and risk factors**. Frontiers in Aging Neuroscience, 2018.

RUSSEL, Stuart; NORVIG, Peter. **Artificial Intelligence: A Modern Approach 4ª ed**. , Prentice Hall, 2020.

HAYKIN, Simon. **Neural Networks and Learning Machines. 3.ed.** Ontario, Canadá: Pearson Education, Inc. 2009.

GOETZ, Christopher; POEWE, Werner; STERN, Matthew B.; FAHN, Stanley; MARTIN, P.; STEBBINS, Glenn T.; SAMPAIO, Cristina; TILLEY, Barbara. **Revision of the Unified Parkinson's Disease Rating Scale**. International Parkinson and Movement Disorder Society, 2008

HOCHREITER, S.; SCHMIDHUBER, J. **Long Short-Term Memory**. Neural Computation, 1997.

BIRKENBIHL, C.; AHMAD, A.; MASSAT, N.J.; et al. **Artificial intelligence-based clustering and characterization of Parkinson's disease trajectories**. Nature Scientific Reports, 2023.

RASTEGAR, Ahmadi D.; HO, N.; HALLIDAY, G.M. et al. **Parkinson's progression prediction using machine learning and serum cytokines**. Nature Partners Journal Parkinsons Disease, 2019.

VATANSEVER, S.; SCHLESSINGER, A.; WACKER, D.; et al. **Artificial intelligence and machine learning-aided drug discovery in central nervous system diseases**: State-of-the-arts and future directions.
Medicinal Research Reviews Volume 41, issue 3. Wiley, 2020.