DETECÇÃO DE EXOPLANETAS UTILIZANDO LSTM

EXOPLANET DETECTION USING LSTM

Silverio, Dênis[a]. Mastrodomenico, Julio Cesar[a].

[a]Centro Universitário Facens - Sorocaba, SP, Brasil
denis.a.silverio@gmail.com, jczanotto@yahoo.com.br

Submetido em: 04 Nov. de 2023. Aceito em: 04 Nov. de 2023.

RESUMO

Nos últimos anos, a astronomia tem experimentado um notável avanço na capacidade de coletar dados, impulsionado pelo desenvolvimento de tecnologias avançadas de observação e sensores. Observatórios espaciais, como o Telescópio Espacial Hubble e o Kepler, juntamente com telescópios terrestres aprimorados, têm gerado uma quantidade massiva de informações. No entanto, esse progresso na aquisição de dados também apresenta o desafio de processar e analisar esses volumes imensos de informações de maneira eficiente.

Este artigo explora a crescente aplicação de algoritmos de inteligência artificial (IA) na extração de conhecimento valioso a partir de dados astronômicos. Em particular, investigamos o uso de uma rede neural recorrente para detectar possíveis exoplanetas por meio da análise das perturbações nas curvas de luz de suas estrelas hospedeiras, com base em dados da missão Kepler. Esta pesquisa visa avaliar a viabilidade da utilização dessa técnica como uma solução para acelerar o processamento da montanha de dados gerados diariamente.

A integração da IA na astronomia não apenas impulsiona o progresso científico, mas também contribui significativamente para lidar com a crescente demanda por análise de dados em grande escala. A exploração dessas tecnologias é essencial para desvendar os mistérios do universo de forma mais eficaz e eficiente.

Palavras-chave: Processamento de Dados. Missão Kepler. Redes Neurais Recorrentes. LSTM. Classificação. Exoplanetas.

ABSTRACT

In recent years, astronomy has witnessed an extraordinary increase in data collection capabilities, driven by the development of advanced observation technologies and sensors. Space observatories, such as the Hubble Space Telescope and Kepler, along with enhanced ground-based telescopes, have generated a massive amount of information. However, this advancement in data collection also brings the challenge of processing and analyzing these vast quantities of information effectively.

This article explores the growing application of artificial intelligence (AI) algorithms to extract valuable knowledge from astronomical data. Specifically, we investigate the use of a recurrent neural network to detect potential exoplanets by analyzing perturbations in the light curves of their host stars, based on data from the Kepler mission. This research aims to assess the feasibility of using this technique as a solution to accelerate the processing of the daily deluge of data.

The integration of AI into astronomy not only drives scientific progress but also significantly contributes to addressing the increasing demand for large-scale data analysis. Exploring these technologies is essential for unraveling the mysteries of the universe more effectively and efficiently.

Keywords: Data Processing. Kepler Mission. Recurrent Neural Network. LSTM. Classification. Exoplanets.

1. INTRODUÇÃO

As estrelas sempre foram um objeto de fascínio para a humanidade, inspirando lendas e mitos, e influenciando nossa cultura e ciência. Constelações foram criadas, calendários precisos auxiliaram na nossa agricultura, navegadores exploraram a imensidão oceanos com base em seu conhecimento sobre as estrelas. Desde então, adquirimos um vasto conhecimento científico e, nos tempos atuais, estamos testemunhando o início de uma exploração espacial sem precedentes, gerando uma quantidade absurda de dados. Satélites nos enviam constantemente informações sobre estrelas, exoplanetas, supernovas, buracos negros e tudo isso precisa ser processado, analisado por meio de diversas técnicas e processos que demandam muito esforço dos cientistas. (BARRY et al., 2019)

Felizmente, a humanidade também deu saltos tecnológicos significativos nas últimas décadas, o que resultou em um substancial aumento no poder de processamento com novas técnicas e algoritmos como a inteligência artificial que surgem a cada ano.

Neste trabalho de pós-graduação, será abordada uma das áreas da astronomia que visa à descoberta de planetas fora do nosso sistema solar. Isso será alcançado por meio da análise dos dados da missão Kepler, que foram disponibilizados ao público pela Agência Espacial Americana (NASA, 2023D).

Esses dados serão processados e utilizados para treinamento em uma rede neural recorrente do tipo LSTM, com o objetivo de desenvolver um modelo capaz de detectar possiveis planetas orbitando-as, servindo como uma pré-análise para mais agilidade do processo. Isso será feito com base na técnica conhecida como trânsito planetário como fonte das sequencias de séries temporais como entrada do modelo de classificação.

2. MISSÃO KEPLER

Em 7 de março de 2009, a NASA lançou a missão Kepler, seu nome foi em homenagem ao astrônomo alemão Johannes Kepler (1571-1630), um dos primeiros a estudar a órbita dos planetas. O telescópio espacial que revolucionou a nossa compreensão do universo, tinha como objetivo encontrar planetas semelhantes à Terra em órbita de outras estrelas.

O Kepler entrou em operação em 11 e maio de 2009 e ficou observando uma região de 100 graus quadrados da área do céu observável da Terra, monitorando a variação de brilho de cerca de 200.000 estrelas.

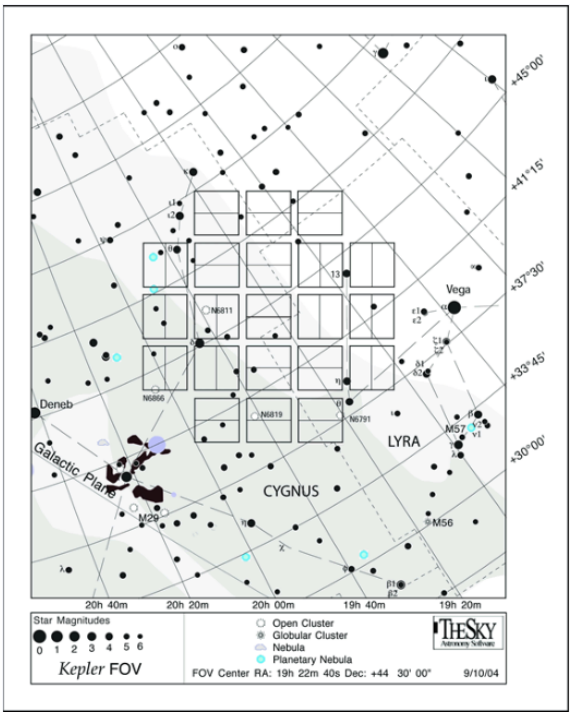
Ao observar a variação de brilho das estrelas, o Kepler era capaz de detectar a pequena variação de brilho que é uma indicação da presença de um planeta.

As descobertas da missão Kepler tiveram um impacto profundo na nossa compreensão do universo. Elas mostraram que planetas semelhantes à Terra são comuns dentre os mais de 2600 exoplanetas encontrados pela missão, fornecendo informações importantes sobre as órbitas e as características físicas dos exoplanetas. Essas informações ajudarão os astrônomos a compreender melhor a formação e a evolução dos planetas.

(NASA, 2023C)

A Figura 1 demonstra a região de monitoramento do satelite Kepler

Figura 1 – Constelação Kepler.



Fonte: (STScI, 2023A).

Todos os dados da missão Kepler estão disponiveis no portal MAST (MAST, 2023), um serviço mantido pela Space Telescope Science Institute (STScI, 2023A) em parceria com a NASA, para a distribuição de dados de satelites de diversas missões.

2.1 Dados da Missão Kepler

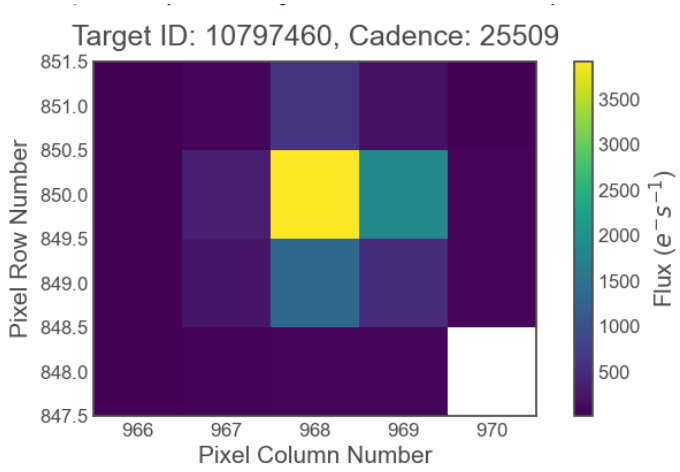
Os dados da missão Kepler foram preservados em um formato de arquivo criado especificamente para atender às necessidades da astronomia, conhecido como Sistema de Transporte de Imagens Flexíveis (NASA, 2023B). Esse formato se mostra altamente eficaz para a retenção, análise e compartilhamento de dados que se organizam na forma de matrizes multidimensionais, como imagens 2D ou tabelas, sendo também capaz de conter informações essenciais relacionadas à fotometria, calibração espacial e outros ajustes fundamentais.

No contexto da missão Kepler, os dados primordiais consistem em sequências de imagens que capturam os pixels circundantes de estrelas de interesse. Isso é feito em intervalos regulares de 30 minutos a 1 minuto, criando, assim, uma série temporal que representa a variação do fluxo luminoso emitido por cada estrela ao longo do tempo.

A cada 90 dias, o satélite executava manobras para assegurar que seus painéis solares permanecessem continuamente direcionados para o sol. Em virtude desse processo, os dados coletados pelo satélite são divididos em quartis, sendo estes intervalos projetados para permitir as devidas correções de orientação e garantir o bom funcionamento da missão entre esses períodos de ajuste.

A Figura 2 demonstra o primeiro quadro da sequência de fotos da estrela KIC 107974601 no quartil número 7.

Figura 2 – Pixel File Frame.



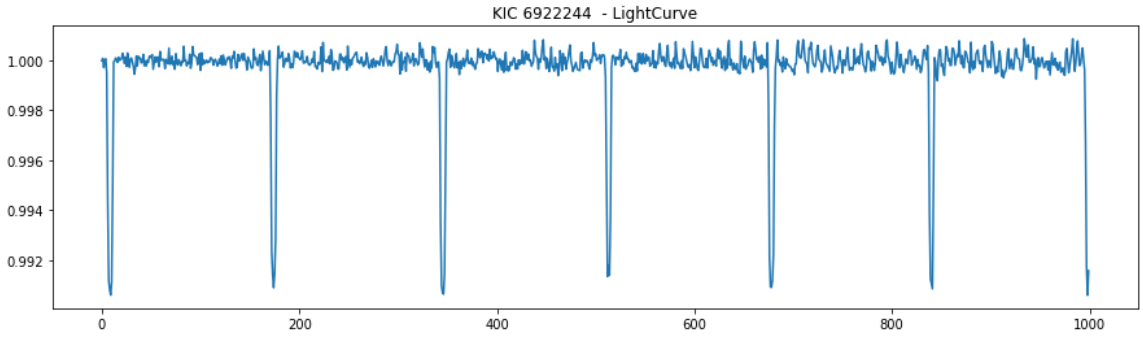
Fonte: Elaborado pelo autor.

3. FOTOMETRIA DE TRANSITO PLANETÁRIO

A fotometria de trânsito planetário envolve medir cuidadosamente o fluxo luminoso de uma estrela ao longo do tempo para detectar variações causadas pelos trânsitos de planetas. Quando um planeta passa na frente da estrela, ocorrem pequenas quedas na luminosidade, semelhantes a pequenos eclipses. Medindo a profundidade e a duração dessas quedas, cientistas obtêm informações cruciais, como o período orbital do planeta, seu tamanho em relação à estrela, sua posição na zona habitável e a possível presença de uma cauda de poeira, semelhante à de um cometa (BUDAJ, 2013)

A Figura 3 ilustra a curva de luz de uma estrela que é conhecida por ter planetas em órbita ao redor dela.

Figura 3 – Curva de Luz Da Estrela KIC-6922244



Fonte: Elaborado pelo autor.

4. COLETA DOS DADOS

O conjunto de dados utilizado foi criado de forma personalizada, a fim de atender às exigências específicas da pesquisa. Não existiam conjuntos de dados prontos contendo informações sobre o fluxo luminoso de estrelas organizadas por pixels e curvas de luz. Por essa razão, cada arquivo estelar foi adquirido, processado e os dados foram subdivididos em segmentos para satisfazer as exigências do modelo de classificação selecionado.

4.1 Nasa Exoplanet Archive

O conjunto de dados utilizado foi construído a partir de dados fornecidos pelo portal NASA Exoplanet Archive(NASA, 2023A), um catálogo eletrônico que concentra informações sobre todos os exoplanetas conhecidos. As informações disponíveis abrange dados importantes, como o tamanho, massa, órbita e temperatura dos exoplanetas, juntamente com detalhes sobre as estrelas que os hospedam.

Para a treinamento do classificador, tivemos que criar o conjunto de dados com base nas estrelas previamente analisadas por astrônomos. Essa escolha se deve ao fato de que nosso classificador utiliza um método de aprendizado supervisionado. Para fazer isso, utilizamos os dados da tabela "cumulative_koi". Nessa tabela, a coluna "koi_disposition"nos informa a situação atual da estrela, enquanto a coluna "koi_score"nos fornece uma medida de confiabilidade nessa classificação.

A classificação encontrada na coluna "koi_disposition"pode ser de três tipos principais:

- **CONFIRMED**: Essa classificação é atribuída quando evidências sólidas de planetas orbitando a estrela foram encontradas. Em outras palavras, há confirmação de exoplanetas nesse sistema.
- **FALSE POSITIVE**: Essa classificação é utilizada quando, em uma análise inicial, a estrela parece ser uma candidata a abrigar planetas (classificação que pode ser encontrada na coluna "koi_pdisposition"). No entanto, análises posteriores mais minuciosas e criteriosas revelam que, na verdade, não há planetas em órbita.
- **CANDIDATE**: Nesse caso, a classificação "CANDIDATE"é atribuída quando é necessária uma análise adicional por astrônomos para determinar se a estrela realmente abriga exoplanetas.

4.2 Download Das Curvas De Luz

Conforme mencionado na seção 2, os dados da missão Kepler foram armazenados em arquivos no formato FIT e estão acessíveis ao público por meio do portal MAST (MAST, 2023). Existem várias maneiras de transferir esses arquivos, conforme explicado no guia fornecido pelo próprio portal (STSCI, 2023B).

Uma dessas opções é usar uma biblioteca Python conhecida como LightKurve, desenvolvida pela LightKurve Collaboration e outros pesquisadores (LIGHTKURVE COLLABORATION et al., 2018). Essa biblioteca foi criada para facilitar a

busca, download e análise de dados da missão Kepler. Ela fornece um conjunto de ferramentas e funções que simplificam o acesso e manipulação desses dados, tornando o processo mais eficiente e amigável para os pesquisadores.

Com a biblioteca LightKurve, é possível realizar o download de todos os trimestres de dados de uma estrela, ou um trimestre especifico utilizando a sua identificação Kepler ID.

5. PRÉ-PROCESSAMENTO

Para cada trimetre de dados baixados com a biblioteca LightKurve, foi necessário realizar uma série de processamentos para ajustar e padronizar os dados antes de proceder com o treinamento.

5.1 Outliers

Devido à extrema sensibilidade das cameras do telescópio Kepler, é comum a presença de ruído nos dados, e que em alguns casos, podem apresentar valores muito discrepantes que se destacam consideravelmente dos demais. Para tratar esses pontos atípicos, aplicamos a identificação de outliers pelo método de amplitude interquartil.

No entanto, no contexto das curvas de luz, não é apropriado remover os outliers do limite inferior. Isso ocorre porque, nesse cenário, os vales na série temporal causados pela transição de um planeta diante da estrela podem ser erroneamente classificados como outliers.

$$Q1 = \text{Percentil}_{25}(x)$$
$$Q3 = \text{Percentil}_{75}(x)$$
$$IQR = Q3 - Q1$$

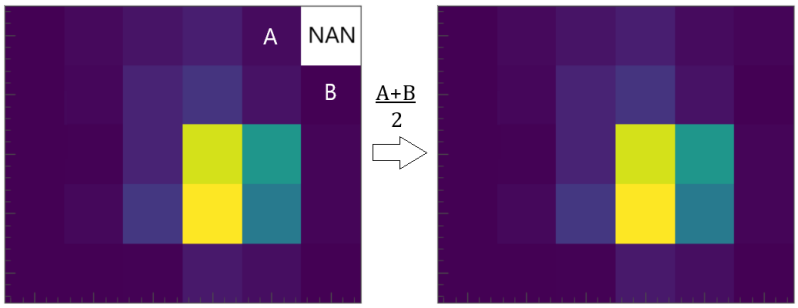
$$h\_bound = Q3 + (1.5 * IQR)$$

5.2 Pixels Nulos

É comum encontrar valores nulos em alguns pixels da serie temporal do fluxo luminoso. Para lidarmos com essa situação, adotamos a abordagem de considerar a média dos pixels adjascentes para cada pixel de valor nulo.

A Figura 4 exemplifica a tratativa para pixels com valores nulos em um frame.

Figura 4 – Tratativa do Pixel Nulo



Fonte: Elaborado pelo autor.

5.3 Padronização do Frame

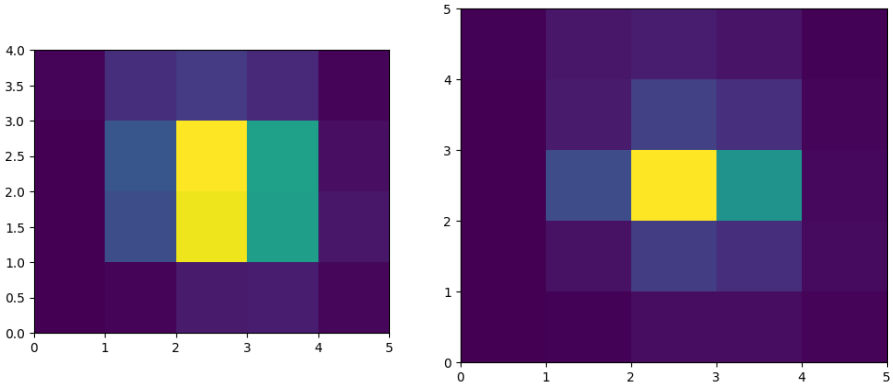
Outro problema frequente, foi à falta de padronização no tamanho das matrizes que representam os frames de fluxo luminoso. Para resolver essa questão, estabelecemos um tamanho padrão de 5x5 para os frames e desenvolvemos um algoritmo para ajustar as dimensões dos frames conforme necessário.

Qualquer frame que não se encaixasse nesse formato foi redimensionado para atender a esse padrão. Utilizamos o método interp2D da biblioteca SciPy para realizar esse redimensionamento. Essa função é projetada especificamente para interpolar dados em duas dimensões, permitindo-nos estimar valores entre os pontos de dados observados. (SciPy, 2023)

Essa abordagem nos permitiu padronizar os frames sem perder informações significativas sobre o fluxo luminoso de cada frame, garantindo que os dados permaneçam consistentes.

A Figura 5 demonstra o redimensionamento de um frame de fluxo luminoso.

Figura 5 – Redimensionamento 4x5 para 5x5



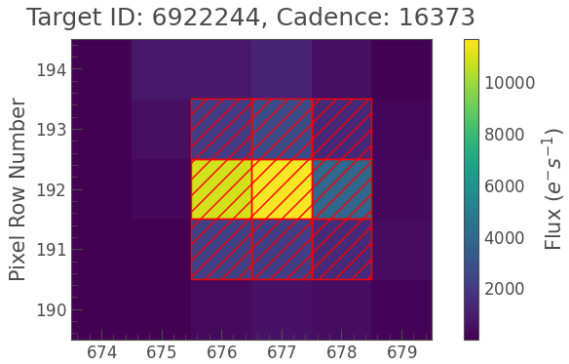
Fonte: Elaborado pelo autor.

Com os frames tratados e redimensionados, a serie temporal foi separada em segmentos de 1000 registros cada, para padronização das sequências. Para cada pixel foi criada uma coluna dedicada que varia do "px0"até "px24". Além disso, foi incluída uma coluna chamada "sap", que representa a Simple Aperture Photometry (Fotometria de Abertura Simples) que a biblioteca lightkurve já fornece.

Fotometria de Abertura é o ato de somar os valores de todos os pixels em uma abertura pré-definida em função do tempo. Ao escolher cuidadosamente o formato da máscara de abertura, você pode evitar contaminantes próximos ou melhorar a intensidade do sinal específico que está tentando medir em relação ao fundo. (NASA, 2023E)

A Figura 6 demonstra a região da máscara de abertura.

Figura 6 – Mascara de Abertura



Fonte: Elaborado pelo autor.

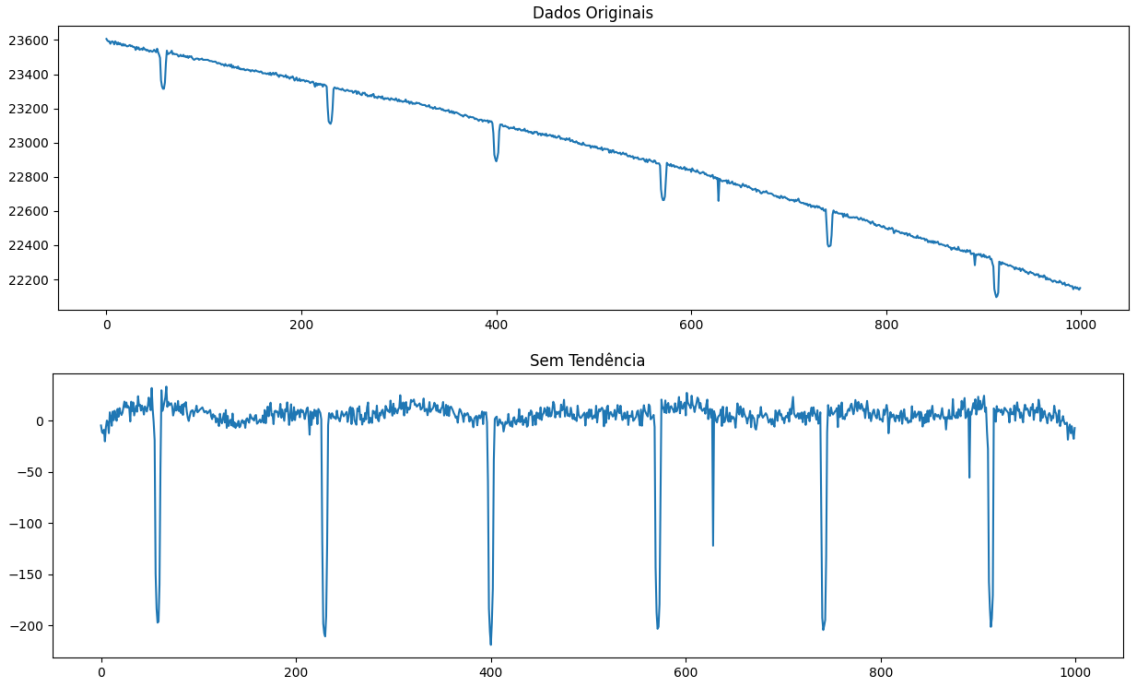
5.4 Remoção de Tendências

Devido à dinâmica do satélite Kepler, as séries temporais geradas podem exibir tendências de aumento ou diminuição ao longo do tempo em cada pixel. Para mitigar essas tendências, foi empregado o uso de técnicas de remoção de tendência.

Com a função polyfit da biblioteca NumPy, é ajustada um polinômio (tendência) de grau 3, aos valores da serie temporal de cada pixel do dataset, obtendo os coeficientes dessa tendência. A curva polinomial de tendência então é calculada e subtraida dos dados originais.

A Figura 7 demonstra a remoção de tendência em uma curva de luz

Figura 7 – Remoção de Tendência



Fonte: Elaborado pelo autor.

6. DATASET

Foram processados dados de um total de 550 estrelas, sendo que cada uma dessas estrelas gerou aproximadamente de 10 a 18 quartis. Esses quartis foram então divididos em sequências de 1000 registros cada, resultando em um volume de dados que ultrapassa a marca de 17 milhões de registros.

A Figura 8 mostra a forma final do dataset.

Figura 8 – Dataset

	px0	px1	px2	px3	px4	...	px20	px21	px22	px23	px24	sap	idx	disposition
0	5.872670	6.658888	26.318502	86.284035	90.025238	...	0.872324	7.654994	68.446404	164.286911	66.486214	1.000445	0	CONFD
1	5.071711	4.494245	29.132658	86.419884	86.229446	...	1.493892	8.369994	66.311485	165.864258	62.561733	0.999429	0	CONFD
2	4.405221	6.581491	29.775621	84.354027	88.305283	...	0.110620	9.016906	69.066322	165.980881	63.982121	0.999994	0	CONFD
3	6.198782	6.737494	30.762922	85.991234	87.175430	...	1.910010	9.172617	67.929398	164.724930	63.526905	0.999703	0	CONFD
4	5.469162	6.695978	29.095543	85.523308	85.683426	...	3.339775	9.137881	66.981316	163.549210	65.630074	1.000209	0	CONFD

Fonte: Elaborado pelo autor.

7. TREINAMENTO

As estrelas do dataset foram categorizadas como "CONFD"quando a existência de planetas em órbita foi confirmada e como "FALSE"quando não foram encontrados planetas orbitando a estrela. Antes do treinamento, todos os valo-

res nas colunas de fluxo luminoso foram padronizados para assegurar que as características tenham uma escala adequada.

O classificador escolhido foi um algoritmo da arquitetura de rede neural recorrente conhecida como LSTM (Long Short-Term Memory), que é amplamente empregada em deep learning. A escolha do LSTM se deve à sua arquitetura que incorpora conexões de feedback, possibilitando o processamento de sequências completas de dados em vez de pontos de dados individuais.

As sequências foram separadas em uma divisão 80 20, onde 80% do dataset foi utilizado para o treinamento e 20% para a validação. Após diversos testes, chegamos em uma configuração da rede neural com os seguintes parâmetros:

A Tabela 1 mostra as configurações utilizadas no classificador LSTM.

Tabela 1 – Configuração do Classificador LSTM.

Configuração	Valor
Células de Memória Ocultas	512
Número de Camadas	3
Taxa de Aprendizado	0.0001
Dropout	20%
Função de Custo	Cross Entropy Loss
Número de Épocas	150
Tamanho do Batch	64
Checkpoint Model	min val loss

Fonte: Elaborado pelo autor.

O classificador foi configurado para salvar o modelo em cada iteração do treinamento, desde que o cálculo da perda na validação atinja o menor valor. Isso assegura que o modelo treinado seja o que possui a menor perda, evitando assim problemas de overfitting que possam surgir durante o treinamento.

8. RESULTADOS

No geral, os resultados mostraram que o modelo tem um desempenho bastante sólido, com boas métricas de precisão e recall para ambas as classes.

Tabela 2 – Relatório da Classificação

	Precision	Recall	F1-Score	Support
CONFD	86%	97%	91%	1742
FALSE	96%	84%	90%	1660
Accuracy	-	-	91%	3402
Macro AVG	91%	90%	90%	3402
Weighted AVG	91%	91%	90%	3402

Fonte: Elaborado pelo autor.

A precisão para ambas as instâncias tem uma alta taxa de acerto na previsão, além de uma alta taxa de recall, o que significa que o modelo tem uma alta capacidade de identificar corretamente cada uma das instâncias. Podemos observar também que precisão para "FALSE"é 10% maior que para "CONFD", porém o recall de "CONFD"é 13% maior do que o de "FALSE". Isso deve-se ao fato de que as classificações para "FALSE", tiveram um número maior de falsos positivos como mostra a tabela confusão abaixo.

A precisão é alta para ambas as classes, indicando que o modelo tem uma taxa significativa de previsões corretas para ambas as instâncias. Além disso, o recall é igualmente alto, o que significa que o modelo é capaz de identificar adequadamente a maioria das instâncias de ambas as classes.

É importante notar que a precisão para a classe "FALSE"é ligeiramente maior do que para a classe "CONFD,"com uma diferença de 10%. No entanto, o recall para a classe "CONFD"supera o recall da classe "FALSE"em 13%. Essa discrepância pode ser atribuída à presença de falsos positivos na classificação da classe "FALSE,"como evidenciado pela matriz de confusão.

A Tabela 3 mostra os resultados de classificação em uma matrix de confusão.

Tabela 3 – Matriz De Confusão

CONFD REAL	1683	59
FALSE REAL	264	1396
	CONFD PREVISTO	FALSE PREVISTO

Fonte: Elaborado pelo autor.

Em resumo, os resultados refletem um equilíbrio entre precisão e recall para ambas as classes como demonstrado no F1-Score de 91%.

9. CONCLUSÃO

A LSTM é uma arquitetura de rede neural recorrente que se mostrou muito eficaz na classificação das series temporais de fluxo de luz, com um desempenho muito bom demonstrados pelas metricas.

A arquitetura de rede neural recorrente LSTM se revelou altamente eficaz na classificação das séries temporais de fluxo de luz da missão Kepler, demonstrando um excelente desempenho, conforme evidenciado pelas métricas de avaliação. Os resultados obtidos demonstram a capacidade da LSTM em lidar com dados sequenciais, fornecendo classificações precisas e confiáveis, o que a torna uma escolha muito interessante para problemas de classificação de séries temporais, como os abordados neste estudo.

REFERÊNCIAS

BARRY, RICHARD K; BABU, JOGESH G; BAKER, JOHN G; FEIGELSON, ERIC D; KAUR, AMANPREET; KOGUT, ALAN J; KRAEMER, STEVEN B; MASON, JAMES P; MEHROTRA, PIYUSH; OLMSCHENK, GREGORY et al. Advanced astrophysics discovery technology in the era of data driven astronomy. *arXiv preprint arXiv:1907.10558*, 2019.

BUDAJ, J. Light-curve analysis of kic 12557548b: an extrasolar planet with a comet-like tail. *Astronomy Astrophysics*, 557, September 2013. doi: 10.1051/0004-6361/201220260.

Lightkurve Collaboration; Cardoso, J. V. D. M.; Hedges, C.; Gully-Santiago, M.; Saunders, N.; Cody, A. M.; Barclay, T.; Hall, O.; Sagear, S.; Turtelboom, E.; Zhang, J.; Tzanidakis, A.; Mighell, K.; Coughlin, J.; Bell, K.; Berta-Thompson, Z.; Williams, P.; Dotson, J.; Barentsen, G. Lightkurve: Kepler and TESS time series analysis in Python. Astrophysics Source Code Library, December 2018.

MAST. Portal mast. Disponível em: `https://mast.stsci.edu/portal/Mashup/Clients/Mast/Portal.html`, Acesso em 20/10/2023, 2023.

NASA. Nasa exoplanet archive. Disponível em: `https://exoplanetarchive.ipac.caltech.edu/index.html`, Acesso em 20/10/2023, 2023a.

NASA. Formato fits. Disponível em: `https://fits.gsfc.nasa.gov/fits_home.html`, Acesso em 20/10/2023, 2023b.

NASA. Missão kepler. Disponível em: `https://science.nasa.gov/mission/kepler/in-depth/`, Acesso em 20/10/2023, 2023c.

NASA. Agencia espacial americana. Disponível em: `https://science.nasa.gov/`, Acesso em 20/10/2023, 2023d.

NASA. Sap. Disponível em: `https://heasarc.gsfc.nasa.gov/docs/tess/LightCurve-object-Tutorial.html`, Acesso em 20/10/2023, 2023e.

SCIPY. Interpolation 2d. Disponível em: `https://docs.scipy.org/doc/scipy/reference/generated/scipy.interpolate.interp2d.html`, Acesso em 20/10/2023, 2023.

STSCI. Constelação kepler. Disponível em: `https://archive.stsci.edu/missions-and-data/kepler`, Acesso em 20/10/2023, 2023a.

STSCI. Guia do portal mast. Disponível em: `https://outerspace.stsci.edu/display/MASTDOCS/Portal+Guide`, Acesso em 20/10/2023, 2023b.