UNIVERSIDADE DE SÃO PAULO
Instituto de Ciências Matemáticas e de Computação

Aplicação de Aprendizado Profundo para detecção de Exoplanetas

Marcelo Daudt

Monografia - MBA em Ciência de Dados (CEMEAI)

ICMC USP SÃO CARLOS

Marcelo Daudt

Aplicação de Aprendizado Profundo para detecção de Exoplanetas

Monografia apresentada ao Centro de Ciências Matemáticas Aplicadas à Indústria do Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo - ICMC/USP, como parte dos requisitos para obtenção do título de Especialista em Ciências de Dados.

Área de concentração: Ciências de Dados

Orientador: Prof. Dr. Afonso Paiva Neto

Versão original

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)

Daudt, Marcelo
D238a Aplicação de Aprendizado Profundo para detecção de Exoplanetas / Marcelo Daudt; orientador Afonso Paiva Neto. -- São Carlos, 2023.
67 p.

Trabalho de conclusão de curso (MBA em Ciência de Dados) -- Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, 2023.

1. Ciência de Dados. 2. Aprendizado Profundo. 3. Exoplanetas. 4. Rede Neurais Convolucionais. 5. Fotometria. I. Paiva Neto, Afonso, 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

Marcelo Daudt

Application of Deep Learning for Exoplanet detection

Monograph presented to the Centro de Ciências Matemáticas Aplicadas à Indústria do Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo - ICMC/USP, as part of the requirements for obtaining the title of Specialist in Data Science.

Concentration area: Data Science

Advisor: Prof. Dr. Afonso Paiva Neto

Original version

São Carlos

2023

Folha de aprovação em conformidade
com o padrão definido
pela Unidade.
No presente modelo consta como
folhadeaprovacao.pdf

Dedico este trabalho a minha querida e amada companheira Tamara, que tanto me apoiou e me incentivou a realizá-lo. Também dedico este trabalho ao meu querido amigo e irmão André, que nos deixou há pouco tempo, mas que muito me ensinou ao longo da sua vida, principalmente a ter coerência nas minhas ações e pensamentos.

AGRADECIMENTOS

Primeiramente, agradeço muito ao professor Afonso Paiva Neto pelas suas valiosas orientações e apoio, fundamentais para o desenvolvimento e conclusão deste trabalho.

Aos professores, funcionários e monitores do Centro de Ciências Matemáticas Aplicadas à Indústria, CEPID-CeMEAI, e do Instituto de Ciências Matemáticas e de Computação da Universidade de São Paulo, ICMC-USP.

Agradeço também aos meus queridos amigos de infância, André (in memorian), Alexandre, Carlito (in memorian), Douglas, Estevan, Guilherme e Serginho, que sempre me apoiaram e incentivaram em diversas etapas importantes da minha vida.

Agradeço imensamente a minha querida e amada companheira Tamara, por toda sua paciência, compreensão, carinho e amor.

E por fim, agradeço mais uma vez a minha mãe e professora Silvia, pelos seus ensinamentos, pela sua dedicação e paciência, em casa e na escola, que fizeram de mim o homem que sou hoje.

"Statistics, the most important science in the whole world: for upon it depends pratical application of every other science and of every art."

Florence Nightingale

RESUMO

DAUDT, M. **Aplicação de Aprendizado Profundo para detecção de Exoplanetas**. 2023. 67p. Monografia (MBA em Ciências de Dados) - Centro de Ciências Matemáticas Aplicadas à Indústria, Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos, 2023.

Atualmente, com a nova geração de telescópios, construídos em solo ou na órbita terrestre, estamos registrando grandes quantidades de sinais (dados) em diversas frequências do espectro eletromagnético conhecido. E como ocorre em diversas áreas do conhecimento, está sendo acumulado um grande volume dados, colocando a Astronomia na era do Big Data. Porém, o problema é que todo esse repositório de dados espaciais, onde existe um crescimento exponencial em volume e variedade de dados, não consegue ser processado, analisado e interpretado com as ferramentas e técnicas tradicionais, transformando-se assim em informação e conhecimento, como, por exemplo, na busca e caracterização de exoplanetas, que são planetas fora do nosso Sistema Solar. Os exoplanetas podem ser identificados através de técnicas de observações de corpos celestes como o Método de Trânsito Planetário, onde são analisados os sinais de luz que chegam até nós, procurando pequenas quedas de brilho de uma estrela quando um planeta passa (transita) na sua frente. Diante deste contexto, se faz cada vez mais necessário o desenvolvimento de novas tecnologias e aplicações capazes de automatizar a análise deste enorme volume de dados. Este trabalho tem como objetivo, sob a perspectiva da Ciência de Dados, estudar um modelo de Aprendizado Profundo, o AstroNet, que é uma Rede Neural Convolucional que pode ser treinada para reconhecer exoplanetas nas leituras de luz registradas durante as missões do Telescópio Espacial Kepler, conhecido como "o caçador de exoplanetas".

Palavras-chave: Exoplaneta. Ciência de Dados. Aprendizado Profundo. Redes Neurais Convolucionais. Fotometria.

ABSTRACT

DAUDT, M. **Application of Deep Learning for Exoplanet detection**. 2023. 67p. Monograph (MBA in Data Sciences) - Centro de Ciências Matemáticas Aplicadas à Indústria, Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos, 2023.

Currently, with the new generation of telescopes, built on the ground or in Earth orbit, we are recording large amounts of signals (data) at different frequencies of the known electromagnetic spectrum. And as in many areas of knowledge, a large volume of data is being accumulated, placing Astronomy in the era of Big Data. However, the problem is that this entire repository of spatial data, where there is an exponential growth in volume and variety of data, cannot be processed, analyzed and interpreted with traditional tools and techniques, thus becoming in information and knowledge, as, for example, in the search and characterization of exoplanets, which are planets outside our Solar System. Exoplanets can be identified through techniques for observing celestial bodies such as the Planetary Transit Method, where the light signals that reach us are analyzed, looking for small drops in brightness of a star when a planet passes (transits) in front of it. Given this context, it is increasingly necessary to develop new technologies and applications capable of automating the analysis of this huge volume of data. This work aims, from a Data Science perspective, to study a Deep Learning model, the AstroNet, which is a Convolutional Neural Network that can be trained to recognize exoplanets in the light readings recorded during the Kepler Space Telescope missions, known as "The Exoplanet Hunter".

Keywords: Exoplanet. Data Science. Deep Learning. Convolutional Neural Network. Photometry.

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

AFP	Astrophysical False Positive
BIC	Bayesian Information Criterion
BNN	Bayesian Neural Networks
CeMEAI	Centro de Ciências Matemáticas Aplicadas à Indústria
CEPID	Centro de Pesquisa, Inovação e Difusão
CNN	Convolutional Neural Network
CSV	Comma-Separated-Values
DCGAN	Deep Convolutional Generative Adversarial Network
DCNN	Deep Convolutional Neural network
DNN	Deep Neural Network
FITS	Flexible Image Transport System
GP	Gaussian Process
HST	Hubble Space Telescope
IA	Inteligência Artificial
IUE	International Ultraviolet Explorer
JWST	James Webb Space Telescope
MAST	Mikulski Archive for Space Telescopes
MIT	Massachusetts Institute of Technology
ML	Machine Learning
NASA	National Aeronautics and Space Administration
PC	Planet Candidate
RF	Random Forest
RNA	Rede Neural Artificial
NTP	Nontransiting Phenomenon

STScI	Space Telescope Science Institute
TCE	Threshold-Crossing Events
TESS	Transiting ExoPlanet Survey Satellite
UNK	Unknown
USP	Universidade de São Paulo

LISTA DE SÍMBOLOS

δ	Letra grega minúscula Delta
λ	Letra grega minúscula Lambda

SUMÁRIO

1 INTRODUÇÃO

Uma das grandes questões da humanidade, e por consequência, objeto de estudo da Astronomia, é saber se estamos ou não sozinhos no Universo. Com o propósito de mapear e encontrar exoplanetas, que são planetas existentes além do nosso **Sistema Solar**, foi construído o **Telescópio Espacial Kepler**, conhecido como "*o caçador de exoplanetas*". Outras missões se seguiram ao **Kepler**, como o **Transiting ExoPlanet Survey Satellite** (TESS) e mais recentemente o **James Webb Space Telescope** (JWST).

A identificação de um exoplaneta se dá através de um método de observação de corpos celestes conhecido como método de trânsito planetário, onde são analisados os sinais de luz que chegam até nós, procurando por minúsculas quedas no brilho de uma estrela quando um planeta passa na frente dela (KEPLER, 2017). Ainda com essa análise dos sinais de luz, também é possível detectarmos a composição do planeta, o seu tamanho e a sua distância da estrela hospedeira.

A ideia da presente pesquisa surge durante a revisão bibliográfica sobre o tema, principalmente após a leitura de dois trabalhos. O primeiro é um artigo da NASA (2017), onde é divulgado que foram descobertos dois novos exoplanetas em torno das estrelas **Kepler-80** e **Kepler-90** após se desenvolver um trabalho em cima dos dados acumulados pelo Kepler e usando a tecnologia de **Inteligência Artificial** (IA) do **Google**. A descoberta surgiu após os pesquisadores *Christopher Shallue* e *Andrew Vanderburg* treinarem um computador para aprender a identificar exoplanetas nas leituras de luz registradas pelo **Kepler**. O segundo trabalho é o próprio artigo publicado pelos pesquisadores Shallue and Vanderburg (2018), onde descrevem o modelo desenvolvido e que apresentou-se altamente efetivo para classificar potenciais sinais de planetas usando **Aprendizado Profundo** (do inglês *deep learning*), um método específico de **Aprendizado de Máquina**.

Com isso, temos como problema de pesquisa: a possibilidade de que esse modelo também pode ser aplicado para recuperar e avaliar a composição atmosférica de um exoplaneta. Ainda, comparado com outros modelos, que usam técnicas diferentes, se o modelo pode ser considerado a melhor opção para a recuperação de exoplanetas.

E uma provável hipótese que apresento é que ao entender de forma ampla este modelo, e através de conceitos e técnicas da **Ciência de Dados**, podemos propor melhorias no modelo, para aumentar a qualidade e a capacidade de análise, interpretação e visualização dos dados recebidos por qualquer telescópio, ao ponto que possa se equivaler ao pesquisador com igual ou maior precisão, permitindo assim que os pesquisadores possam se dedicar a outras atividades.

A pesquisa terá então como objeto de estudo o modelo criado por Shallue and

Vanderburg (2018), o **AstroNet**, e procurará reproduzi-lo na íntegra, utilizando-se, primeiramente, um conjunto de dados pré-computado para treinamento e testes do modelo, sendo este disponibilizado pelos próprios desenvolvedores. Posteriormente, serão utilizadas duas bases de dados com dados reais do telescópio **Kepler** e que são disponibilizadas publicamente através do **Mikulski Archive Space Telescopes (MAST)** e do **Nasa Exoplanet Archive**. A partir destas bases, será gerado um novo conjunto de dados para treinamento e testes do **AstroNet**.

1.1 Justificativa

Atualmente, com a nova geração de telescópios, construídos em solo ou na órbita terrestre, estamos registrando grandes quantidades de sinais (dados) em diversas frequências do espectro eletromagnético conhecido, desde os raios gama e alfa, passando por ultravioleta, luz visível, infravermelho, micro-ondas, até as bandas de rádios.

E como ocorre em diversas áreas do conhecimento, está sendo acumulado um grande volume dados, colocando a **Astronomia** na era do **Big Data**.

> *Astronomical data, already amounting to petabytes, continue to increase with the advent of new instruments. Astronomy, like many other scientific disciplines, is facing a data tsunami that necessitates changes to the means and methodologies used for scientific research. This new era of astronomy is making dramatic improvements in our comprehensive investigations of the Universe.* (ZHANG; ZHAO, 2015)

Porém, o problema é que todo esse repositório de dados espaciais, onde, existe um crescimento exponencial em volume e variedade de dados, não consegue ser processado, analisado e interpretado, transformando-se assim em informação e conhecimento, como, por exemplo, na busca e caracterização de exoplanetas.

Diante deste contexto, se torna cada vez mais necessário o desenvolvimento de processos que automatizem a análise deste enorme volume de dados, recuperando assim informações valiosas. Portanto, há ainda muita pesquisa a ser realizada e demanda para o desenvolvimento de novas tecnologias e ferramentas, capazes de auxiliarem cada vez mais os astrônomos, e trabalhos como de Shallue and Vanderburg (2018), com o seu **AstroNet**, são de extrema importância para a **Ciência Exoplanetária**, o que incentiva ainda mais a produção de outras pesquisas na área.

1.2 Objetivos

Levantar os principais conceitos e práticas relacionados à **Ciência de Dados** e estudar o modelo existente desenvolvido por Shallue and Vanderburg (2018), o **AstroNet**,

propondo melhorias no modelo, como a possibilitando de que ele também seja capaz de recuperar e analizar a composição atmosférica de um exoplaneta.

1.2.1 Objetivos Específicos

- A partir de uma revisão de literatura, respaldada em autores consagrados da área e publicações recentes, apresentar os conceitos e práticas da **Ciência de Dados**, bem como conhecer o estado da arte com relação às pesquisas modelos desenvolvidos para a detecção de exoplanetas;
- Por meio de um estudo de caso, após o levantamento dos conceitos e análise das tecnologias existentes para o processamento de recuperação de exoplanetas, analisar e reproduzir o modelo de Shallue and Vanderburg (2018), o **AstroNet**, para identificação automática de exoplanetas, propondo um modelo computacional que possa melhorar a qualidade e a capacidade de análise, recuperação e visualização dos dados produzidos pelos modernos telescópios em atividade, principalmente pelo **Telescópio Espacial Kepler**.

1.3 Estrutura do Trabalho

Este trabalho está organizado da seguinte forma: no **Capítulo 1**, são apresentados a introdução, jutificativa e objetivos deste trabalho. Em seguida, no **Capítulo 2**, é apresentado todo o referência teórico utilizado no desenvolvimento deste trabalho, onde a revisão bibliográfica teve como objetivo conhecer o estado da arte em relação às pesquisas e os modelos computacionais desenvolvidos para a identificação de exoplanetas.

No **Capítulo 3**, são apresentados conceitos básicos necessários para um melhor entendimento deste trabalho, como, por exemplo, **Redes Neurais Profundas** e **Especificidades da Astronomia**. Neste capítulo também é apresentada toda a arquitetura do modelo estudado, o **AstroNet**, bem como a descrição dos aspectos computacionais utilizados para seu funcionamento.

A seguir, no **Capítulo 4**, é apresentada a descrição de como os experimentos foram realizados, desde a obtenção dos dados até o treinamento e teste do modelo, bem como os resultados obtidos.

Por fim, no **Capítulo 5**, são apresentadas as considerações finais sobre o trabalho e também são apresentadas algumas sugestões para o desenvolvimento de trabalhos futuros derivados do modelo estudado.

2 REFERENCIAL TEÓRICO

A revisão bibliográfica deste trabalho teve como objetivo conhecer o estado da arte com relação às pesquisas e ferramentas (modelos) desenvolvidas para a recuperação de exoplanetas a partir dos conjuntos de dados disponibilizados publicamente das missões realizadas pelos telescópios espaciais da **NASA**, como o **Telescópio Espacial Kepler**. Com isso, também buscou-se responder a questão: "Quais abordagens da ciência de dados têm sido utilizadas para a detecção de exoplanetas?".

Como critérios para seleção dos trabalhos que foram utilizados como referência neste trabalho, levou-se em consideração a relevância dos mesmos com o tema e o fator de impacto das revistas científicas em que estes trabalhos foram publicados. A seleção de artigos também considerou o seu período de criação, os últimos cinco anos (2018 - 2022), para se observar o que está sendo publicado recentemente com relação a recuperação de exoplanetas que utiliza-se de conceitos e ferramentas da **Ciência de Dados**. Ainda, como último critério para seleção dos artigos, foram priorizados aqueles em que o código fonte do modelo desenvolvido pelos autores foram disponibilizados publicamente para a comunidade e são de acesso aberto.

Inicialmente, a revisão se deu a partir do artigo publicado por **Christopher Shallue**, engenheiro de software e pesquisador de **Inteligência Artificial** do **Google**, e **Andrew Vanderburg**, astrofísico e professor assistente de física no **Massachusetts Institute of Technology** (MIT), onde apresentam um método bem-sucedido que classifica sinais de potenciais planetas utilizando algoritmos de **Aprendizado de Máquina** e implementados no **TensorFlow**, que a partir de dados reais do **Telescopio Espacial Kepler**, treinaram uma **Rede Neural Convolucional Profunda**, do inglês *Deep Convolutional Neural network* (DCNN), para identificar se um determinado sinal (curvas de luz) é um exoplaneta em trânsito ou um falso positivo causado por fenômenos astrofísicos (como eclipses estelares, manchas solares ou estrelas variáveis) ou ruídos instrumentais. Com esse trabalho, eles conseguiram encontrar dois novos exoplanetas, sendo um em torno da estrela **Kepler-80** e o outro em órbita a estrela **Kepler-90**. Este modelo, denominado **AstroNet**, tem seu código fonte disponibilizado publicamente[1], permitindo assim a sua investigação, bom como a possibilidade de melhorias e adaptações (SHALLUE; VANDERBURG, 2018).

Os dois pesquisadores continuam seu trabalho na área de detecção de exoplanetas e no desenvolvimento do **AstroNet**, publicando ainda mais mais três artigos sobre o tema em colaboração com outros pesquisadores. O segundo artigo aplica o modelo de rede neural convolucional em dados da **missão K2** (**AstroNet-K2**), que é uma missão estendida do **Telescópio Kepler**, onde também encontraram dois novos planetas (DATTILO *et al.*,

[1] https://github.com/cshallue/exoplanet-ml

2019). O terceiro artigo é uma adaptação do modelo agora para trabalhar com dados de outro outro telescópio espacial, o **TESS**, sendo a primeira rede neural a ser treinada e testada em dados reais deste telescópio (**AstroNet-Triage**) (YU *et al.*, 2019). Por fim, o quarto artigo publicado, treinaram seus modelos de aprendizado de máquina em dados simulados, porém utilizando outro método para detecção de exoplanetas, com observações de velocidade radial, que é o movimento da estrela em torno do centro de massa produzido pela órbita do planeta em torno desta estrela, ou seja, a presença de um exoplaneta é deduzida a partir do deslocamento nas linhas espectrais da estrela hospedeira devido ao efeito **Doppler** (BEURS *et al.*, 2021).

Os trabalhos de **Shallue** e **Vanderburg**, direta ou indiretamente, influenciam trabalhos de outros pesquisadores. Ansdell *et al.* (2018) expandiram o modelo de **Shallue** e **Vanderburg**, incluindo conhecimento de domínio científico na arquitetura da rede e representações de entrada, melhorando a classificação de trânsito de exoplanetas com aprendizado profundo, conseguindo aumentar o desempenho geral do modelo de 95,8% para 97,5% de precisão e de 95,5% para 98,0% de precisão média. O código e os dados usados no trabalho estão disponíveis publicamente[2] (ANSDELL *et al.*, 2018).

Outros trabalhos também utilizaram aprendizado de máquina com rede neural para desenvolverem seus modelos de detecção de exoplanetas, porém com pequenas alterações nas estratégias adotadas para esse fim.

Cui *et al.* (2021) exploram um modelo de detecção de objetos bidimensionais (2D) para identificação de sinais de planetas em trânsito, ou seja, a detecção é realizada na imagem - as curvas de luz são plotadas em uma imagem -, de forma que a série temporal 1D seja convertida em uma imagem 2D, o que permite que algoritmos de visão computacional sejam aplicados a dados de curva de luz. De acordo com os autores, essa é uma abordagem direta e corresponde à intuição da percepção visual humana. Para comparação, o modelo de **Shallue** e **Vanderburg** utilizam Redes Neurais Convolucionais 1D, que trabalham nos dados do Kepler com visualizações locais e globais para curvas de luz dobradas de candidatos a trânsito. O modelo de treinamento e detecção de exoplanetas são implementados em uma aplicação denominada **Deep-Transit**, desenvolvido em **Python** e disponibilizado publicamente[3] (CUI *et al.*, 2021).

Dvash *et al.* (2022) também utilizam um modelo de rede neural para detecção de exoplanetas baseado em imagens, que marca amostras que foram obtidas durante os trânsitos, semelhante a tarefa de identificar o contexto semântico de cada pixel de uma imagem, chamada de "**segmentação semântica**" em visão computacional, e geralmente realizada por **Redes Neurais Profundas**. Ainda, a rede neural que desenvolveram também faz uso de conceitos de **Aprendizado Profundo** (*deep learning*) como **U-Nets**

[2] https://gitlab.com/frontierdevelopmentlab/exoplanets
[3] https://github.com/ckm3/Deep-Transit

(rede neural convolucional desenvolvida para segmentação de imagens), **Rede Adversarial Generativa**, do inglês *Generative Adversarial Networs* (GAN), e **Perda Adversária**. Detalhes técnicos das arquiteturas utilizadas e o código desenvolvido pelos autores estão disponíveis publicamente no **GitHub**[4] (DVASH *et al.*, 2022).

Valizadegan *et al.* (2022) desenvolveram o **ExoMiner**, um classificador de aprendizado profundo baseado em um novo modelo de **Rede Neural Profunda**, do inglês *Deep Neural Network* (DNN), que procura imitar o processo pelo qual os especialistas de domínio examinam sinais de trânsito, que verificam vários tipos de testes de diagnósticos em formas de valores escalares e dados de séries temporais. De acordo com os autores, acreditam que classificadores DNN mais precisas podem ser desenvolvidos se a vasta quantidade de conhecimento de domínio utilizada no projeto do *pipeline* também for usada como guia para projetar a arquitetura da DNN. O **ExoMiner** conseguiu recuperar 93,6% de todos os exoplanetas no conjunto de teste. O modelo foi aplicado em dados do telescópio **Kepler**, mas dizem que também pode ser transferido para classificar sinais do telescópio **TESS** (VALIZADEGAN *et al.*, 2022).

Alguns trabalhos, ainda utilizando rede neural, procuraram caracterizar os exoplanetas, treinando seus modelos para recuperarem dados atmosféricos ou tecnoassinaturas dos exoplanetas já detectados.

Em Cobb *et al.* (2019), nos é apresentado o **Plan-Net**, um modelo de **Aprendizado de Máquina** (ML), do inglês *Machine Learning*, para recuperação atmosférica de exoplanetas baseado em um conjunto de **Redes Neurais Bayesianas**, do inglês *Bayesian neural networks* (BNNs), que gera inferências, segundo os autores, mais precisas do que uma **Floresta Aleatória**, do inglês *Random Forest* (RF), desenvolvida em trabalhos anteriores. Também introduziram em seu modelo uma nova função de perda para BNNs que aprende correlações entre as saídas do modelo. Ainda falam que a incorporação de conhecimento específico de domínio em modelos de ML pode gerar resultados melhores, oferecendo assim *insights* sobre a covariância dos parâmetros atmosféricos. O código do **Plan-Net** está disponível para acesso público[5] (COBB *et al.*, 2019).

Yip *et al.* (2021) apontam que a caracterização de atmosferas em exoplanetas é uma disciplina que ainda está dando seus primeiros passos. O trabalho é desenvolvido a partir de análises da espectroscopia de trânsito, que consiste em observar trânsitos em diferentes comprimentos de onda, permitindo aos astrônomos detectar as composições químicas dos planetas, como vapor de água, moléculas contendo carbono, óxidos e espécies alcalinas na atmosfera. Porém, os autores apontam que apesar de seu alto poder preditivo, as DNNs também são famosas por serem "**caixas pretas**". Partindo destas premissas, os autores apresentaram uma série de metodologias gerais de avaliação que podem ser

[4] https://github.com/StrudelTAU/ShallowTransitsDL
[5] https://github.com/exoml/plan-net

aplicadas a qualquer modelo treinado. O trabalho então treinou três arquiteturas DNN populares diferentes para recuperação parâmetros atmosféricos de espectros de exoplanetas e mostraram que todas as três alcançam um bom desempenho preditivo. Também realizaram uma extensa análise das previsões de DNNs, para obterem informações dos limites de credibilidade dos parâmetros atmosféricos para um determinado instrumento e modelo. Por fim, realizaram uma análise de sensibilidade baseada em perturbação para identificar a quais características do espectro o resultado da recuperação é mais sensível. Suas conclusões foram de que para diferentes moléculas, as faixas de comprimento de onda para as quais as previsões de DNNs são mais sensíveis coincidem de fato com suas regiões de absorção características. A implementação do modelo está disponível no **Github**[6] (YIP *et al.*, 2021).

Pinchuk and Margot (2022), desenvolveram seu trabalho no contexto de buscas de tecnoassinatura de rádio, descrevendo uma aplicação de **Rede Neural Convolucional**, do inglês *Convolutional Neural network* (CNN), para a extração de **Interferência de Radiofrequência**, do inglês *Radio Frequency Interference* (RFI), em dados de rádio telescópios. Esses sinais devem ser cuidadosamente analisados, pois podem determinar se são ou não de natureza antropogênica, ou seja, oriundos de uma civilização inteligente. O trabalho incluiu o desenvolvimento de um **filtro de direção de origem**, do inglês *direction-of-origin* (DoO), utilizando técnicas modernas de visão computacional, onde um sinal é classificado como RFI se for detectado em várias direções no céu. Os autores então projetaram e treinaram uma CNN que pode determinar se o sinal da primeira imagem também está presente na segunda imagem, podendo assim, aplicar essa rede para determinar se um sinal detectado é persistente em uma e apenas uma direção no céu. Todos os modelos implementados utilizando o **TensorFlow**, bem como o código para reprodução do modelo final estão disponíveis para acesso público[7] (PINCHUK; MARGOT, 2022).

Porém, outros trabalhos, com o objetivos de criarem aplicações para detecção de exoplanetas, ou caracterizarem estes, adotaram outros métodos para esses fins.

Zingales and Waldmann (2018) desenvolveram uma aplicação chamada **ExoGAN** (*Exoplanet Generative Adversarial Network*), um modelo de aprendizado profundo com **Rede Adversarial Generativa Convolucional Profunda**, do inglês *Deep Convolutional Generative Adversarial Network* (DCGAN), que dada a entrada de um espectro observado é capaz de reconhecer características moleculares, abundância de gases traços atmosféricos e parâmetros planetários usando aprendizado não supervisionado, realizando uma interpretação das medidas espectroscópicas de transmissão e emissão. O conjunto de treinamento do **ExoGAN** contempla uma ampla gama de químicas atmosféricas e tipos de planetas para realizar estas interpretações. Segundo os autores, as recuperações

[6] https://github.com/ucl-exoplanets/Spectra_Sensitivity_analysis
[7] https://github.com/UCLA-SETI-Group/doom/releases/tag/v1.0.1

do **ExoGAN** constituem uma melhoria significativa da velocidade em relação às recuperações tradicionais e podem ser usadas como uma análise atmosférica final ou fornecer restrições prévias à recuperação subsequente. Com as observações que serão feita pelos novos telescópios **James Webb Space Telescope** (JWST) e **ARIEL**, a modelagem da espectroscopia atmosférica exoplanetária através dos chamados algoritmos de recuperação atmosférica ganharão cada vez mais destaque. O código fonte do modelo desenvolvido pelos autores são de acesso aberto e foi disponibilizado publicamente para a comunidade[8] (ZINGALES; WALDMANN, 2018).

Para Gordon, Agol and Foreman-Mackey (2020), os **Processos Gaussianos**, do inglês *Gaussian Processes* (GPs), são comumente utilizados como modelo de variabilidade estocástica em séries temporais astrofísicas, empregados para explicar a variabilidade estelar correlacionada nas curvas de luz de trânsito planetário. E devido aos avanços nos métodos de **GPs**, incluindo o método da **celerite**, cada vez mais é aplicado de forma eficiente este método em curvas de luz. O trabalho dos autores se concentrou no método do **Processo Gaussiano** de modelagem de ruído correlacionado, apresentando uma extensão do código **celerite**, que pode ser utilizado para modelar ruído em duas dimensões. Com isso, usaram a extensão para simular a variabilidade estelar de vários comprimentos de onda em observações de trânsito, demonstrando que ao modelar com precisão a correlação entre comprimentos de onda, é possível mitigar os efeito da variabilidade estelar na detecção de exoplanetas em trânsito e na medição de suas propriedades. Todo o código está disponível publicamente na forma de um pacote **Python** instalável *pip* chamado **specgp**[9] (GORDON; AGOL; FOREMAN-MACKEY, 2020).

Os autores Robnik and Seljak (2020) argumentam que sempre que forem necessários parâmetros de trânsito planetário de alta precisão, uma análise conjunta da variabilidade de estrelas e trânsito de planetas deve ser realizada, juntamente com uma análise de probabilidade de ruído **não Gaussiana** adequada. Para isso, primeiramente construíram um modelo de distribuição de probabilidade de ruído que permite analisar rigorosamente as detecções de planetas e fazer uso de estatísticas robustas, como remoção de *outliers* desnecessários, que tornam a distribuição de probabilidade de ruído **não Gaussiana**. Em seguida, desenvolveram uma análise de verossimilhança de sinal com base nessa distribuição de probabilidade ruído verdadeiro, na qual o sinal é modelado como uma soma da variabilidade estelar e trânsitos planetários. Neste trabalho os autores definiram ruído como a diferença entre os dados observados e o sinal que combina a variabilidade estelar com os trânsitos do planeta. Já para o modelo de variabilidade estelar, foi desenvolvida uma análise de **Processo Gaussiano** utilizando uma abordagem de filtro Wiener baseada em Fourier, onde o espectro de potência é não paramétrico e aprendido a partir dos dados. Por fim, desenvolveram uma otimização de alta dimensão da função objetivo, onde

[8] https://github.com/orgs/ucl-exoplanets
[9] https://github.com/exoplanet-dev/exoplanet

otimizaram em conjunto todos os parâmetros do modelo, incluindo milhares de modos de variabilidade de estrelas e parâmetros de trânsito planetário. Todo o código fonte deste trabalho foi disponibilizado publicamente[10] (ROBNIK; SELJAK, 2020).

O trabalho de Günther and Daylan (2021), apresenta uma aplicação desenvolvida em **Python** e disponibilizado publicamente [11], com código fonte aberto, que permite inferência simultânea, ou seja, conjunta, de estrelas e exoplanetas a partir de dados fotométricos e de velocidade radial. Oferecendo uma seleção de modelos orbitais e de trânsito/eclipse e modelos de ruído sistemáticos, a aplicação apresenta estimativa de parâmetros e seleção de modelo **Baysiano**, permitindo que que o usuário escolha entre um ajuste de **Markov Chain Monte Carlo** ou vários algoritmos de **Amostragem Aninhada**. Ambos pegam amostras justas do posterior do modelo selecionado, enquanto o último também fornece uma estimativa de baixa variância da evidência **Bayesiana** para comparação de modelos estatísticos e é mais robusto para alta dimensionalidade (GüNTHER; DAYLAN, 2021).

Como se observa, nos últimos anos os astrônomos vêm lutando com aplicações de metodologias estatísticas sofisticadas afim de analisar seus enormes conjuntos de dados e resolver problemas astrofísicos complexos (FEIGELSON; BABU, 2014). E como podemos observar, a **Ciência Exoplanetária** já é um dos campos de pesquisa em expansão mais rápida da **Astronomia**, sendo que mais de cinco mil exoplonetas já foram confirmados, o que fornece a motivação necessária para o seu maior crescimento, através do desenvolvimento de novas pesquisas apoiadas cada vez mais pela **Ciência de Dados**. Diante deste contexto, este trabalho reproduzirá, na íntegra, o modelo **AstroNet**, criado por Shallue and Vanderburg (2018), ferramenta esta desenvolvida com o propósito de identificar exoplanetas em dados produzidos pelo **Telescópio Espacial Kepler**.

[10] https://github.com/JakobRobnik/Kepler-data-analysis
[11] https://github.com/MNGuenther/allesfitter

3 METODOLOGIA

Este trabalho é um estudo de caso do modelo **AstroNet** (SHALLUE; VANDERBURG, 2018), que identifica automaticamente exoplanetas em sinais (**Curvas de Luz**) produzidos pelo **Telescópio Espacial Kepler**, e foi desenvolvido por **Christopher Shallue**, engenheiro de *software* e pesquisador de **Inteligência Artificial** do **Google**, e **Andrew Vanderburg**, astrofísico e professor assistente de física no **Massachusetts Institute of Technology (MIT)**, e buscou responder "como" a utilização de um modelo de **Aprendizado de Máquina Profundo** pode resolver e ajudar no tratamento dos terabytes de dados produzidos pelos telescópios espaciais, o chamado **"Big Data Astronômico"**.

O **AstroNet** foi escolhido pela facilidade de acesso às informações, principalmente ao seu código fonte, permitindo assim uma completa investigação do trabalho realizado e possibilitando a elaboração de uma nova solução para o problema.

3.1 CONCEITOS BÁSICOS

Primeiramente, para melhor entendimento do modelo **AstroNet**, que será descrito com maiores detalhes mais a frente neste capítulo, vamos passar por alguns conceitos básicos de **Ciência de Dados**, como o **Aprendizado de Máquina** (ML), do inglês *Machine Learning*, e as **Redes Neurais Convolucionais**.

3.1.1 Aprendizado de máquina

A **Ciência de Dados** lida principalmente com a transformação de problemas empresariais em problemas de dados, coletando, entendendo, limpando e formatando dados, e o **Aprendizado de Máquina** é um dos recursos que os cientistas de dados utilizam para resolver esses problemas de dados. Por isso, é essencial o conhecimento de **ML** para esses profissionais (GRUS, 2016).

Mas antes de definirmos propriamente o **Aprendizado de Máquina**, vamos definir o que é modelo. Grus (2016) diz que modelo "é simplesmente a especificação de uma relação matemática (ou probabilística) existente entre variáveis diferentes".

Com isso, agora podemos definir **Aprendizado de Máquina** como a aplicação de técnicas computacionais na tentativa de encontrar padrões ocultos em dados, ou seja, é a criação e o uso de modelos que são aprendidos a partir dos dados (relação entre variáveis) (AMARAL, 2016). Grus (2016) pode complementar ao dizer que "normalmente, nosso objetivo será usar os dados existentes para desenvolver modelos que possamos usar para prever possíveis saídas para dados novos".

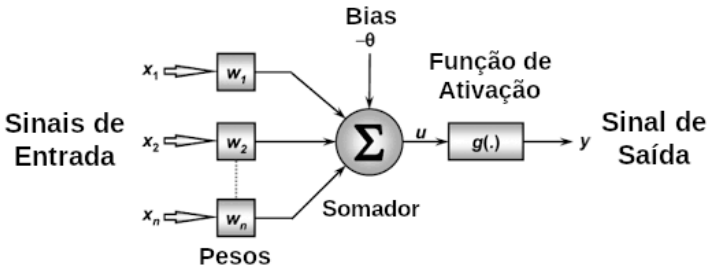
A título de curiosidade, em alguns contextos, o **ML** também pode ser chamado de "modelo preditivo" ou "mineração de dados". Ainda, o **ML** está estreitamente associado com outras ciências, como a **Estatística** e a **Inteligência Artificial**, sendo que alguns consideram o **ML** uma área da **Inteligência Artificial** (AMARAL, 2016).

O **Aprendizado de Máquina** é o emprego de um algoritmo para extrair informações de um conjunto de dados bruto, representando-os em algum tipo de modelo matemático, de modo que, utilizaremos este modelo para fazer inferências a partir de novos conjuntos de dados. Para isso, existem vários algoritmos, mas um tipo em especial que está se destacando dos demais, são as **Redes Neurais Artificiais** (ACADEMY, 2022).

3.1.2 Redes neurais artificiais

Uma das técnicas empregadas em **Aprendizado de Máquina** são as **Redes Neurais Artificiais** (RNA), que nascem da busca pela criação de máquinas inteligentes, ou que tenham comportamentos inteligentes, e para isso se baseiam nos neurônios do nosso próprio cérebro (KOVáCS, 2006). Ou seja, um **neurônio artificial**, ou **neurônio matemático**, simula um neurônio do cérebro humano.

Figura 1 – Neurônio artificial Perceptron.

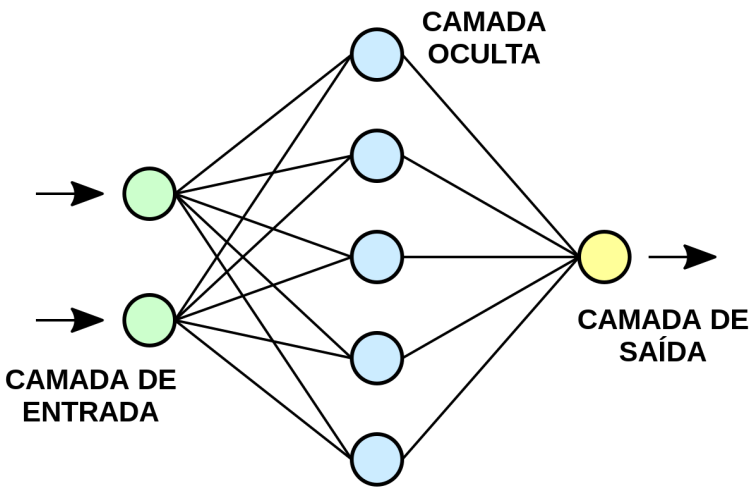


Fonte: Adaptado de <https://pt.wikipedia.org/wiki/Perceptron_multicamadas>. Acesso em: 19 set. 2022.

Na década de 1950, **Frank Rosenblatt**, um neurobiologista da **Universidade de Cornell**, desenvolve um modelo matemático de um **neurônio artificial** que denominou ***Perceptron*** (**Figura 1**). Uma **RNA** é uma rede com uma topologia onde os **neurônios artificiais** estão dispostos em várias camadas (**Figura 2**), sendo que:

> Os neurônios que recebem diretamente as entradas da rede constituem o que se chama de **camada de entrada**. Os neurônios que recebem como entradas as saídas daquelas da camada de entrada constituem a segunda camada e assim sucessivamente até a camada final que é a **camada de saída**. As camadas internas que não são nem a de entrada e nem a de saída são geralmente referidas como **camadas ocultas**. (KOVáCS, 2006)

Figura 2 – Rede neural artificial.



Fonte: Adaptado de <https://pt.wikipedia.org/wiki/Rede_neural_artificial>.
Acesso em: 19 set. 2022.

Na **Figura 2**, os círculos são os **neurônios artificiais**, onde cada neurônio é caracterizado pelo **peso**, **bias** e uma **função de ativação**, como pode-se observar em detalhe na **Figura 1**. Os dados de entrada são introduzidos pela **camada de entrada**, então os neurônios fazem uma modificação linear nessa entrada através de seus **pesos** e **bias**. A modificação não-linear é feita pela **função de ativação**. Com isso, a informação se move da **camada de entrada** para as **camadas ocultas**, onde são feitos os processamentos e enviados para a saída final, a **camada de saída**. Ou seja, para que a aprendizagem possa funcionar, os pesos entre as conexões dos neurônios são ajustados durante o treinamento, causando alterações correspondentes na saída da rede (ACADEMY, 2022).

3.1.3 Funções de ativação

Uma **função de ativação** é um componente matemático essencial na estrutura de uma **RNA**, o que permite que problemas complexos possam ser resolvidos, e são responsáveis pela decisão se um neurônio será ativado ou não. Sem essas funções, uma **RNA** seria apenas um modelo de regressão linear.

> A função de ativação é a transformação não linear que fazemos ao longo do sinal de entrada. Esta saída transformada é então enviada para a próxima camada de neurônios como entrada. Quando não temos a função de ativação, os pesos e bias simplesmente fazem uma transformação linear. Uma equação linear é simples de resolver, mas é limitada na sua capacidade de resolver problemas complexo. (ACADEMY, 2022)

Uma **função de ativação** ao fazer essa transformação não-linear nos dados de

entrada, torna-o capaz de aprender e executar tarefas mais complexas, que funcionem em tarefas complicadas, como, por exemplo, a classificação de imagens (**Visão Computacional**), tarefa esta que é realizada neste trabalho (ACADEMY, 2022).

Existem diversos tipos de **funções de ativação**. agora veremos os tipos que são utilizados pelo modelo **AstroNet**.

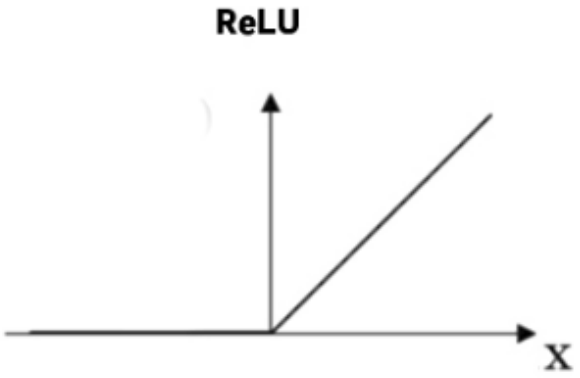
3.1.3.1 Função ReLU

O primeiro tipo é a **ReLU**, do inglês *Rectified Linear Unit* (unidade linear retificada), que é a **função de ativação** mais amplamente utilizada ao se projetar redes neurais atualmente. Essa função é mais fácil de se treinar e por muitas vezes obtém melhores resultados, pois por serem quase lineares, conseguem preservar muitas das propriedades que fazem com que os modelos lineares fiquem mais simples de se otimizar e com uma boa capacidade de generalização.

A sua principal vantagem sobre outras **funções de ativação** é que ela não ativa todos os neurônios ao mesmo tempo, ou seja, se a entrada for, por exemplo, negativa, ela será convertida em 0 (zero) e o neurônio não será ativado. Isso significa que, ao mesmo tempo, apenas alguns neurônios serão ativados, o que torna a rede dispersa, eficiente e fácil para a computação (ACADEMY, 2022). A função **ReLU** é definida como:

$$f(x) = max(0, x) \tag{3.1}$$

Figura 3 – Gráfico da função de ativação ReLU.



Fonte: Elaborado pelo autor.

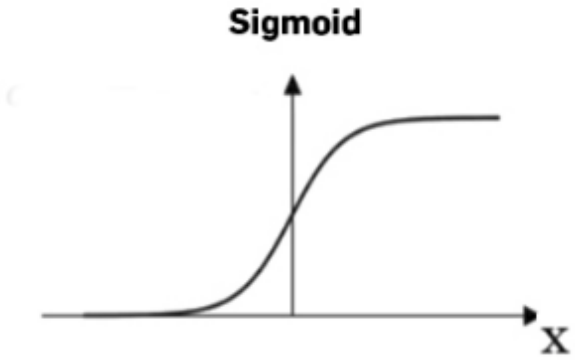
3.1.3.2 Função Sigmoide

O segundo tipo é a **Sigmoide**, que é uma **função de ativação** suave e é continuamente diferenciável. Esta é uma característica interessante da função **Sigmoide**, pois significa basicamente que quando existem vários neurônios em uma rede neural com a

função de ativação Sigmoide, a saída não é linear. A função varia de 0 (zero) a 1 (um) tendo um formato **S**. A função **Sigmoide** é definida como:

$$f(x) = \frac{1}{(1 + e^{-x5})} \tag{3.2}$$

Figura 4 – Gráfico da função de ativação sigmoide.

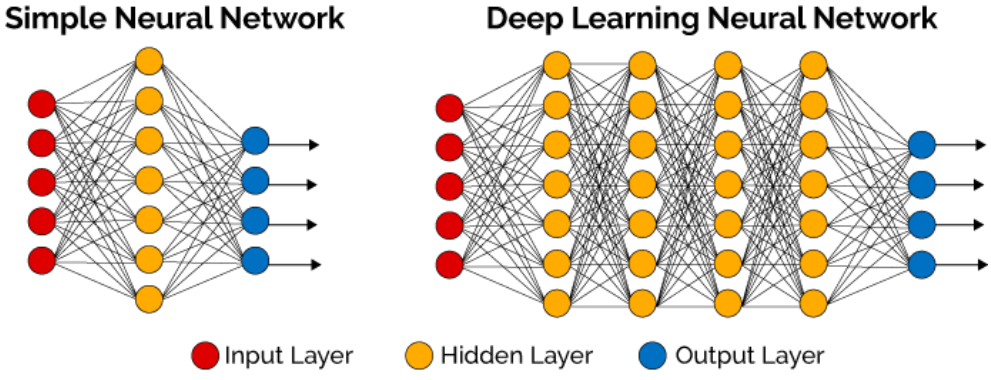


Fonte: Elaborado pelo autor.

3.1.4 Aprendizado profundo ou *deep learning*

O **Aprendizado Profundo**, ou *Deep Learning*, que é uma subárea do **ML**, e está sendo o responsável pelos avanços recentes que estão ocorrendo em **Visão Computacional**, desde o reconhecimento de fala, processamento de linguagem natural até o reconhecimento de áudio, e se baseia no conceito de **Redes Neurais Artificiais**.

Figura 5 – Rede neural simples Vs Rede neural profunda (*deep learning*).



Fonte:
<https://www.deeplearningbook.com.br/o-que-sao-redes-neurais-artificiais-profundas/>.
Acesso em: 19 set. 2022.

No **Aprendizado Profundo**, as redes neurais, também conhecidas como **Redes Neurais Profundas**, utilizam camadas de **neurônios artificiais** para processar dados,

como reconhecer objetos visualmente. Essa informação é transmitida através de cada camada, onde a saída da camada anterior fornece a entrada para a próxima camada. A primeira camada em uma rede é chamada de **camada de entrada**, enquanto a última é chamada de **camada de saída**. Todas as camadas entre as duas são referidas como **camadas ocultas**. Cada camada é tipicamente um algoritmo simples e uniforme contendo um tipo de **função de ativação**. Diferentemente de uma **Rede Neural Simples**, uma **Rede Neural Profunda** é caracterizada por muitas **camadas ocultas**, como representada na **Figura 5**.

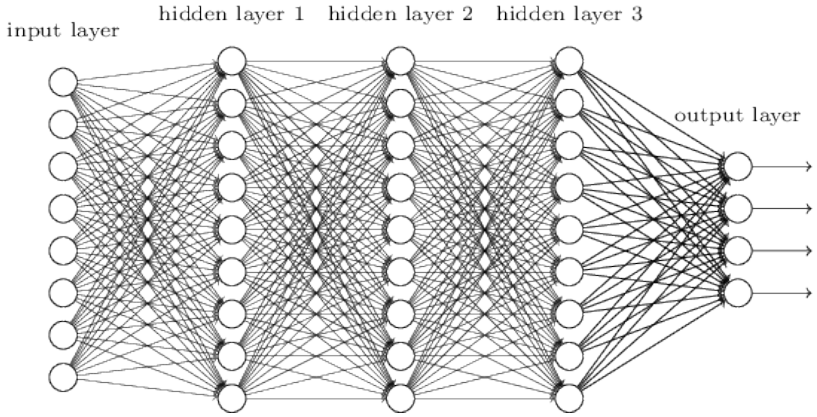
3.1.5 Redes neurais convolucionais

Uma **Rede Neural Convolucional** (CNN), do inglês *Convolutional Neural Network*, nada mais é que um algoritmo que ao receber de entrada uma imagem atribui "importância (pesos e vieses que podem ser aprendidos) a vários aspectos/objetos da imagem e ser capaz de diferenciar um do outro" (ACADEMY, 2022). Seu pré-processamento é relativamente menor se comparado a outros algoritmos de classificação, e, ao passo que "nos métodos primitivos os filtros são feitos à mão, com treinamento suficiente" (ACADEMY, 2022), uma **CNN** é capaz de aprender esses filtros.

A arquitetura de uma **CNN**:

> É semelhante ao padrão de conectividade de neurônios no cérebro humano e foi inspirada na organização do Visual Cortex. Os neurônios individuais respondem a estímulos apenas em uma região restrita do campo visual conhecida como Campo Receptivo. Uma coleção desses campos se sobrepõe para cobrir toda a área visual. (ACADEMY, 2022)

Figura 6 – Arquitetura de uma Rede Neural Convolucional.



Fonte: <https://www.deeplearningbook.com.br/introducao-as-redes-neurais-convolucionais/>. Acesso em: 19 set. 2022.

Essas redes usam uma arquitetura especial que é particularmente bem adaptada para classificar imagens (**Figura 6**) e usam três ideias básicas: campos receptivos locais, pesos compartilhados e *pooling*. O uso dessa arquitetura torna as **Redes Neurais Convolucionais** rápidas de treinar. Isso, por sua vez, nos ajuda a treinar **Redes Neurais Profundas** de muitas camadas, que são muito boas na classificação de imagens. Hoje, as **Redes Neurais Convolucionais**, ou alguma variante próxima, são usadas na maioria das redes neurais para reconhecimento de imagens.

3.2 ESPECIFIDADES DA ASTRONOMIA

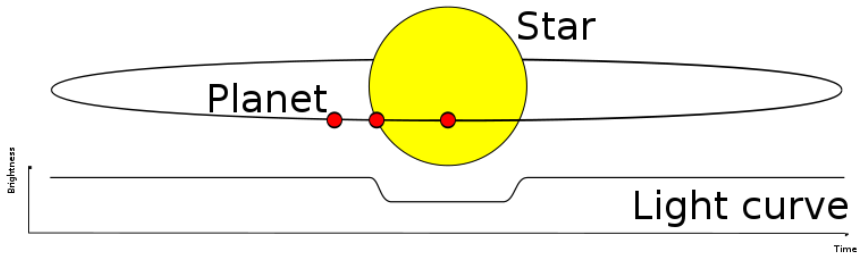
Como o **AstroNet** trabalha na classificação automática de possíveis sinais planetários (exoplanetas), através de uma **Rede Neural Convolucional**, iremos conceituar brevemente o **Método de Trânsito Planetário** e as **Curvas de Luz**.

3.2.1 Método de Trânsito Planetário

O método mais utilizado pelos astrônomos na identificação de planetas fora do nosso **Sistema Solar**, os exoplanetas, é o **Método de Trânsito Planetário**, e que consiste em encontrar um escurecimento periódico da luz à medida que um planeta passa na frente de sua estrela hospedeira, projetando assim sua sombra em nossos telescópios.

Ou seja, pelo **Método de Trânsito**, os exoplanetas são detectados quando eles passam (transitam) na frente da sua estrela hospedeira. Quando isso ocorre é verificada uma minúscula queda no brilho da estrela, registrando assim os eventos periódicos de escurecimento.

Figura 7 – Método de trânsito planetário.



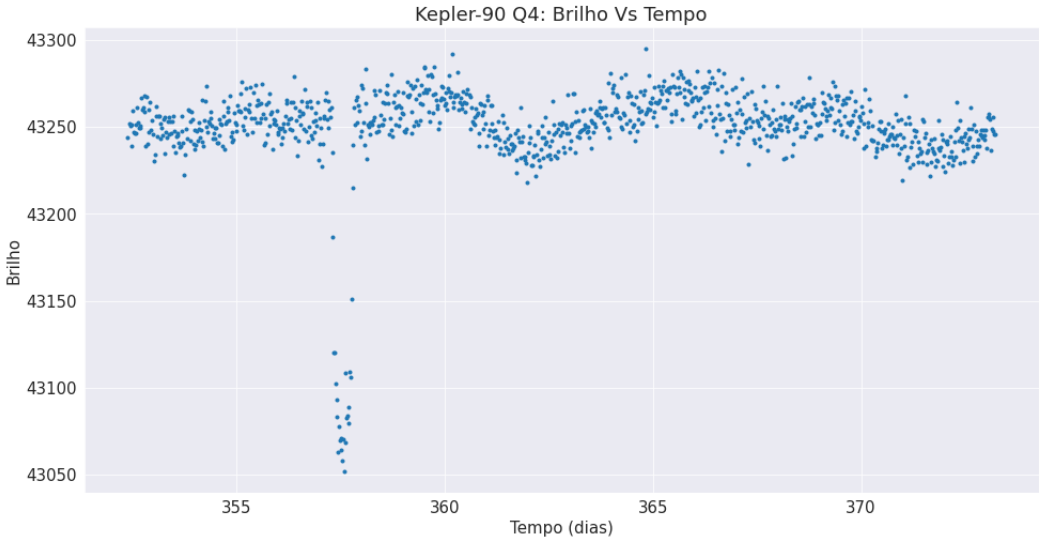
Fonte: <https://pt.wikipedia.org/wiki/Ficheiro:Planetary_transit.svg>.
Acesso em: 19 set. 2022.

3.2.2 Curvas de Luz

Na **Figura 7**, a linha desenhada abaixo do planeta e da estrela é chamada de **Curva de Luz** (*light curve*). A **Curva de Luz** é um gráfico do brilho da estrela ao longo do tempo e é a medida que o telescópio **Kepler** faz para descobrir exoplanetas. A queda

na luz que acontece quando o planeta passa na frente da estrela é chamada de “trânsito”. Os trânsitos também podem fornecer informações sobre o tamanho e a órbita do planeta.

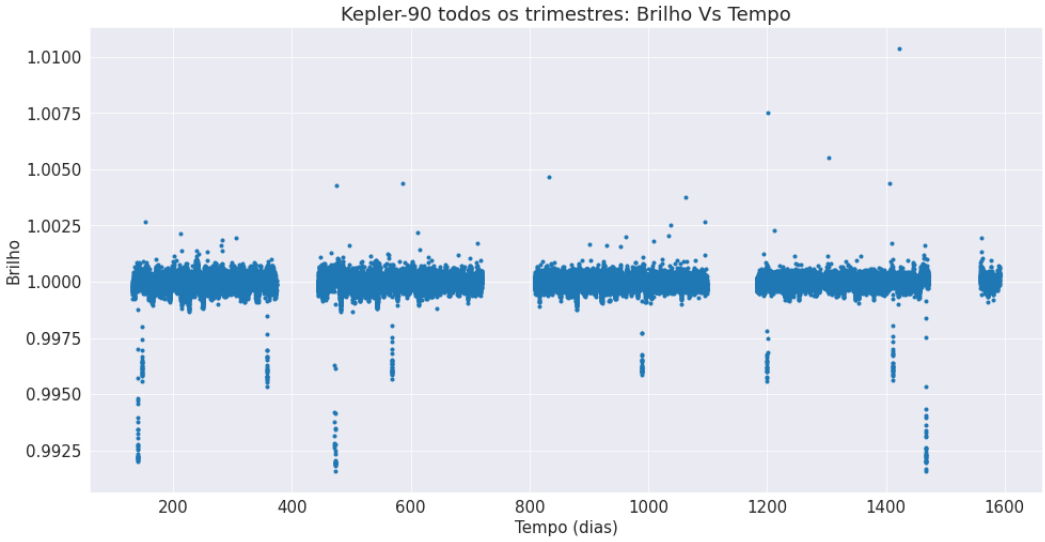
Figura 8 – Gráfico de saída Kepler-90 - trimestre 4.



Fonte: Elaborado pelo autor.

No primeiro gráfico (**Figura 8**), é apresentada uma **Curva de Luz** da estrela **Kepler-90**, de segmento único e de aproximadamente 20 dias, onde podemos observar o trânsito de um planeta, o **Kepler-90 g**. Também podemos observar que o brilho da estrela não é um plano ao longo do tempo, ocorrendo uma variação natural em seu brilho, mesmo quando longe do trânsito do planeta.

Figura 9 – Gráfico de saída Kepler-90 - todos trimestres.



Fonte: Elaborado pelo autor.

No segundo gráfico (**Figura 9**), é mostrada a **Curva de Luz** completa ao longo de toda a missão do **Kepler**, que foi de aproximadamente quatro anos. Nele podemos observar o trânsito de dois planetas, o **Kepler-90 h** (o trânsito mais profundo), e o **Kepler-90 g** (o segundo trânsito mais profundo).

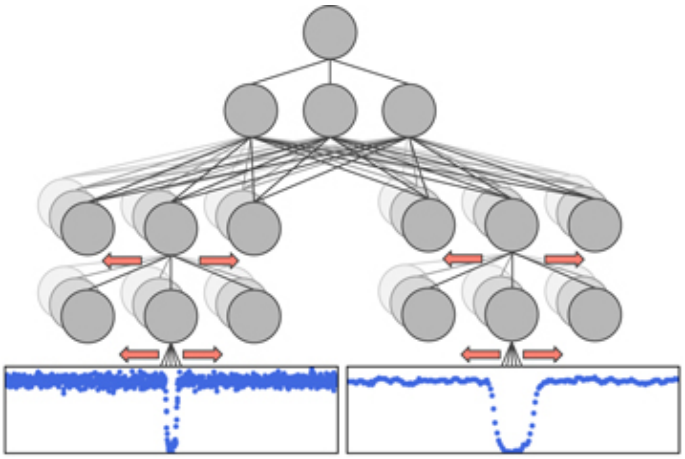
3.3 ASTRONET

Neste item focaremos na descrição do melhor modelo do **AstroNet**, que foi desenvolvido utilizando-se um tipo de **Rede Neural Profunda** com **arquitetura convolucional**.

3.3.1 Arquitetura da rede neural

Inicialmente foram consideradas, pelos pesquisadores, três tipos de arquiteturas de **redes neurais** para classificar automaticamente as **Curvas de Luz** do **Kepler**: **arquitetura linear**, **arquitetura totalmente conectada** e **arquitetura convolucional**. Para cada tipo foram consideradas três opções de entradas diferentes: apenas a **visualização global**, apenas a **visualização local** e as **visualizações globais e locais**.

Figura 10 – Arquitetura de rede neural convolucional para classificação de curvas de luz, com visualizações de entrada global e local.

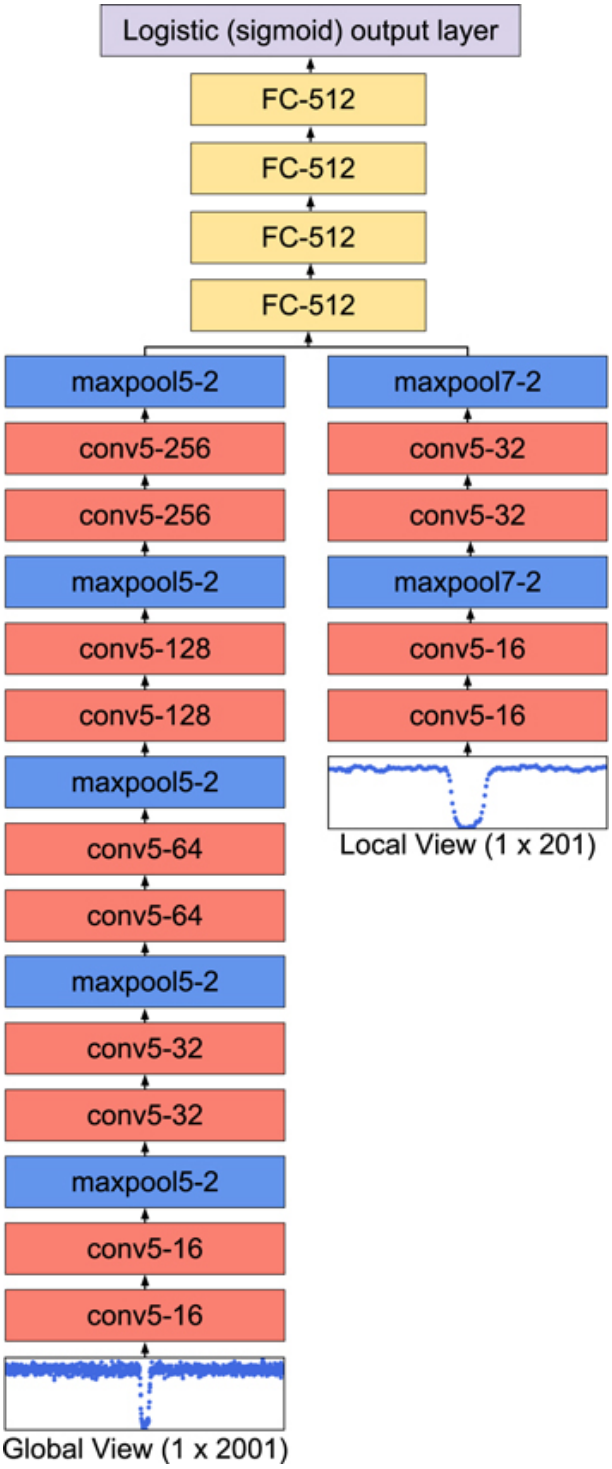


Fonte: (SHALLUE; VANDERBURG, 2018).

Neste trabalho focaremos apenas no melhor modelo desenvolvido pelos pesquisadores, que se utiliza da **arquitetura convolucional** com **visualização de entrada global e local** (**Figura 10**).

As **Redes Neurais Convolucionas** têm sido bem-sucedidas em aplicações com dados de entrada estruturados espacialmente na classificação de imagens. O modelo **AstroNet** usa uma **Rede Neural Convolucional Unidimensional** com camada de *pooling* (*pool* máximo, ou *max-pooling*). Esta arquitetura assume que as **Curvas de Luz**

Figura 11 – Arquitetura do modelo de rede neural de melhor desempenho. Camadas convolucionais são denotadas **conv<tamanho do kernel>-<número de mapas de características>**, camadas de pool máximo são denotadas **max-pool<comprimento da janela>-<comprimento do passo>**, e camadas totalmente conectadas são denotadas **FC-<número de unidades>**.



Fonte: (SHALLUE; VANDERBURG, 2018).

de entrada podem ser descritas por características espacialmente locais, e que a saída da rede deve ser invariável para pequenas traduções da entrada. Se as visualizações global e local estiverem presentes, passamos os dois vetores através de colunas convolucionais disjuntas antes de combiná-los em camadas totalmente conectadas compartilhadas. A sua configuração exata é apresentada na **Figura 11**.

Todas as camadas ocultas da rede neural **AstroNet** utilizam a **função de ativação ReLU** (retificador linear) e a camada de saída da rede utiliza a **função de ativação Sigmoide**. A saída da rede é a probabilidade prevista de que a entrada (uma **Curva de Luz**) seja um planeta em trânsito. Valores próximos de 1 (um) indicam alta confiança de que a entrada é realmente um planeta em trânsito. Já valores próximos de 0 (zero) indicam alta confiança de que a entrada é um falso positivo.

3.3.2 Pré-processamento e conjunto de treinamento

A rede neural do **AstroNet** utiliza **Aprendizado de Máquina Supervisionado**, ou seja, é fornecido a rede neural um conjunto de dados com exemplos já rotulados com os quais ela pode aprender.

Neste trabalho, realizamos dois experimentos, sendo que no primeiro experimento, utilizamos um conjunto de treinamento já pré-computado pelos criadores do **AstroNet**.

Para iniciar a geração desse conjunto de treinamento, atividade essa realizada em nosso segundo experimento, utilizamos um conjunto de dados disponibilizado em formato **CSV** (formato de arquivo que significa "valores separados por vírgulas", do inglês *Comma-Separated-Values*) e pode ser baixado no **Nasa Exoplanet Archive**[1].

Neste conjunto de dados **CSV** encontram-se os chamados **TCEs**, do inglês *Threshold-Crossing Events*, ou os "eventos de cruzamento de limiar", que são os eventos periódicos de escurecimento potencialmente consistentes da estrela-alvo com sinais produzidos por planetas em trânsito, ou seja, são sinais periódicos detectados que podem ser consistentes com planetas em trânsito do *pipeline* **Kepler**, que normalmente foram identificados manualmente por humanos para remover falsos positivos causados por variabilidade astrofísica e artefatos instrumentais.

Esses **TCEs** já estão rotulados, sendo que este trabalho foi realizado pelo **Autovetter Planet Candidate Catalog** (CATANZARITE, 2015) e foram produzidos por verificação manual e outros diagnósticos.

Os rótulos dos **TCEs** no conjunto de dados encontram-se na coluna "av_training_set" e possui três valores possíveis: candidato a planeta (**PC**, do inglês *planet candidate*), falso positivo astrofísico (**AFP**, do inglês *astrophysical false positive*) e fenômeno sem trânsito (**NTP**, do inglês *nontransiting phenomenon*). Os **TCEs** que estavam rotulados

[1] https://exoplanetarchive.ipac.caltech.edu/index.html. Acesso em 24 ago. 2022.

como “desconhecido” (**UNK**, do inglês *unknown*) foram ignorados. A **Tabela 1** apresenta as colunas encontradas no conjunto de dados.

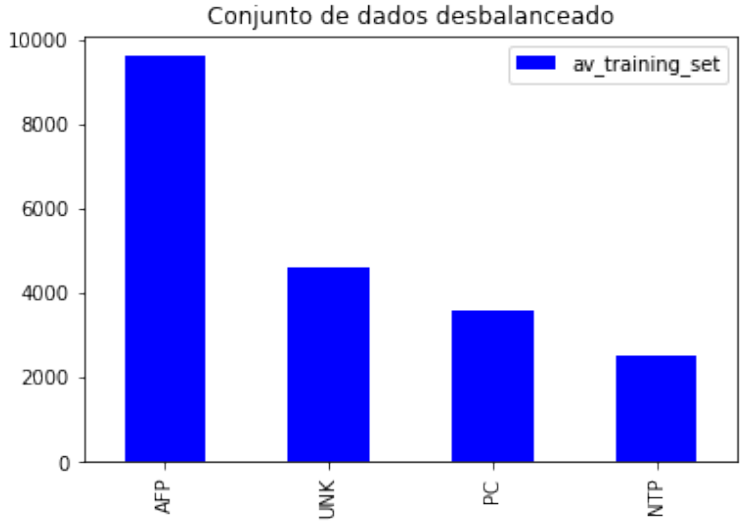
Tabela 1 – As principais colunas do arquivo CSV

Campo	**Descrição**
ROWID	ID inteiro da linha na tabela TCE
KEPID	ID Kepler da estrela alvo
TCE_PLNT_NUM	número TCE dentro da estrela alvo
TCE_PERIOD	período do evento detectado, em dias
TCE_TIME0BK	o tempo correspondente ao centro do primeiro evento detectado no dia juliano baricêntrico (BJD) menos um deslocamento constante de 2.454.833,0 dias
TCE_DURATION	duração do evento detectado, em horas
AV_TRAINING_SET	rótulo do conjunto de treinamento do Autovetter; um de PC (candidato a planeta), AFP (falso positivo astrofísico), NTP (fenômeno não transitório), UNK (desconhecido)

Fonte: Elaborado pelo autor.

No total, o conjunto de dados apresenta 3.600 candidatos a planetas (PC), 9.596 falsos positivos astrofísicos (AFP) e 2.541 fenômenos sem trânsito (NTP).

Figura 12 – Gráfico da frequência dos rótulos no conjunto de dados CSV.



Fonte: Elaborado pelo autor.

Como se observa (Figura 12), o conjunto de dados **CSV** é altamente desbalanceado, fazendo com que a precisão não seja uma medida muito útil para avaliação do modelo, além de causar o deslocamento de uma superfície de decisão aprendida, favorecendo a classe majoritária. Mas Shallue and Vanderburg (2018) argumentam que é possível se obter uma alta precisão simplesmente classificando tudo como não positivo, ou seja, não planeta.

Em seguida, durante a fase de pré-processamento realizado pelo próprio **AstroNet**, os rótulos são binarizados como:

- **planetas**: candidato a planeta - "1 = *planet*";
- **não planeta**: AFP/NTP - "0 = *not planet*".

Para completar o conjunto de treinamento do modelo, também são utilizadas as **Curvas de Luz** (imagens) das estrelas correspondentes aos **TCEs** do conjunto de dados **CSV**, geradas pela missão **Kepler** e disponíveis no **Mikulski Archive Space Teles- copes**[2] (MAST). As **Curvas de Luz** são explicadas com mais detalhes na próxima subseção.

Por fim, os dois conjuntos de dados (tabela e imagens) são agrupados e então dividido aleatoriamente em três subconjuntos, sendo: treinamento (80%), validação (10%) e teste (10%), gerando assim oito arquivos para o treinamento, um arquivo para testes e um arquivo para validação do modelo.

Os conjuntos de validação e de testes, foram utilizados por Shallue and Vanderburg (2018), respectivamente, durante o desenvolvimento do modelo para a escolha dos hiperparâmetros e para a avaliação de desempenho final do modelo.

3.3.3 Curvas de luz do Kepler

As **Curvas de Luz** baixadas da plataforma do **MAST**, são produzidas pela equipe do *pipeline* **Kepler**, que calibra os *pixels* das imagens das **Curvas de Luz**, identifica aberturas fotométricas ideais, realiza fotometria de abertura simples (estacionária) e remove artefatos instrumentais de modo comum.

Cada **Curva de Luz** consiste em medições de fluxo integradas e espaçadas em intervalos de 29,4 minutos por até quatro anos, o que equivale a aproximadamente a 70.000 pontos. As **Curvas de Luz** são então armazenadas em arquivos **.FITS**, do inglês **Flexible Image Transport System**, que é o formato de dados padrão usado em astronomia.

Na sequência, algumas etapas adicionais são necessárias para preparar as **Curvas de Luz** que são utilizadas como entradas para a **rede neural** do **AstroNet**. Para cada **TCE** do conjunto de treinamento, são removidos pontos correspondentes a trânsitos de outros planetas já confirmados no sistema estelar observado (alvo). Depois, essa **Curva de Luz** é "achatada", ou seja, é removida a variabilidade de baixa frequência, ajustando um *spline*[3] básico à **Curva de Luz** e dividindo-o pelo *spline* de melhor ajuste. Para

[2] https://archive.stsci.edu/missions-and-data. Acesso em 19 set. 2022.

[3] Um ***spline*** é uma curva definida matematicamente por dois ou mais pontos de controle. Os pontos de controle que ficam na curva são chamados de nós. Os demais pontos definem a tangente à curva em seus respectivos nós.

preservar os trânsitos, são removidos os pontos em trânsito do **TCE** enquanto o *spline* é ajustado e interpolado linearmente sobre esses trânsitos, ajuste esse feito de forma iterativa, removendo assim os *outliers* e reajustando o *spline* enquanto é interpolado sobre esses *outliers*, evitando assim que o *spline* seja "puxado" por pontos discrepantes como, por exemplo, impactos de raios cósmicos.

Cada **Curva de Luz** do **Kepler** tem diferentes características de variabilidade estelar de baixa frequência, portanto, usar um espaçamento de ponto de interrupção para todas as **Curvas de Luz** é abaixo do ideal. Em vez disso, é definido o espaçamento ideal dos pontos de quebra de *spline* para cada **Curva de Luz** ajustando *splines* com diferentes espaçamentos de ponto de quebra, calculando o "**critério de informação Bayesiano**" (BIC, do inglês *Bayesian information criterion*) para cada *spline* e escolhendo o espaçamento de ponto de quebra que minimiza o **BIC**.

3.3.4 Representações de entrada

Para cada **TCE**, foram preparadas duas representações de **Curvas de Luz** normalizadas em profundidade e dobradas em fase ("achatamento"), ou seja, foram geradas entradas para a **rede neural** dobrando cada **Curva de Luz** achatada no período **TCE** (com o evento de trânsito centralizado) e *binning*[4] para produzir um **vetor 1D**.

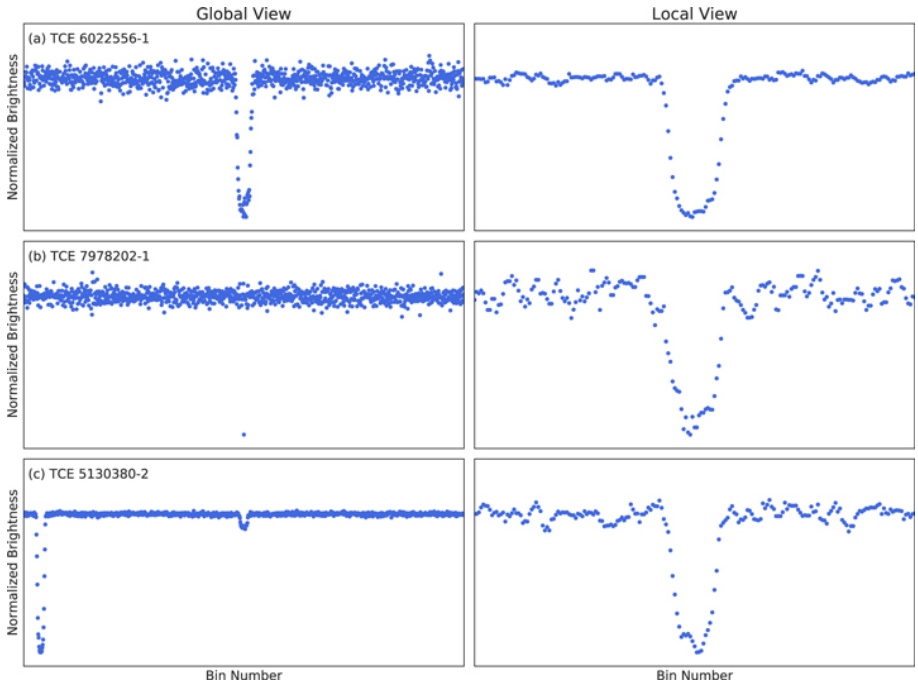
Para agrupar uma **Curva de Luz** dobrada ("achatamento"), foi definida uma sequência de intervalos uniformes no eixo do tempo com largura $[\delta]$ e distância $[\lambda]$ entre os pontos médios e calculado o fluxo médio dos pontos dentro de cada intervalo. Se escolhermos $[\delta] = [\lambda]$, então os intervalos dividem o eixo do tempo: cada ponto cai precisamente em um *bin*. Se escolhermos $[\delta] > [\lambda]$, os *bins* se sobrepõem, o que reduz a dispersão e torna alguns trânsitos mais visíveis. A escolha mais simples de $[\lambda]$ seria uma constante fixa, mas isso resultaria em uma ampla faixa de comprimentos de entrada devido à ampla faixa de períodos **TCE**. Ao invés disso, foram escolhidos dois valores específicos de **TCE** para $[\lambda]$, sendo que cada um deles gera uma "visão" diferente da **Curva de Luz**.

A "visão global" (*global view*) é uma representação de comprimento fixo de toda a **Curva de Luz**. Para gerar uma "visão global" da **Curva de Luz**, define-se $[\lambda]$ como uma fração do período **TCE**. Então todas as **Curvas de Luz** são agrupadas no mesmo comprimento e cada *bin* representa o mesmo número de pontos, em média, nas **Curvas de Luz**. Uma desvantagem é que os **TCEs** de longo período podem acabar com trânsitos muito estreitos que caem inteiramente dentro de um pequeno número de *bins* (Figura 13 (b)).

Já a "visão local" (*local view*) é uma representação de comprimento fixo de uma

[4] O ***binning***, ou o armazenamento de dados, também chamado de armazenamento discreto de dados ou agrupamento de dados, é uma técnica de pré-processamento de dados usada para reduzir os efeitos de pequenos erros de observação.

Figura 13 – Representações de curvas de luz usadas como entradas para o modelo de rede neural.



Fonte: (SHALLUE; VANDERBURG, 2018).

janela ao redor do trânsito detectado. Para gerar uma "visão local" do trânsito, escolhe-se [λ] como uma fração da duração do **TCE**. É considerado então k durações de trânsito em cada lado do evento, de modo que o trânsito ocupe uma fração fixa do vetor resultante. Essa técnica representa **TCEs** de curto e longo período igualmente, mas analisa apenas parte da **Curva de Luz** e, portanto, pode perder informações importantes, como eclipses secundários (Figura 13 (c)).

Finalmente, todas as **Curvas de Luz** são normalizadas para terem mediana igual a **0** (zero) e valor mínimo igual a **-1** (menos um), fazendo com que todos os **TCEs** tenham uma profundidade de trânsito fixa.

3.4 Bases de Dados

Este trabalho faz uso das bases de dados do **NASA Exoplanet Archive** e do **Arquivo Mikulski para Telescópios Espaciais** (MAST), do inglês *Mikulski Archive for Space Telescopes*.

A primeira base da dados é o **NASA Exoplanet Archive**, operado pelo **Instituto de Tecnologia da California**, sob contrato com a **NASA** no âmbito do **Programa de Exploração de Exoplanetas**, é um catálogo astronômico *on-line* de exoplanetas e estrelas e um serviço de dados, que agrupa e correlaciona dados astronômicos e informações sobre exoplanetas e suas estrelas hospedeiras, fornecendo ferramentas para se trabalhar

com esses dados.

Este arquivo se dedica a coletar e a servir, publicamente, importantes conjuntos de dados que podem ser utilizados na busca e caracterização de exoplanetas e suas estrelas hospedeiras. Esses dados incluem **parâmetros estelares**, como posições, magnitudes e temperaturas, **parâmetros de exoplanetas**, como massas e parâmetros orbitais, e **dados de descoberta/caracterização**, como curvas de velocidade radial, curvas de luz fotométricas, imagens e espectros.

A outra base de dados, o **MAST** é uma plataforma criada para distribuição dos dados produzidos pelos telescópios espaciais, onde temos acesso a todos os dados produzidos pelo telescópio **Kepler** e que foram utilizados tanto para o treinamento do modelo **AstroNet**, como para realizar previsões com novos dados nunca vistos antes pelo modelo.

Em 1997, o **Space Telescope Science Institute** (STScI), da **NASA**, já operava como distribuidor dos dados do **Telescópio Espacial Hubble** (HST) e dados da missão **International Ultraviolet Explorer** (IUE). Após o fim da missão **IUE**, a **NASA** busca uma forma de arquivar esses dados e combiná-los com os dados do **HST**. Assim, a **NASA** fez do **STScI** o seu centro de arquivo para todos dados de missões espaciais semelhantes ao **HST** e a **IUE**, que trabalharam na faixa de ultravioleta, óptica e infravermelho próximo, e o nome **Multimission Archive at STScI** (**MAST**) foi escolhido para transmitir o foco mais amplo do arquivo estendido.

Em abril de 2012 o arquivo foi renomeado para **Mikulski Archive for Space Telescopes**, permanecendo sua sigla, como forma de homenagear a senadora **Barbara A. Mikulski**, por suas realizações ao longo da sua carreira e por se tornar a mulher com mais tempo de serviço na história do **Congresso dos EUA**.

O **MAST** suporta uma variedade de arquivos de dados astronômicos, principalmente com foco em conjunto de dados (*datasets*) cientificamente relacionados nas partes ópticas, ultravioleta e infravermelho próximo do espectro. Assim, arquiva uma variedade de dados espectrais (faixa de comprimento de onda de dados espectroscópicos e faixa de magnitude versus resolução espectral) e de imagens (campo de visão para detectores de imagem **HST** e fluxo mínimo detectável versus comprimento de onda para Missões de Pesquisa). Quanto ao formato dos arquivos, quase todos os conjuntos de dados arquivados são armazenados usando o formato **Flexible Image Transport System** (**FITS**).

Atualmente o **MAST** hospeda dados de mais de uma dúzia de missões como **James Webb**, **Hubble**, **TESS**, **Kepler**, **Gaia** e, no futuro, **Roman**. Embora não haja nenhum custo envolvido na recuperação de dados do **MAST**, é solicitado aos pesquisadores incluir um reconhecimento nas publicações que fazem uso do **MAST**.

Figura 14 – Tela inicial do *website* do **MAST**

Fonte: <https://archive.stsci.edu>.
Acesso em: 18 set. 2022.

3.5 MÉTRICAS DE AVALIAÇÃO

Apresentamos aqui as métricas utilizadas para a avaliação do desempenho do **AstroNet**, aplicadas originalmente pelos pesquisadores nas três arquiteturas de redes neurais que desenvolveram, para assim poderem escolher o melhor modelo.

Todas as métricas utilizadas variam no intervalo [0,1] (zero e um), o que significa que quanto mais próximo de 1 (um), melhor será o modelo.

Particularmente, foram atribuídos a cada sinal (objeto) uma das quatro categorias:

- **Verdadeiro Positivo (VP)**: sinais classificados corretamente como “candidato a planeta”;
- **Verdadeiro Negativo (VN)**: sinais classificados corretamente com “não planeta”;
- **Falso Positivo (FP)**: sinais classificados incorretamente como “candidato a planeta”;
- **Falso Negativo (FN)**: sinais classificados incorretamente como “não planeta”.

3.5.1 Precisão (*precision*)

É a fração de sinais classificados como planetas que são planetas verdadeiros (também conhecido como confiabilidade). A fórmula utilizada para o cálculo da precisão é apresentada a seguir:

$$PRECISÃO = \frac{VP}{VP + FP}\,, \tag{3.3}$$

Quanto maior o valor da precisão, ou seja, próximo ao valor um, melhor a capacidade do modelo classificar uma observação como positiva que realmente é positiva.

3.5.2 Revocação (*recall*) ou Sensibilidade

É a fração de planetas verdadeiros que são classificados como planetas (também conhecido como completude). A fórmula utilizada para o cálculo da revocação é apresentada a seguir:

$$REVOCAÇÃO = \frac{VP}{VP + FN}\,, \tag{3.4}$$

Quanto maior o valor da revocação, ou seja, próximo ao valor um, melhor a capacidade do modelo em prever a classe positiva.

3.5.3 Acurácia (*accuracy*)

É a fração de classificações corretas. A fórmula utilizada para o cálculo da acurácia é apresentada a seguir:

$$ACURÁCIA = \frac{VP + VN}{VP + VN + FP + FN} \tag{3.5}$$

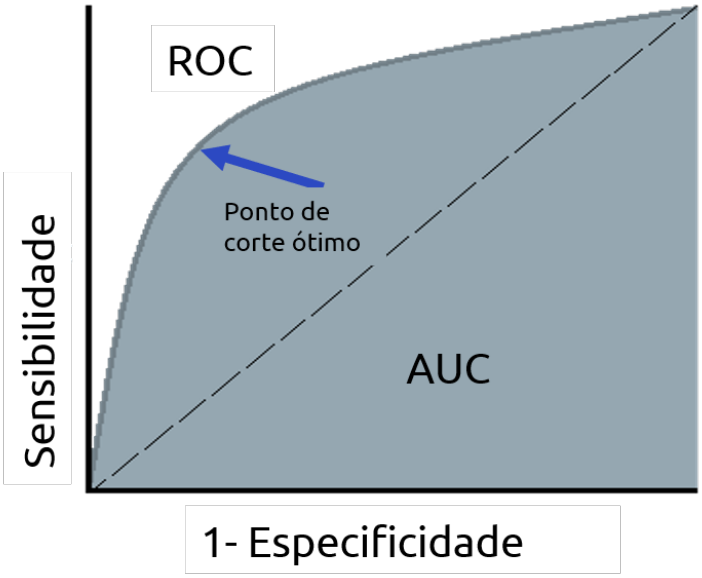
Quanto maior o valor da acurácia, ou seja, próximo ao valor um, melhor a capacidade do modelo fazer previsões corretas.

3.5.4 AUC (*Area Under the Curve*)

É a área sob a curva característica de operação do receptor, que é equivalente à probabilidade de um planeta selecionado aleatoriamente ter uma pontuação mais alta do que um falso positivo selecionado aleatoriamente.

O AUC é uma medida decorrente da curva ROC (*Receiver Operating Characteristic Curve*), que é uma curva gerada pela taxa de verdadeiros positivos (sensibilidade, igual a revocação) e pela taxa de falsos positivos (1 – especificidade) para diferentes pontos de

Figura 15 – AUC e Curva ROC.



Fonte: (SCUDILIO, 2020).

cortes (**Figura 15**). A fórmula utilizada para o cálculo da especificidade é apresentada a seguir:

$$ESPECIFICIDADE = \frac{VN}{VN + FP}, \tag{3.6}$$

3.6 VALIDAÇÃO

Os pesquisadores utilizaram o conjunto de validação durante o desenvolvimento para escolher os hiperparâmetros do modelo. Com isso, utilizaram a plataforma **Vizier**[5] do **Google** para otimização de caixa preta e ajustar automaticamente os hiperparâmetros, incluindo os hiperparâmetros para as representações de entrada (por exemplo, número de compartimentos, largura do compartimento), arquitetura do modelo (por exemplo, o número de camadas conectadas, número de camadas convolucionais, tamanho do *kernel*), e treinamento (por exemplo, probabilidade de *dropout*).

Cada "estudo" do **Google Vizier** treinou vários milhares de modelos para encontrar a configuração do hiperparâmetro que maximizasse a área sob a curva característica de operação do receptor sobre o conjunto de validação. Cada modelo foi treinado em uma única Unidade Central de Processamento (CPU, do inglês *Central Processing Unit*), e o tempo de treinamento variou de 10 minutos a 5 horas, dependendo do tamanho do modelo. O melhor modelo levou 90 minutos para treinar. Para acelerar a busca de hiperparâmetros,

[5] https://cloud.google.com/ai-platform/optimizer/docs/overview/. Acesso em 5 out. 2022.

utilizaram 100 CPUs por estudo para treinar modelos individuais em paralelo. Foram executados muitos estudos do **Vizier** durante o desenvolvimento do modelo à medida que iam melhorando iterativamente as representações de entrada e decisões de design. Por fim, ajustaram cada combinação de arquitetura e representação de entrada separadamente.

3.7 ASPECTOS COMPUTACIONAIS

Para o desenvolvimento deste trabalho, o modelo implementado foi baseado no código disponibilizado publicamente por Shallue and Vanderburg (2018) em repositório *online* **GitHub**[6] e utilizou-se da linguagem **Python** e do **TensorFlow**[7], que é uma biblioteca de *software* gratuita e de código aberto para **Aprendizado de Máquina** e **Inteligência Artificial**, com foco no treinamento e inferência de redes neurais profundas, originalmente desenvolvido por pesquisadores e engenheiros da equipe do **Google Brain**, para uso interno do **Google**[8].

Ainda, diante de todo conhecimento e experiência que desenvolvemos e descrevemos com o **AstroNet**, procuramos implementar este modelo em **Docker**[9], que é um *software* de código aberto utilizado para implementar aplicações dentro de contêineres virtuais e possibilitar sua portabilidade e reprodução em outros ambientes. Assim, foi gerada uma imagem com todos os pacotes necessários instalados para execução do **AstroNet**, de forma que esta imagem ficará disponível publicamente, também em repositório *online*, para que outros pesquisadores que se interessem pelo mesmo assunto, possam reproduzi-lo e, caso achem necessário, realizarem suas próprias adaptações e chegando a novas conclusões e modelos melhores.

6 https://github.com/cshallue/exoplanet-ml/. Acesso em 24 ago. 2022.
7 TensorFlow. Disponível em: https://www.tensorflow.org/.
8 https://en.wikipedia.org/wiki/TensorFlow/. Acesso em 5 out. 2022.
9 Docker. Disponível em: https://www.docker.com/.

4 RESULTADOS

Nesse capítulo apresentamos os resultados obtidos após os experimentos realizados com a reimplementação e reavaliação do melhor modelo do **AstroNet** desenvolvido por Shallue and Vanderburg (2018), que utiliza uma arquitetura de **Rede Neural Convolucional** com **visualizações de entrada globais e locais**. Sua configuração exata foi apresentada no capítulo anterior (**Figura 11**).

4.1 Descrição do experimento

O desenvolvimento dos experimentos foram realizados em contêineres **Docker**, executados a partir de uma única imagem criada com os seguintes pacotes instalados:

- Python versão 3.6;
- TensorFlow versão 1.15.0;
- Pandas;
- NumPy;
- SciPy;
- AstroPy;
- PyDl;
- Bazel;
- Vim.

O código do modelo **AstroNet** é desenvolvido em **Python** puro e essa linguagem foi mantida nos experimentos, porém foi necessário pequenas adaptações para ser executado, de forma compilada, utilizando-se a ferramenta **Bazel**[1], nos contêineres **Docker**.

Tanto a imagem criada em **Docker**, bem como o código adaptado, os dados pré-treinados e a documentação para utilização do **AstroNet**, foram disponibilizados publicamente no **DockerHub**[2] e no **GitHub**[3], respectivamente. Com isso, outros pesquisadores também podem realizar seus próprios experimentos.

[1] Bazel é uma ferramenta de *software* de código aberto, desenvolvida pelo Google, para automatizar processos e testes de criação de software. Disponível em: https://bazel.build/.

[2] https://hub.docker.com/r/mdaudt/astronet-image

[3] https://github.com/marcelodaudt/astronet

Foram realizados dois experimentos, se diferenciando entre eles pela forma como os dados foram obtidos para treinar e testar o modelo, sendo utilizado no primeiro experimento um conjunto de dados pré-computado e no segundo experimento obtendo-se novos dados e gerando um conjunto de dados balanceado.

4.1.1 Treinamento

No primeiro experimento foi utilizado para treinar o modelo, um conjunto de dados pré-computado, obtido no **Github** do **Christopher Shallue**, no formato **TFRecord** para **TensorFlow**, que são arquivos de formato simples para armazenar uma sequência de registros binários (serializados). Este conjunto de dados pré-computado possuí os dados desbalanceados e foi dividido aleatoriamente em três subconjuntos: treinamento (80%), validação (10%) e teste (10%).

No segundo experimento, um novo conjunto de dados foi obtido através da plataforma **MAST**, sendo composto por um subconjunto de todos os dados do telescópio **Kepler**. Esse conjunto de dados é uma tabela no formato **CSV** com 20.367 **TCEs**[4] rotulados, sendo que estes rótulos foram atribuídos por humanos e classificam um **TCE** em:

- **Planeta**: PC = *planet candidate*;
- **Não Planeta**: AFP = *astrophysical false positive*, NTP = *non-transiting phenomena*.

Deste conjunto de dados com 20.367 **TCEs**, foi selecionada uma amostra aleatória com 1.000 exemplos de cada rótulo, tendo como propósito o treinamento e a avaliação do modelo com os dados agora balanceados (Figura 16).

Os **TCEs** que possuíam os rótulos "desconhecido" (UNK = *unknow*) foram desprezados quando do processamento dos dados pelo modelo, ao final desta etapa.

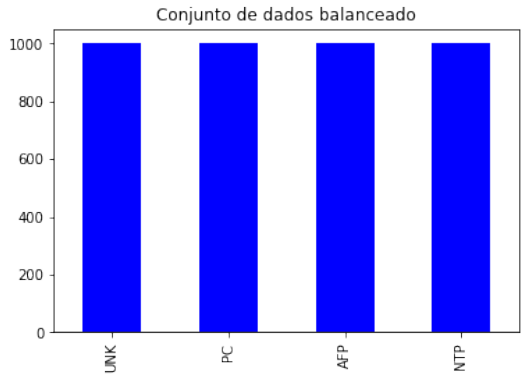
Em seguida, um segundo conjunto de dados foi obtido, que são as **Curvas de Luz** das estrelas correspondentes aos **TCEs** do conjunto de dados em **CSV**. A partir deste arquivo **CSV**, é gerado um *script*, criado pelo próprio modelo, que tem como função baixar todas as curvas de luz da plataforma **MAST**.

Este conjunto de dados possui mais de 11GB de tamanho e está no formato **.FITS**. Observando que cada arquivo **.FITS** corresponde a um quarto (trimestre) de uma estrela específica, porém alguns quartos podem ser divididos em vários arquivos **.FITS**.

Por fim, todos os dados, dos dois conjuntos de dados, são processados para gerarem arquivos no formato **TFRecord**, sendo oito arquivos para o treinamento (80%) e um

[4] TCE (Evento de Cruzamento de Limiar): é um sinal periódico que foi detectado em uma curva de luz de uma estreal específica e está associado a um período, uma duração, uma época e possivelmente metadados adicionais.

Figura 16 – Gráfico da frequência dos rótulos no conjunto de dados após balanceamento.



Fonte: Elaborado pelo autor.

arquivo para a validação (10%). Também foi gerado um arquivo que será utilizado para testes do modelo (10%).

Porém, como não haveria tempo hábil para a realização deste experimento com os 1.000 exemplos de cada rótulo, optou-se por gerar esse novo conjunto de dados balanceado para o treinamento do modelo com somente 50 exemplos, porém os resultados não foram nada satisfatórios.

O modelo foi treinado, nos dois experimentos, por 625 épocas (passos) e seu tempo de execução médio foi de 40 minutos.

4.1.2 Avaliação

Para a realização desta etapa do experimento, o modelo foi avaliado utilizando-se os arquivos no formato **TFRecord** (test-00000-of-00001), citados anteriormente.

No **Experimento 1**, os resultados obtidos pelo modelo neste conjunto de teste são apresentados na **Tabela 2**. Nesta tabela também são apresentados os resultados obtidos por Shallue and Vanderburg (2018), como forma de comparação.

Tabela 2 – Acurácia e AUC no Experimento 1 - dados desbalanceados.

	Experimento 1	**Shallue and Vanderburg (2018)**
Acurácia	0.963786	0.962515
AUC	0.987954	0.988882
Perda	0.113198	0.112445
Entropia Cruzada ponderada	0.114984	0.112952

Fonte: Elaborada pelo autor.

Lembrando que o modelo foi treinado para classificar se um determinado **TCE** é um sinal de planeta real ou foi provocado por algum outro fenômeno, como, por exemplo,

uma estrela binária.

A matriz de confusão da **Tabela 3**, resultante do **Experimento 1**, nos mostra um visão geral de como o modelo classificou os 1.574 exemplos de TCEs do conjunto de testes comparado à rotulação verdadeira da base de dados.

Tabela 3 – Matriz de confusão (0=Não Planeta, 1=Planeta) - Experimento 1.

		Valor Predito	
		0	**1**
Valor Real	**0**	1170.0	44.0
	1	13.0	347.0

Fonte: Elaborada pelo autor.

No **Experimento 2**, os resultados obtidos pelo modelo em um conjunto de dados balanceado não foram satisfatórios, como podem ser observados na **Tabela 4**. Existe assim uma necessidade de uma maior investigação para entender se esse baixo desempenho se deu pela quantidade baixa de exemplos, ou existe a necessidade de ajuste no código do modelo.

Tabela 4 – Acurácia e AUC no Experimento 2 - dados balanceados.

	Experimento 2
Acurácia	0.750000
AUC	0.772727
Perda	3.261106
Entropia Cruzada ponderada	3.261106

Fonte: Elaborada pelo autor.

A **Tabela 5**, mostra a matriz de confusão resultante do **Experimento 2**, onde o modelo classificou 15 exemplos de **TCEs** do conjunto de testes em comparação com a rotulação verdadeira da base de dados.

Tabela 5 – Matriz de confusão (0=Não Planeta, 1=Planeta) - Experimento 2.

		Valor Predito	
		0	**1**
Valor Real	**0**	9.0	2.0
	1	2.0	3.0

Fonte: Elaborada pelo autor.

4.1.3 Predição

Para realizar essa etapa do experimento, inicialmente foi utilizado um **TCE** específico da estrela **Kepler-90** (ID 11442793) no nosso modelo do **Experimento 1** já treinado, obtendo-se o seguinte valor de predição: **0.906339**. Ou seja, o modelo teve cerca de **90%** de confiança de que o **TCE** de entrada é um planeta. Para comparação, a predição apresentada por Shallue and Vanderburg (2018) para a mesma estrela, teve o valor de **0.9480018**, ou seja, teve uma confiança maior, cerca de **95%**.

Dando continuidade aos experimentos, foram feitas predições de mais onze **TCEs**, selecionados a partir de uma lista de resultados preditivos realizados por Shallue and Vanderburg (2018), que são considerados planetas confirmados ou possíveis candidatos a planetas e que o modelo previu de serem planetas com uma probabilidade maior que **0.5**.

Tabela 6 – Resultados das predições nos dois experimentos.

ID	**Experimento 1**	**Experimento 2**	**Shallue and Vanderburg (2018)**
11442793	0.868	0.000	0.942
8480285	0.915	0.000	0.941
11568987	0.944	0.000	0.920
11030475	0.903	0.000	0.858
4548011	0.834	0.000	0.852
10130039	0.837	2.371	0.764
9896018	0.832	0.000	0.707
4851530	0.839	1.454	0.613
8804283	0.867	0.000	0.604
6508221	0.755	9.963	0.595
11968463	0.610	0.000	0.509
10600261	0.542	1.445	0.507

Fonte: Elaborado pelo autor.

A tabela **Tabela 6** apresenta os resultados obtidos nos dois experimentos e quais os resultados obtidos por Shallue and Vanderburg (2018).

Como trabalhos futuros e dando continuidade aos experimentos, serão realizadas previsões com um novo conjunto de dados do **Kepler**, especificamente da sua missão estendida, denominada **K2**, que também pode ser obtido no **MAST**.

Também queremos, se possível, realizar experimentos com conjuntos de dados de outros telescópios, como, por exemplo, o telescópio **TESS** (*Transiting Exoplanet Survey Satellite*). Com isso, queremos saber se o modelo **AstroNet** continuará obtendo os excelentes resultados como os encontrados até o momento.

5 CONCLUSÃO

Fica claro que a utilização de **Aprendizado de Máquina**, especificamente as **Redes Neurais Convolucionais**, aplicadas na **Astronomia**, mostra-se uma ferramenta competente e eficiente para o tratamento do enorme repositório da dados espaciais produzidos pelo telescópio espacial **Kepler**. Porém, como a utilização de **Aprendizado de Máquina**, especificamente na identificação de exoplanetas é recente, muita pesquisa ainda pode ser realizada e a **Ciência de Dados** pode contribuir muito como o desenvolvimento da área.

A realização deste trabalho teve como um dos seus objetivos, conhecer o estado da arte na área de identificação de exoplanetas com a utilização de **Aprendizado de Máquina Profundo**, aliada a aquisição de conhecimento tecnológico para seu funcionamento.

O trabalho também teve como objetivo analisar o modelo **AstroNet**, onde foram realizados dois experimentos, sendo que o primeiro utilizou um conjunto de dados pré-computado pelos autores do **AstroNet** para treinamento e testes do modelo. O segundo experimento utilizou um conjunto de dados para treinamento e testes que foram criados a partir de dados obtidos na plataforma **MAST** e processados pelo próprio modelo.

Os resultados do treinamento e dos testes no **Experimento 1**, se mostraram bastante semelhantes com os descritos no artigo de Shallue and Vanderburg (2018). Porém, os resultados de predição obtidos neste experimento não apresentaram as mesmas semelhanças, sendo que 75% das predições tiveram resultados melhores dos que apresentadas por Shallue and Vanderburg (2018).

Com relação ao **Experimento 2**, onde foi gerado um novo conjunto de dados balancedos, com 50 exemplos de cada rótulo, tantos os resultados dos testes como das predições realizadas pelo modelo, não se mostraram nada satisfatórios, ficando assim, como sugestão para trabalhos futuros, a necessidade de uma melhor investigação dos motivos para esse baixo desempenho.

Ainda, devido a sua atual complexidade e algumas dificuldades computacionais encontradas para execução do **AstroNet**, naturais devido ao espaço de tempo entre sua criação e experimentos atuais, não houve tempo hábil para desenvolvimento de uma nova proposta de modelo, ficando como outra sugestão para trabalhos futuros.

Por fim, como última sugestão de trabalho futuro, podemos aperfeiçoar o modelo **AstroNet**, para que este também seja capaz de identificar se o planeta está na **Zona Habitável** da sua estrela hospedeira e qual é a sua composição atmosférica, somente quando este for identificado como positivo. Podemos utilizar como base para a caracterização

da composição atmosférica e da habitabilidade ideais de um exoplaneta, trabalhos como os realizados pela Dra. Raissa Estrela, pesquisadora do **Laboratório de Propulsão a Jato da NASA**, que em sua tese de doutorado, intitulada *Exoplanets atmospheres and habitability* (ESTRELA, 2020), analisa dados do **Kepler** e pode auxiliar na definição das melhores variáveis a serem analisadas. Também há um artigo publicado em colaboração, intitulado *Detection of as atmosphere on a rocky exoplanet* (SWAIN *et al.*, 2021), onde é relatada a detecção de uma atmosfera em um exoplaneta rochoso, GJ 1132 b, que é um exoplaneta semelhante à Terra em termos de tamanho e densidade.

REFERÊNCIAS

ACADEMY, D. S. **Deep Learning Book**. 2022. Available at: https://www.deeplearningbook.com.br/. Access at: 11 set. 2022.

AMARAL, F. **Introdução a ciência de dados : mineração de dados e Big Data**. Rio de Janeiro: Alta Books, 2016.

ANSDELL, M. *et al.* Scientific domain knowledge improves exoplanet transit classification with deep learning. **The Astrophysical Journal**, American Astronomical Society, v. 869, n. 1, p. L7, dec 2018. Available at: https://doi.org/10.3847/2041-8213/aaf23b.

BEURS, Z. L. d. *et al.* Identifying Exoplanets with Deep Learning. IV. Removing Stellar Activity Signals from Radial Velocity Measurements Using Neural Networks. **Bulletin of the AAS**, v. 53, n. 1, jan 11 2021. Available at: https://baas.aas.org/pub/2021n1i332p04.

CATANZARITE, J. H. Autovetter planet candidate catalog for q1-q17 data release 24. **Astronomy & Astrophysics**, 2015.

COBB, A. D. *et al.* An ensemble of bayesian neural networks for exoplanetary atmospheric retrieval. **The Astronomical Journal**, American Astronomical Society, v. 158, n. 1, p. 33, jun 2019. Available at: https://doi.org/10.3847/1538-3881/ab2390.

CUI, K. *et al.* Identify light-curve signals with deep learning based object detection algorithm. i. transit detection. **The Astronomical Journal**, American Astronomical Society, v. 163, n. 1, p. 23, dec 2021. Available at: https://doi.org/10.3847/1538-3881/ac3482.

DATTILO, A. *et al.* Identifying exoplanets with deep learning. II. two new super-earths uncovered by a neural network in k2 data. **The Astronomical Journal**, American Astronomical Society, v. 157, n. 5, p. 169, apr 2019. Available at: https://doi.org/10.3847/1538-3881/ab0e12.

DVASH, E. *et al.* Shallow transits—deep learning. II. identify individual exoplanetary transits in red noise using deep learning. **The Astronomical Journal**, American Astronomical Society, v. 163, n. 5, p. 237, apr 2022. Available at: https://doi.org/10.3847/1538-3881/ac5ea2.

ESTRELA, R. de L. F. **Exoplanets atmospheres and habitability**. 2020. 108 p. Tese (Doutorado) — Universidade Presbiteriana Mackenzie, São Paulo, 2020.

FEIGELSON, E. D.; BABU, G. J. **Modern statistical methods for astronomy : with R applications**. 3. ed. Cambridge: Cambridge University Press, 2014.

GORDON, T. A.; AGOL, E.; FOREMAN-MACKEY, D. A fast, two-dimensional gaussian process method based on celerite: Applications to transiting exoplanet discovery and characterization. **The Astronomical Journal**, American Astronomical Society, v. 160, n. 5, p. 240, nov 2020. Available at: https://doi.org/10.3847/1538-3881/abbc16.

GRUS, J. **Data science do zero**. Rio de Janeiro: Alta Books, 2016.

GüNTHER, M. N.; DAYLAN, T. Allesfitter: Flexible star and exoplanet inference from photometry and radial velocity. **The Astrophysical Journal Supplement Series**, American Astronomical Society, v. 254, n. 1, p. 13, apr 2021. Available at: https://doi.org/10.3847/1538-4365/abe70e.

KEPLER, S. O. **Astronomia e astrofísica**. 4. ed. São Paulo: Editora Livraria da Física, 2017.

KOVáCS, Z. L. **Redes neurais artificiais : fundamentos e aplicações**. São Paulo: Livraria da Física, 2006.

NASA. **Artificial Intelligence, NASA Data Used to Discover Eighth Planet Circling Distant Star**. 2017. Available at: https://www.nasa.gov. Access at: 12 jun. 2022.

PINCHUK, P.; MARGOT, J.-L. A machine learning–based direction-of-origin filter for the identification of radio frequency interference in the search for technosignatures. **The Astronomical Journal**, American Astronomical Society, v. 163, n. 2, p. 76, jan 2022. Available at: https://doi.org/10.3847/1538-3881/ac426f.

ROBNIK, J.; SELJAK, U. Kepler data analysis: Non-gaussian noise and fourier gaussian process analysis of stellar variability. **The Astronomical Journal**, American Astronomical Society, v. 159, n. 5, p. 224, apr 2020. Available at: https://doi.org/10.3847/1538-3881/ab8460.

SCUDILIO, J. **Qual a melhor métrica para avaliar os modelos de Machine Learning?** 2020. Available at: https://www.flai.com.br/juscudilio/qual-a-melhor-metrica-para-avaliar-os-modelos-de-machine-learning/. Access at: 18 set. 2022.

SHALLUE, C. J.; VANDERBURG, A. Identifying exoplanets with deep learning: A five-planet resonant chain around kepler-80 and an eighth planet around kepler-90. **The Astronomical Journal**, American Astronomical Society, v. 155, n. 2, p. 94, jan 2018. Available at: https://doi.org/10.3847/1538-3881/aa9e09.

SWAIN, M. R. *et al.* Detection of an atmosphere on a rocky exoplanet. **The Astronomical Journal**, The American Astronomical Society, v. 161, n. 5, p. 213, apr 2021. Available at: https://dx.doi.org/10.3847/1538-3881/abe879.

VALIZADEGAN, H. *et al.* ExoMiner: A highly accurate and explainable deep learning classifier that validates 301 new exoplanets. **The Astrophysical Journal**, American Astronomical Society, v. 926, n. 2, p. 120, feb 2022. Available at: https://doi.org/10.3847/1538-4357/ac4399.

YIP, K. H. *et al.* Peeking inside the black box: Interpreting deep-learning models for exoplanet atmospheric retrievals. **The Astronomical Journal**, American Astronomical Society, v. 162, n. 5, p. 195, oct 2021. Available at: https://doi.org/10.3847/1538-3881/ac1744.

YU, L. *et al.* Identifying exoplanets with deep learning. III. automated triage and vetting of tess candidates. **The Astronomical Journal**, American Astronomical Society, v. 158, n. 1, p. 25, jun 2019. Available at: https://doi.org/10.3847/1538-3881/ab21d6.

ZHANG, Y.; ZHAO, Y. Astronomy in the big data era. **Data Science Journal**, n. 14, p. p. 1–9, 2015.

ZINGALES, T.; WALDMANN, I. P. ExoGAN: Retrieving exoplanetary atmospheres using deep convolutional generative adversarial networks. **The Astronomical Journal**, American Astronomical Society, v. 156, n. 6, p. 268, nov 2018. Available at: https://doi.org/10.3847/1538-3881/aae77c.