Universidade de São Paulo
Instituto de Astronomia, Geofísica e Ciências Atmosféricas
Departamento de Astronomia

Clara Amorim Navarro

Identificação de Estrelas Be em Levantamentos Fotométricos utilizando Aprendizado de Máquina

São Paulo
2025

Clara Amorim Navarro

Identificação de Estrelas Be em Levantamentos Fotométricos utilizando Aprendizado de Máquina

Trabalho de Conclusão de Curso apresentado ao Instituto de Astronomia, Geofísica e Ciências Atmosféricas da Universidade de São Paulo como requisito parcial para a obtenção do título de Bacharel em Astronomia.

Vertente: Computacional
Orientador(a): Prof. Dr. Alex Cavaliéri Carciofi (IAG/USP)

São Paulo
2025

Ao meu pai,
que me ensinou a sonhar,
e segue sendo minha *Polaris*.

À minha amiga Bruna,
que me presenteou com um "Pálido Ponto Azul",
e todo um universo de saudade.

Agradecimentos

Às minhas irmãs mais novas, Catharina e Helena, por serem minha maior felicidade, e darem sentido à minha vida. Cada conquista minha é também de vocês;

À minha mãe, Elizabete, por sempre encontrar um caminho onde parecia não haver nenhum. Obrigada por todo amor, pelos sacrifícios e por sempre acreditar em mim;

Ao meu orientador, Prof. Alex, pelos conselhos, pela paciência, e por nunca duvidar das minhas capacidades, mesmo quando eu mesma duvidei. Serei sempre grata por cada ensinamento e oportunidade;

Aos meus amigos de uma vida, em especial Isa, Matheus, Lu e Leo, por crescerem comigo e acompanharem minha trajetória, dos primeiros passos até hoje;

Aos meus colegas de grupo e aos de turma, Felipe, Luana, Tajan, Vic, Alice, Gui, André, Pamela, Amanda, Matheus, Ariane, Carlos, Du, Rafa e Vitor, pelo ambiente acolhedor de aprendizado mútuo, e cumplicidade de quem enfrentou as mesmas provas, prazos e desafios;

Aos meus gatos, Morgana e Jupi, pelos ronronos terapêuticos e "ajuda" na escrita;

Ao Pedro Yudi, cujo apoio foi tão fundamental, que merece um volume inteiro de agradecimentos. Você viveu meus desafios como seus, e meus sonhos como nossos. Esta conquista é tanto sua quanto minha, e sem você nada disso seria possível. Obrigada por tudo — pelo que estas palavras podem dizer, e por tudo aquilo que só o coração sabe;

À FAPESP, pelo apoio financeiro, sob o projeto nº: 2023/12720-0;

Ao Instituto de Astronomia, Geofísica e Ciências Atmosféricas da USP e funcionários – professores, técnicos, administrativos, e equipes de limpeza e segurança – pela infraestrutura, suporte e pelo ambiente acadêmico que foram fundamentais para a minha formação.

"There is no problem in science that can be solved by a man that cannot be solved by a woman."

Vera Rubin

Resumo

Estrelas Be são objetos astrofísicos caracterizados pela presença transitória de um disco circunstelar gasoso, que produz linhas de emissão no espectro e padrões complexos de variabilidade fotométrica. Tradicionalmente, sua identificação depende de espectroscopia, o que limita sua detecção em grandes *surveys*. Com o crescimento de levantamentos fotométricos massivos, como o *Legacy Survey of Space and Time* (LSST), conduzido pelo *Vera C. Rubin Observatory*, torna-se necessário desenvolver métodos automatizados para classificar essas estrelas com base em curvas de luz.

Neste trabalho é investigado o uso de aprendizado de máquina supervisionado para essa tarefa, utilizando curvas de luz do *Optical Gravitational Lensing Experiment* (OGLE), rotuladas manualmente por Figueiredo et al. (2025). Foram aplicados modelos tradicionais (*Random Forest*, *eXtreme Gradient Boosting*, *k-Nearest Neighbors*, *Support Vector Machine* e *MultiLayer Perceptron*) que usam valores numéricos, além de redes neurais convolucionais (CNNs), que processam as imagens das curvas de luz. Os modelos tradicionais alcançaram acurácias entre 81% e 86%, e os atributos com maior importância foram os variogramas e índice Stetson J. A CNN binária obteve acurácia de 88%, superando ligeiramente os modelos tradicionais, enquanto uma versão multiclasse, que também classificava as orientações das estrelas, atingiu 71% de acurácia.

Os resultados demonstram que ambas as abordagens são viáveis para identificação fotométrica de estrelas Be, com a CNN mostrando potencial para classificação direta a partir de imagens, sem necessidade de extração manual de atributos. O *pipeline* metodológico desenvolvido constitui uma base sólida para aplicação futura em projetos de larga escala como o LSST.

Abstract

Be stars are astrophysical objects characterized by the transient presence of a gaseous circumstellar disk, which produces emission lines in their spectra and complex patterns of photometric variability. Traditionally, their identification relies on spectroscopy, which limits detection in large surveys. With the growth of massive photometric projects, such as the Legacy Survey of Space and Time (LSST) conducted by the Vera C. Rubin Observatory, it becomes necessary to develop automated methods to classify these stars based on light curves.

This work investigates the use of supervised machine learning for this task, using light curves from the Optical Gravitational Lensing Experiment (OGLE), manually labeled by Figueiredo et al. (2025). Traditional models (Random Forest, eXtreme Gradient Boosting, k-Nearest Neighbors, Support Vector Machine, and MultiLayer Perceptron), which operate on numerical features, were applied alongside convolutional neural networks (CNNs) that process images of the light curves. The traditional models achieved accuracies between 81% and 86%, and the most relevant features were variograms and the Stetson J index. The binary CNN obtained an accuracy of 88%, slightly outperforming the traditional models, while a multiclass version, which also classified stellar orientations, reached 71% accuracy.

The results show that both approaches are viable for the photometric identification of Be stars, with CNNs demonstrating potential for direct classification from images without the need for manual feature extraction. The methodological pipeline developed here provides a solid foundation for future applications in large-scale projects such as the LSST.

Lista de Figuras

Lista de Tabelas

Sumário

Capítulo 1

Introdução

1.1 Estrelas Be e a era da fotometria massiva

A descoberta das estrelas Be remonta ao século XIX, quando o padre Angelo Secchi observou uma "linha luminosa muito brilhante" no espectro da estrela γ Cassiopeiae (B0.5 IV), hoje reconhecida como a primeira estrela Be identificada. Inicialmente, a classificação dessas estrelas era meramente taxonômica: qualquer estrela do tipo espectral B que apresentasse linhas de emissão no espectro, especialmente nas linhas de Balmer do hidrogênio, era agrupada sob a designação "Be". Com o tempo, no entanto, ficou evidente que esse grupo abrigava objetos com naturezas muito distintas, o que levou à diferenciação entre subclasses, como as estrelas B[e], as estrelas Ae/Be de Herbig e outras classes peculiares.

Dentre essas, existem as chamadas estrelas Be clássicas, que são o foco deste trabalho. Estas são estrelas de tipo B na sequência principal ou próximas a ela, que exibem rotação elevada e, em algum momento de sua vida, formam um disco circunstelar gasoso, associado às linhas de emissão observadas no espectro, a partir de material ejetado da própria estrela.

A identificação tradicional dessas estrelas é espectroscópica, pois detecta suas linhas de emissão características. No entanto, esse método apresenta limitações práticas e logísticas: é caro, demanda tempo de telescópio em instrumentos especializados e, mais significativamente, não é viável para aplicação em larga escala. A fotometria, por outro lado, permite obter dados de milhares de estrelas em uma observação e acompanhar sua variabilidade ao longo do tempo, de forma mais acessível graças ao menor custo e à maior abundância dos dados.

Nesse âmbito, entram levantamentos fotométricos como o OGLE ("*Optical Gravitational Lensing Experiment*", Udalski et al. 1992), ASAS ("*All Sky Automated Survey*",

Pojmanski 1997) e KELT ("*Kilodegree Extremely Little Telescope*", Pepper et al. 2004), que atualmente fornecem bases de dados para o estudo de variabilidade estelar. A próxima fronteira na escala de volume de dados será inaugurada com a chegada do LSST ("*Legacy Survey of Space and Time*", Ivezic et al. 2019) do Observatório Vera C. Rubin, que, sozinho, produzirá milhões de alertas de variabilidade por noite. Diante desse momento de explosão de dados na Astronomia, torna-se cada vez mais necessário o desenvolvimento de métodos alternativos e automatizados para identificar objetos de interesse científico, por exemplo, estrelas Be, e usando dados mais acessíveis e abundantes, como curvas de luz.

No caso das estrelas Be, a presença (ou ausência) de um disco circunstelar, bem como os eventos que levam à sua formação ou dissipação, formam assinaturas características nas suas curvas de luz, as quais refletem fases de atividade do disco, erupções de brilho, construção e dissipação de material, entre outros. Como diferentes padrões de variabilidade são reconhecidos como indicativos da natureza Be, isso permite que seja possível identificar manualmente (com certa precisão) essas estrelas a partir de suas curvas de luz. Um exemplo desse tipo de abordagem é o trabalho de Figueiredo et al. (2025), no qual cerca de 3000 curvas de luz do levantamento OGLE foram analisadas visualmente, resultando na identificação de 1751 candidatas a estrelas Be.

A Figura 1.1, retirada de Figueiredo et al. (2025), apresenta algumas das assinaturas típicas das curvas de luz de estrelas Be. Nela, observa-se a separação dos dados fotométricos em diversos estágios de atividade, relacionados ao ciclo de vida do disco circunstelar, como crescimento, dissipação, platô, entre outros. As fases de variabilidade correspondentes são indicadas pela cor dos pontos, e linhas verticais e horizontais sinalizam as diferentes fases de observação ou as transições entre os estágios de variabilidade. Cada painel mostra um exemplo distinto, apresentando padrões usuais de variabilidade observados nestas estrelas.

Amostras rotuladas manualmente, como a criada por Figueiredo et al. (2025), são valiosas para a criação de técnicas computacionais de classificação automática, pois fornecem um padrão de referência confiável, fundamentado na experiência humana. O uso de métodos de aprendizado de máquina (*machine learning*, ML) nesses dados permite avaliar o quanto os algoritmos reproduzem esses critérios de forma automatizada e possibilita sua aplicação futura em levantamentos de larga escala, como o LSST.

Assim, nesse cenário, o aprendizado de máquina surge como uma alternativa promissora para a classificação de estrelas Be em levantamentos fotométricos. Ao treinar modelos

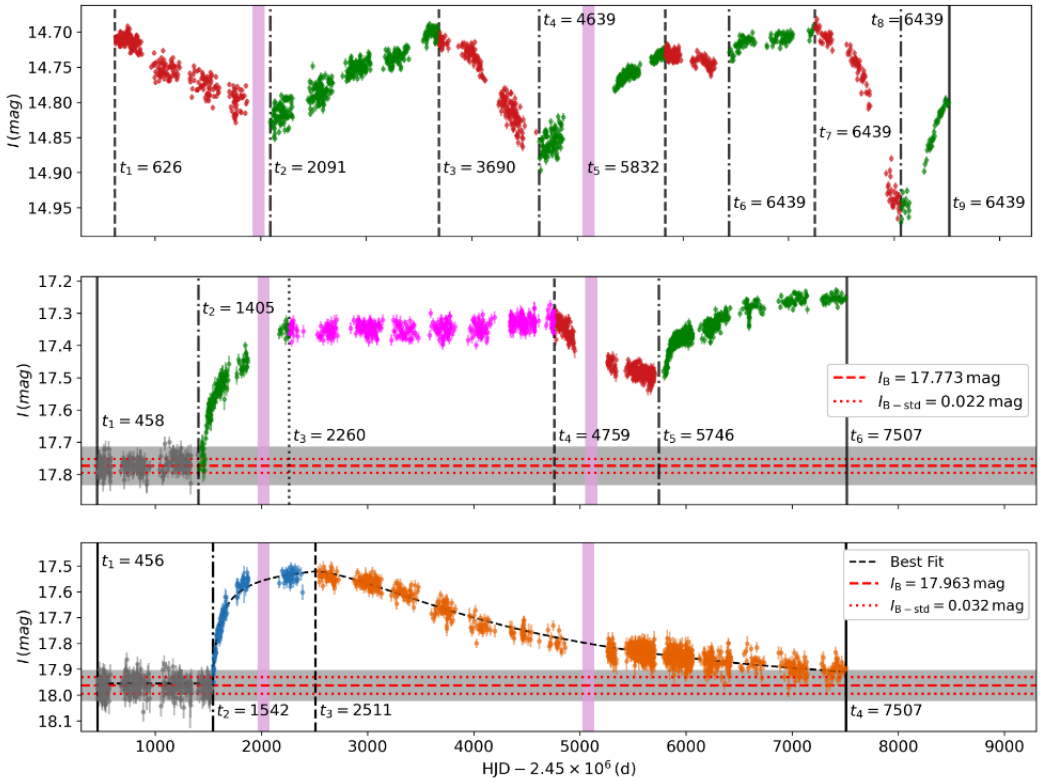

Figura 1.1: Exemplos de curvas de luz de estrelas Be, classificadas segundo as fases de variabilidade do disco circunstelar. Cada cor representa uma fase: cinza (linha de base, sem disco), azul (crescimento do disco), laranja (dissipação do disco), magenta (platô), verde (formação isolada de disco) e vermelho (dissipação isolada). Linhas verticais indicam transições entre estágios com as coloridas representando diferentes fases observacionais, enquanto linhas horizontais mostram o nível médio de brilho no estado de linha de base. Extraído, com permissão, de Figueiredo et al. (2025)

.

capazes de reconhecer padrões complexos em curvas de luz, é possível automatizar parte do trabalho que, tradicionalmente, dependeria de análise visual por especialistas. A ideia central é aproveitar exemplos rotulados (como os identificados por Figueiredo et al. 2025) para ensinar algoritmos a distinguir entre curvas de luz de estrelas Be e não Be.

A monografia está organizada da seguinte forma. O Capítulo 2 reúne os fundamentos teóricos, abordando tanto as propriedades das estrelas Be e sua variabilidade fotométrica quanto os conceitos de aprendizado de máquina essenciais para a compreensão deste trabalho. O Capítulo 3 descreve a metodologia empregada, incluindo a preparação dos dados, a definição de atributos e modelos utilizados. O Capítulo 4 apresenta e discute os resultados obtidos no problema de classificação das estrelas Be, nos dados rotulados de Figueiredo et al. (2025), incluindo uma discussão crítica sobre a metodologia e o conjunto de dados utilizados. Por fim, o Capítulo 5 apresenta as conclusões gerais e perspectivas para trabalhos futuros.

Capítulo 2

Fundamentos teóricos

2.1 Variabilidade fotométrica de estrelas Be

As propriedades observacionais das estrelas Be estão ligadas à presença e à evolução de um disco circunstelar gasoso, cuja formação e dissipação geram assinaturas características em suas curvas de luz. Essas variações fotométricas refletem uma gama de processos físicos e podem assumir diferentes morfologias conforme o ângulo de observação e o estágio de atividade do sistema. Em sistemas vistos de cima (*pole-on*), a presença do disco resulta em um aumento no brilho e em um leve avermelhamento. Já em sistemas observados de forma equatorial (*edge-on*), o disco pode obscurecer parcialmente a estrela, levando ao avermelhamento e à redução do brilho (Haubois et al. 2012, Rímulo et al. 2018).

Trabalhos clássicos, como Mennickent, R. E. et al. (2002) e Sabogal et al. (2005), propuseram classificações empíricas para organizar os diferentes padrões de variabilidade observados em estrelas Be nos dados dos levantamentos OGLE e ASAS. Nesses estudos, os autores mostraram que as curvas de luz dessas estrelas podem exibir os seguintes comportamentos: *outbursts*[1] abruptos de brilho seguidos de declínio suave, aumentos prolongados de brilho, variações quase periódicas ou flutuações irregulares em múltiplas escalas de tempo.

Embora úteis para descrever a diversidade de comportamentos fotométricos, essas classificações dependem fortemente da interpretação visual do avaliador, e não traduzem necessariamente os processos físicos subjacentes. Mais recentemente, trabalhos como o de Figueiredo et al. (2025) buscaram introduzir dimensões adicionais de informação, a partir da análise de variações simultâneas de brilho e cor nas curvas de luz.

[1] *Outburst*: episódio de aumento repentino de brilho, normalmente associado a eventos de ejeção de massa na estrela.

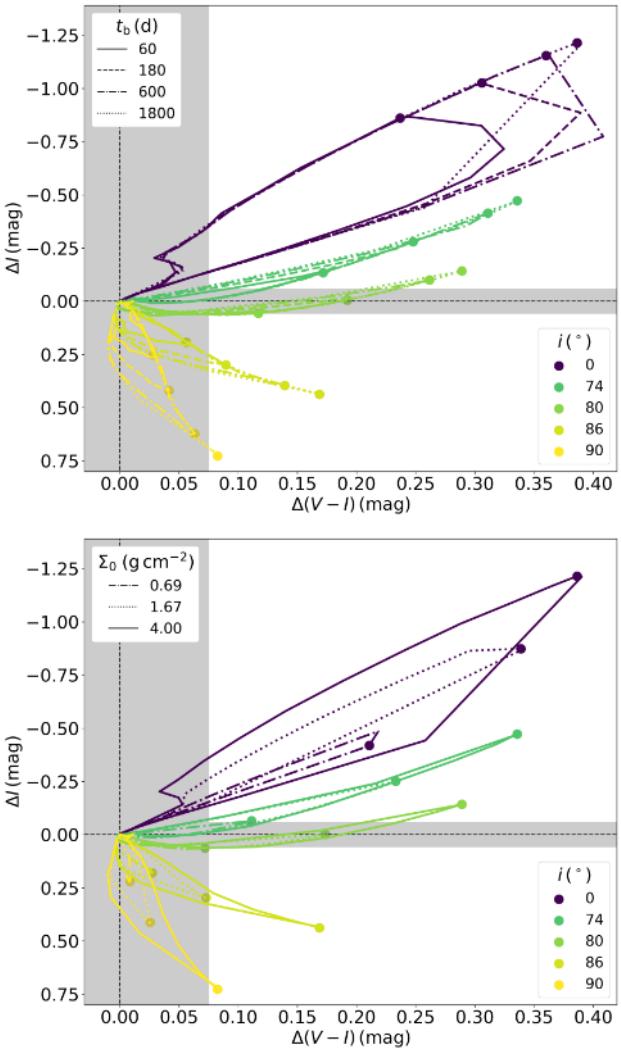

Figura 2.1: Simulações de sistemas Be no plano $\Delta(V-I)$ por ΔI para diferentes ângulos de inclinação e propriedades do disco. Extraído, com permissão, de Figueiredo et al. (2025). No painel superior, tem-se tempo de formação e ângulos de inclinação selecionados. No inferior, variam-se as densidades mantendo um tempo de formação fixo em $t_b = 1800$ dias, para os mesmos ângulos selecionados. Os círculos indicam a transição entre as fases de formação e dissipação do disco. Esses modelos foram utilizados como base para classificar as geometrias observacionais (ângulos de inclinação) das curvas reais.

No caso deste trabalho específico, o método adotado envolveu uma análise baseada em diagramas de cor-magnitude ($\Delta(V-I)$ por ΔI), construídos a partir dos dados fotométricos do OGLE. Esses diagramas observacionais foram comparados com diagramas simulados (vide Figura 2.1) que preveem o comportamento esperado para diferentes ângulos de inclinação e estágios de atividade do disco. Com base nessa comparação, os autores estabeleceram a orientação provável (*pole-on vs. edge-on*) de cada sistema e identificaram quais curvas de luz apresentavam variações compatíveis com os diferentes estágios de atividade de estrelas Be. A partir desses resultados, foi então criada uma amostra rotulada, na

qual as curvas de luz foram classificadas manualmente como candidatas a estrelas Be ou não-Be, que abriu a possibilidade, explorada neste trabalho, de aplicar métodos de aprendizado de máquina para reproduzir e ampliar o processo de identificação/classificação de estrelas Be.

Para realizar essa transição da inspeção visual para a classificação computacional automatizada, é necessário compreender os princípios que guiam o aprendizado dos algoritmos a partir dos dados. Por isso, a Seção 2.2 apresenta, de forma sucinta, os fundamentos teóricos do aprendizado de máquina necessários para compreender o resto do texto.

2.2 Fundamentos de aprendizado de máquina

2.2.1 Definição e tipos de aprendizado

Em termos gerais, o aprendizado de máquina é uma área da inteligência artificial voltada para o desenvolvimento de modelos computacionais capazes de reconhecer padrões e extrair informações relevantes diretamente a partir dos dados. Em vez de seguir instruções pré-definidas (Samuel 1995), esses modelos aprendem por meio de exemplos e ajustam seus parâmetros internos para realizar tarefas como classificação, regressão ou agrupamento.

Podemos distinguir dois paradigmas principais: o aprendizado supervisionado, em que o modelo é treinado com exemplos rotulados para aprender a relacionar entradas a saídas conhecidas; e o não supervisionado, no qual não há rótulos e o algoritmo busca agrupamentos naturais nos dados. Este trabalho concentra-se em modelos supervisionados, dada a existência de um conjunto rotulado de curvas de luz.

2.2.2 Tratamento dos dados e escolha de atributos

Independentemente da abordagem de aprendizado utilizada, é fato que o desempenho dos modelos é dependente da qualidade e da representação dos dados (para uma análise detalhada sobre como diferentes dimensões de qualidade afetam tarefas de aprendizado supervisionado e não supervisionado, ver Mohammed et al. 2025). Por isso, é importante que sempre haja uma etapa de preparação da amostra, que pode incluir desde o tratamento de valores ausentes, até a remoção de ruído, normalização dos dados, entre outros.

Outro aspecto importante é a definição das *features* (atributos), que são valores numéricos extraídos dos dados e fornecidos aos modelos para que estes façam suas tarefas. No caso de

curvas de luz, esses atributos podem representar propriedades como amplitude, periodicidade e brilho médio. A escolha desses valores é fundamental, já que influencia diretamente a performance dos modelos.

2.2.3 Modelos utilizados neste trabalho

Neste trabalho, foram empregados diferentes algoritmos de aprendizado de máquina supervisionado, sendo eles:

- *Random Forest (RF, Breiman 2001)*: algoritmo baseado em um conjunto de árvores de decisão, construídas de forma aleatória. Cada árvore contribui com um "voto", e a resposta final é dada pela maioria.
- *k-Nearest Neighbors (KNN, Cover e Hart 1967)*: classifica um novo dado com base nos rótulos dos k dados mais próximos (em termos de distância) no espaço de atributos.
- *Support Vector Machine (SVM, Cortes e Vapnik 1995)*: encontra uma superfície que melhor separa as classes no espaço de atributos. Pode usar funções *kernel*[2] para fazer separações mais complexas.
- *Multi-Layer Perceptron (MLP, David E. Rumelhart 1986)*: uma rede neural artificial com várias camadas.
- *eXtreme Gradient Boosting (XGBoost, Chen e Guestrin 2016)*: algoritmo também baseado em árvores, construindo-as de forma sequencial, com cada nova árvore corrigindo os erros da anterior (método de *boosting*).
- *Convolutional Neural Network (CNN, O'Shea e Nash 2015)*: classe de rede neural que processa diretamente imagens (ou dados em grade), em vez de medidas numéricas isoladas (como no caso dos outros modelos acima, chamados daqui em diante de modelos clássicos ou tradicionais). Pode ser construída com diferentes tipos e ordens de camadas, com cada uma aplicando operações diferentes nos dados.

[2] Kernel: função que transforma os dados para um espaço de dimensão maior onde a separação entre classes torna-se linear.

2.2.4 Métricas de avaliação

Uma vez treinado, o modelo precisa ser avaliado de forma objetiva. Para tarefas supervisionadas de classificação, podemos utilizar as seguintes métricas bem estabelecidas de classificação multiclasse (Powers 2020):

- *Acurácia*: mede a proporção de previsões corretas em relação ao total de exemplos.
- *Precisão*: indica a fração de exemplos classificados como positivos que realmente pertencem à classe. No contexto deste trabalho, entre todas as vezes que o modelo disse que uma estrela é Be, esta métrica mede quantas realmente são.
- *Revocação*: em inglês *recall*, mede a fração de exemplos positivos corretamente identificados. Isto é, entre todas as estrelas Be da base, quantas foram reconhecidas como tais pelo modelo. Essa métrica é relevante quando o objetivo é não deixar escapar objetos de interesse, como é o caso neste trabalho.
- *F1-score*: corresponde a uma média equilibrada entre precisão e revocação.

2.2.5 Redução de dimensionalidade

Em muitos casos, técnicas de redução de dimensionalidade, como a análise de componentes principais (PCA, do inglês *Principal Component Analysis*), são aplicadas para visualizar e interpretar o comportamento dos dados em espaços de menor dimensão, facilitando o entendimento dos resultados obtidos.

Nesse método, as novas "componentes principais" são combinações lineares das *features* originais, construídas para tentar capturar as direções onde há a maior variância possível dos dados. Como cada componente explica uma fração específica da variância total, isso permite identificar quais direções no espaço de atributos concentram mais informação relevante para o problema.

Capítulo 3

Metodologia

Neste capítulo, são detalhadas as etapas metodológicas empregadas para aplicar as técnicas de ML ao problema de classificação fotométrica de estrelas Be. Devido à natureza distinta dos dados de entrada utilizados pela CNN em comparação com os demais algoritmos, as seções subsequentes são organizadas de forma separada: uma dedicada aos modelos tradicionais, e outra à CNN.

3.1 Modelos tradicionais

3.1.1 Organização e tratamento dos dados

O ponto de partida deste trabalho foi um conjunto de cerca de 3000 curvas de luz provenientes do levantamento fotométrico OGLE, previamente tratadas e analisadas visualmente por Figueiredo et al. (2025). Diferentemente de um banco já classificado, essas curvas não possuíam rótulos diretos de classificação como estrelas Be ou não, mas possuíam as seguintes informações, identificadas visualmente por Figueiredo et al. (2025):

- "Metadados" correspondentes a comportamentos fotométricos específicos: com base nas características visuais da curva, Figueiredo et al. (2025) atribuiu múltiplas letras a cada curva, com cada letra correspondendo a um comportamento observado:
  i. "a" = presença de atividade identificada;
  ii. "r" = atividade recorrente identificada (periodicidade);
  iii. "m" = evento isolado identificado;
  iv. "b" = linha de base identificada;
  v. "n" = estrela descartada (não-Be) por análise de cor.

- Orientação do sistema: as curvas já estavam classificadas com orientação *pole-on*, *edge-on* ou indeterminado (*unclear*).

Com base nessas informações, foi possível separar o conjunto em estrelas rotuladas como candidatas a Be ou não candidatas, o que era necessário para tornar o problema tratável como uma tarefa de aprendizado supervisionado. As estrelas candidatas a Be foram definidas como aquelas que, concomitantemente: não foram descartadas por cor (isto é, não possuíam o metadado "n"), apresentavam orientação bem definida (*pole-on* ou *edge-on*, pois parte da análise de Figueiredo et al. (2025) dependia desta informação), e exibiam ao menos um metadado indicativo de atividade ("a", "r", "m" ou "b"). As demais curvas, incluindo aquelas com orientação indeterminada ou sem indícios de atividade clássica de estrelas Be, foram agrupadas na classe não Be.

Após esta etapa, foram obtidos então dois subconjuntos a partir da amostra original: um contendo as estrelas candidatas a Be (com rótulo igual a 1) e outro representando a classe não Be (com rótulo igual a 0), composta por objetos de variabilidade diversa.

3.1.2 Seleção e normalização de atributos (features)

Com os rótulos estabelecidos, o próximo passo consistiu em escolher, para representar cada curva de luz, um conjunto de atributos numéricos que fossem capazes de descrever sua variabilidade. Essa seleção foi baseada em intuições físicas relacionadas à variabilidade de estrelas Be, e também buscando maximizar o desempenho dos modelos de forma empírica.

Inicialmente, buscou-se empregar métricas já consolidadas na literatura de variabilidade estelar, seguindo outros trabalhos que já lidaram com o problema de classificação supervisionada em estrelas Be e variáveis no geral, como Pérez-Ortiz et al. (2017) e Debosscher et al. (2007). Destes trabalhos, foram aproveitadas as métricas:

- **Estatísticas descritivas básicas**: mediana e desvio absoluto mediano (MAD), para medidas da amplitude e dispersão da magnitude;
- **Correlação fotométrica**: índices Stetson J[3] e K[4], para medidas de similaridade entre observações sucessivas.

[3] Índice de Stetson J: mede a correlação temporal entre variações simultâneas em duas bandas fotométricas.

[4] Índice de Stetson K: descreve a distribuição (curtose) das variações de magnitude normalizadas.

- **Forma da distribuição**: *Octile Skewness* (OS) e *Left/Right Octile Weight* (LOW e ROW), para medidas de assimetria da curva de luz;

Além dessas *features* aproveitadas da literatura, foi desejado incorporar outros valores que trouxessem informações sobre a variabilidade em função do tempo. Por isso, primeiro, foi adaptado o conceito de variograma, bastante utilizado em geofísica e geoestatística, substituindo o domínio espacial para o domínio temporal das curvas de luz. Essa adaptação já foi feita em trabalhos como Eyer e Genton (1999), seguindo a estimativa de variograma feita no livro de geoestatística Matheron (1962).

Adicionalmente, foi incorporada uma análise no espaço de frequências por meio do periodograma de Lomb-Scargle (Lomb 1976, Scargle 1982), que é comumente empregado para detectar periodicidades em séries temporais irregulares. Entre os parâmetros derivados, foram considerados, como atributos para os modelos, a frequência correspondente ao maior pico de amplitude e o valor da amplitude nesse pico.

Para normalizar os atributos, foi utilizada a função `StandardScalar()` da biblioteca em *Python* `scikit-learn` (Pedregosa et al. 2011). Todos os atributos foram normalizados de forma a seguir uma distribuição normal padrão, com média igual a zero e desvio padrão igual a um. Isto foi feito pois garante um melhor funcionamento para os algoritmos baseados em escala/distância (como o KNN e SVM), além de otimizar e melhorar a convergência de algoritmos baseados em gradiente (como o XGBoost e MLP).

Uma PCA dos dados no espaço das *features* escolhidas foi feita para avaliar a capacidade de separação entre as classes (candidatas a Be e não-Be) neste espaço — isto é, para verificar se os atributos selecionados são eficazes para discriminar entre os dois grupos.

3.1.3 Treinamento dos modelos

Com as *features* extraídas e normalizadas, é possível aplicar, então, os modelos de aprendizado de máquina ao problema de classificação de estrelas Be. Nesta etapa, foram testados os cinco algoritmos clássicos de aprendizado supervisionado apresentados na Seção 2.2.3: RF, XGBoost, KNN, SVM e MLP, cada um representando um paradigma diferente de aprendizado supervisionado.

Os modelos foram treinados utilizando o conjunto rotulado descrito anteriormente, com uma divisão de 80% dos dados para treinamento e 20% para teste, de forma aleatorizada. Ademais, a otimização dos hiperparâmetros de cada modelo foi realizada utilizando o

método `GridSearchCV` novamente da biblioteca `scikit-learn`, que testa diferentes combinações de parâmetros e avalia o desempenho com validação cruzada.

3.1.4 Métricas e estratégias de avaliação

A avaliação do desempenho de cada modelo supervisionado foi conduzida com base em métricas de classificação binária, obtidas a partir da matriz de confusão[5] dos modelos. As métricas utilizadas foram acurácia, precisão, revocação e *F1-score*, descritas na Subseção 2.2.4.

Além da avaliação de desempenho para cada modelo, também foi analisada a importância relativa das *features* empregadas, estimada pelo método de importância por permutação. Nesse método, o valor de uma única *feature* é aleatoriamente embaralhado, sem alterar o restante dos dados. A redução consequente na acurácia do modelo é usada como medida da relevância do atributo em questão para a classificação dos modelos.

A metodologia descrita até aqui diz respeito aos modelos tradicionais, que operam sobre os atributos numéricos definidos. Além disso, também foi desenvolvida uma abordagem paralela baseada em redes neurais convolucionais (CNNs), que atuam diretamente nos gráficos (imagens) de magnitude por tempo das curvas de luz. As especificidades dessa segunda estratégia são abordadas na Seção 3.2.

3.2 Rede neural convolucional (CNN)

3.2.1 Preparação do conjunto de imagens

A primeira etapa correspondeu à geração de imagens, para servir de entrada ao modelo. Para isso, foram gerados gráficos de magnitude por tempo na banda I do OGLE, devido à sua maior cadência observacional. Todas as curvas foram representadas na mesma escala temporal, correspondente às fases observacionais do levantamento OGLE. A grande maioria das curvas incluídas no conjunto possui observações em todas essas fases, e por isso essa uniformização no eixo das abscissas não resultou em perda significativa de informação.

As imagens foram geradas em preto e branco, e sem eixos e rótulos, mas foram incluídas

[5] Matriz de confusão: visualização em tabela que avalia o desempenho de um modelo de classificação, mostrando a contagem de Verdadeiros Positivos (TP), Falsos Positivos (FP), Verdadeiros Negativos (TN) e Falsos Negativos (FN).

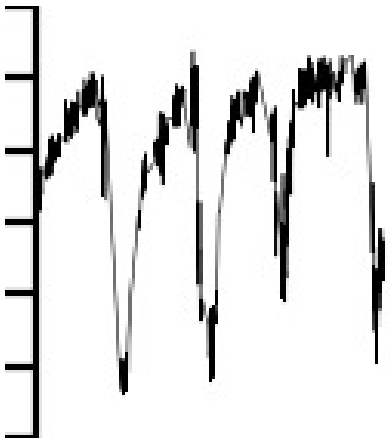

Figura 3.1: Exemplo de curva de luz processada para os modelos de CNN. A imagem foi gerada com resolução reduzida, em preto e branco, sem eixos ou rótulos, e com marcações horizontais a cada 0,1 mag no eixo y como referência visual da escala de variação.

pequenas marcações horizontais a cada 0.1 mag no eixo das ordenadas, para que o modelo tivesse uma referência visual da escala de variação de brilho. A escala de magnitude foi distinta para cada curva, correspondendo aos valores entre o máximo e mínimo de magnitude para cada estrela. Neste caso, o modelo foi guiado pelas variações de magnitude, e não pelos valores absolutos.

Foram testadas diferentes resoluções de imagem, variando o tamanho físico e a densidade de *pixels* (parâmetros *"figsize"* e *"dpi"* da biblioteca `matplotlib`, Hunter 2007). Observou-se empiricamente que o aumento da resolução elevava consideravelmente o tempo de treinamento, o que faz sentido, pois implica em um número maior de parâmetros a serem ajustados na camada linear/densa da rede neural. Por exemplo, imagens três vezes maiores resultam em aproximadamente nove vezes mais pesos a serem otimizados, o que acarreta em um tempo de execução e custo computacional proporcionalmente maiores. Por isso, optou-se pela menor resolução possível das imagens, mas que ainda proporcionasse um bom desempenho dos modelos. Um exemplo de imagem utilizada está disponível na Figura 3.1.

Para fornecer os dados como entrada aos modelos, as imagens finais foram convertidas em valores de cor (na escala de cinza, com 0 correspondendo a um *pixel* completamente branco, e 1 a um completamente preto) — isto é, cada imagem foi convertida em uma matriz na qual cada elemento correspondia ao valor de cor de um *pixel* — utilizando a função `imread` da biblioteca `scikit-image` (Pedregosa et al. 2011).

Cada imagem foi então associada ao seu respectivo rótulo: 0 para estrelas não Be e 1

para candidatas a Be, de forma similar ao que foi feito para os modelos tradicionais. Os dados foram separados aleatoriamente em 70% para treinamento, 20% para validação e 10% para teste.

3.2.2 Arquitetura da rede

A arquitetura da rede foi definida a partir de experimentação prática e inspirada em trabalhos recentes que também aplicaram redes convolucionais a curvas de luz do OGLE, como Monsalves et al. (2024).

A rede proposta consiste em duas camadas convolucionais[6] (com respectivamente 8 e 16 filtros e tamanho de *kernel* igual a 3), alternadas por uma camada de *max-pooling*[7] (com janela 2x2 e passo 2), e seguida por uma camada totalmente conectada[8] (linear). Cada convolução foi seguida por uma função de ativação ReLU[9], para introduzir não linearidade na rede, e a saída final é obtida por uma função sigmoide[10], para garantir que a previsão final do modelo seja um valor entre 0 e 1. Os valores obtidos são então arredondados para 0 ou 1, correspondendo à classificação final da rede.

3.2.3 Treinamento do modelo, ajuste de hiperparâmetros e avaliação

O treinamento foi realizado com *batch size*[11] igual a 50 e diferentes valores de taxa de aprendizado (*learning rate*)[12], explorados em uma grade de valores. Para cada combinação, o modelo foi treinado até atingir um critério de parada por detecção de platô: o treinamento era interrompido caso a função de perda não apresentasse melhora no conjunto de validação após três épocas[13] consecutivas. O número máximo de épocas foi fixado em 1000, mas todos os modelos convergiram antes desse limite. Com isso, a taxa de aprendizado e o número de épocas para o modelo foram ajustados empiricamente, de forma a buscar a combinação

[6] Camada convolucional: usando operações matriciais, aplica filtros (*kernels*) sobre diferentes regiões da imagem para extrair características locais.

[7] Camada de *pooling*: camada que reduz a dimensionalidade espacial da representação.

[8] Camada totalmente conectada: camada que gera a classificação final.

[9] Função de ativação ReLU: retorna o valor de entrada se for positivo, e zero caso contrário.

[10] Função de ativação sigmoide: transforma qualquer valor em um número entre 0 e 1. Valores próximos de 0 ou 1 indicam maior confiança na classificação.

[11] *Batch size*: tamanho do subconjunto de dados usado em cada iteração de treinamento.

[12] Taxa de aprendizado (*learning rate*): tamanho do passo a cada iteração de treinamento.

[13] Época: corresponde a uma iteração do modelo pelo conjunto de treino.

que gerasse melhor desempenho.

O desempenho da CNN foi avaliado utilizando as mesmas métricas descritas na Seção 3: acurácia, precisão, revocação e *F1-score*.

3.2.4 Extensão da CNN para classificação de orientação das estrelas

Com o desenvolvimento do trabalho, buscou-se também investigar se uma rede neural convolucional seria capaz de estimar não apenas a classificação entre Be e não Be, mas também a orientação das estrelas Be (Figueiredo et al. 2025) em dois regimes distintos: *pole-on* (que cobre um grande intervalo de ângulos entre aproximadamente 0° e 60°) e *edge-on* (que cobre ângulos entre 80° e 90°). Essa informação é bastante relevante, já que a inclinação do sistema em relação à linha de visada altera consideravelmente a aparência fotométrica das curvas. Esse tipo de análise também abre espaço para estudos posteriores, como estimativas de outras características das estrelas, juntamente à classificação binária.

Para testar essa possibilidade, foi desenvolvida uma segunda rede convolucional, utilizando as mesmas imagens de entrada descritas anteriormente, mas com uma nova configuração de rótulos: 0 para estrela não Be, 1 para estrela Be com orientação *pole-on*, e 2 para estrela Be com orientação *edge-on*.

Essa abordagem conjunta foi escolhida por dois motivos principais. Primeiro, as classificações de orientação não são independentes da classificação Be/não-Be: o procedimento de Figueiredo et al. (2025), que permitiu a determinação visual da orientação, só pode ser aplicado às estrelas já identificadas como Be. Segundo, a subrepresentação de estrelas *edge-on* no conjunto original (correspondendo a apenas cerca de 10% dos dados) tornaria inviável treinar uma rede dedicada exclusivamente à estimativa de orientação, pois o modelo não teria exemplos suficientes para aprender adequadamente essa característica.

O processo de treinamento, avaliação e a arquitetura base permaneceram os mesmos, mas foram feitas modificações na camada de saída e na função de ativação final. A função sigmoide, utilizada na versão binária, foi substituída por uma função *softmax*[14]. Com essa mudança, a camada de saída, que na rede anterior produzia um escalar (correspondente à probabilidade de ser Be), passou a gerar um vetor no qual cada elemento corresponde à probabilidade da curva de luz pertencer a uma das classes definidas anteriormente.

[14] Função de ativação *softmax*: transforma a saída da camada linear em um vetor de probabilidades normalizadas entre 0 e 1, cuja soma é igual a 1. Usada para problemas multiclasse.

Capítulo 4

Resultados e Discussão

O desenvolvimento da metodologia descrita no Capítulo 3 foi, por si só, um dos principais resultados do trabalho. O *pipeline* completo usado (desde a rotulagem, seleção de atributos, otimização dos modelos e desenvolvimento de ambas as CNNs) é uma estrutura metodológica que pode ser aprimorada e reutilizada em futuras aplicações, particularmente em outros levantamentos (salvo ajustes necessários), como o LSST. No presente capítulo, são apresentados e discutidos os resultados obtidos a partir de sua aplicação.

4.1 Modelos tradicionais

Os modelos tradicionais de aprendizado supervisionado apresentaram desempenhos consistentes na tarefa de classificação entre estrelas candidatas a Be e não Be com dados do OGLE. As métricas médias obtidas para os diferentes algoritmos testados (RF, XGBoost, KNN, SVM e MLP) estão resumidas na Tabela 4.1.

Tabela 4.1 - Desempenho dos modelos tradicionais de aprendizado supervisionado.

Algoritmo	**Acurácia**	**Precisão**	**Revocação**	***F1-score***
RF	0.86	0.85	0.93	0.89
XGBoost	0.85	0.86	0.92	0.89
KNN	0.84	0.85	0.90	0.87
SVM	0.81	0.84	0.85	0.85
MLP	0.85	0.85	0.92	0.88

Os resultados indicam um desempenho geral bom entre os modelos testados, com valores de *F1-score* entre 0.85 e 0.89. Ademais, os classificadores baseados em conjuntos de árvores, mais especificamente RF e XGBoost apresentaram os melhores resultados em

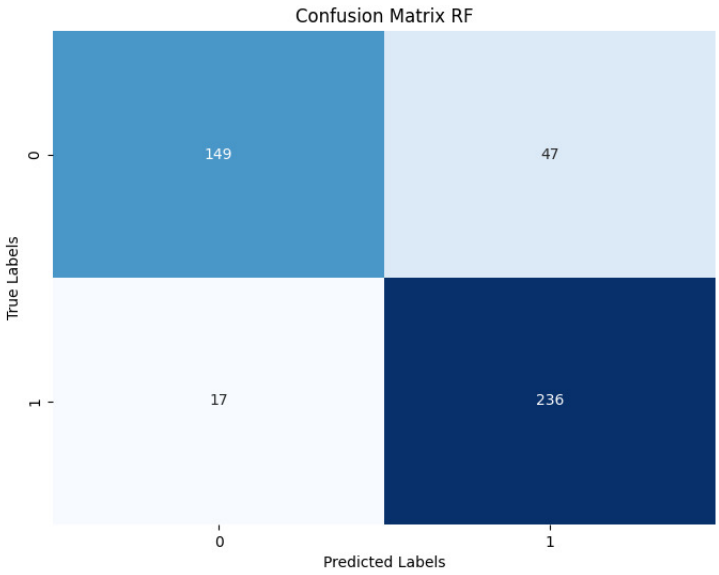

Figura 4.1: Matriz de confusão do modelo *Random Forest* (RF) para a classificação binária. As classes 0 e 1 representam as categorias analisadas, com 0 sendo não Be e 1 sendo Be. Valores na diagonal principal (149 e 236) correspondem às previsões corretas (verdadeiros negativos e verdadeiros positivos, respectivamente); já os valores na diagonal secundária (47 e 17) representam os erros de classificação (falsos positivos e falsos negativos, respectivamente).

média, com os maiores valores de revocação e *F1-score*.

Os valores elevados de revocação (≥ 0.85) mostram que os modelos são capazes de identificar eficientemente as estrelas Be presentes em nosso conjunto de dados. Em particular, a matriz de confusão do modelo RF (que obteve o melhor desempenho geral), disponível na Figura 4.1, contém 236 verdadeiros positivos (TP) e apenas 17 falsos negativos (FN), o que mostra que o modelo recupera a maior parte das estrelas candidatas a Be. Isso é muito bom, pois, quando se lida com objetos raros como no caso deste trabalho, é importante minimizar a quantidade de objetos descartados (FN), mesmo que seja necessário fazer uma análise posterior para descartar objetos adicionais que foram classificados como Be erroneamente. Neste sentido, foram obtidos 149 verdadeiros negativos (TN) e 47 falsos positivos (FP), mostrando que o algoritmo tende a classificar algumas estrelas não Be como candidatas a Be, mas mantém uma taxa de erro aceitável.

Uma vez que o desempenho de todos os modelos foi muito similar, acredita-se que a etapa de escolha de atributos seja mais importante do que a escolha dos algoritmos em si para este problema específico de classificação. A importância relativa das *features* foi

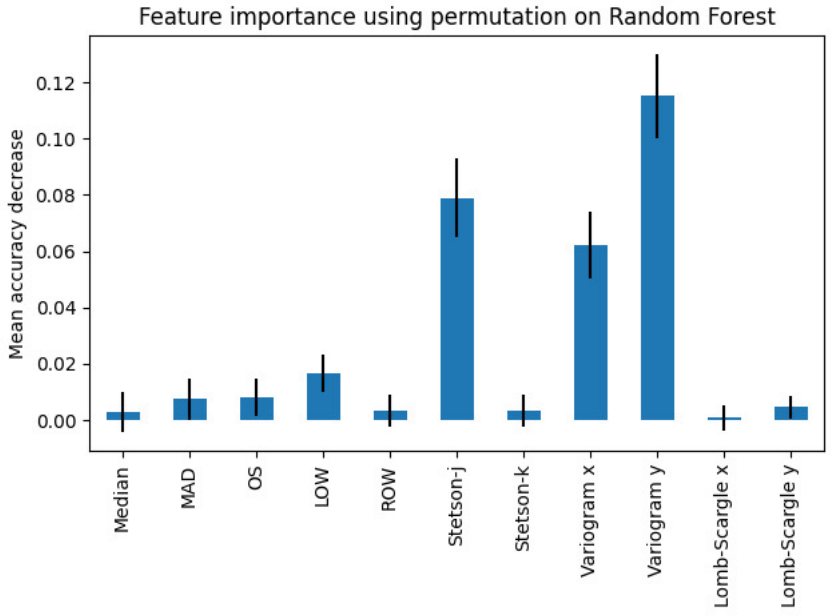

Figura 4.2: Importância das *features* calculada por permutação para o modelo *Random Forest* (RF). O eixo vertical representa a diminuição média na acurácia do modelo quando cada atributo é embaralhado, e as barras pretas correspondem ao erro.

avaliada por meio do método de permutação no modelo *Random Forest*. Os resultados (Figura 4.2) mostraram que as métricas com maior impacto na acurácia do modelo foram, em ordem decrescente de importância, o variograma em y (na escala de magnitude), seguido pelo índice de Stetson J e o variograma em x (na escala temporal). Além disso, também foi realizada uma análise exploratória por PCA, utilizando as *features* normalizadas. Os resultados indicaram que as três primeiras componentes principais concentraram cerca de 60% da variância total dos dados, com maior contribuição dos variogramas e do índice Stetson J nas componentes, o que reforça suas importâncias já observadas na análise de permutação.

Esses resultados evidenciam aspectos relevantes, ainda mais quando se considera o procedimento de classificação feito por humanos. As métricas de variograma em y e x e o índice de Stetson J são as únicas dentre as utilizadas que possuem informação temporal explícita sobre a curva de luz: o variograma em y e x medem, respectivamente, quantidades relacionadas à maior variação em magnitude e sua escala de tempo, e o índice Stetson J tem relação com a correlação entre observações sucessivas (no tempo). Como humanos especializados na tarefa de identificar estrelas Be a partir de suas curvas de luz também utilizam a dinâmica temporal das variações em suas análises, é natural que as *features* que

incorporem essa informação sejam muito importantes para os classificadores automáticos.

Vale comentar que, enquanto o índice Stetson J capta a correlação temporal entre medidas consecutivas, o Stetson K descreve apenas a forma da distribuição das magnitudes (sem informação temporal). Sabendo disso, também é coerente que o Stetson J tenha se mostrado muito mais relevante que o Stetson K para os modelos.

Surpreendentemente, os atributos derivados a partir do periodograma de Lomb-Scargle não foram muito úteis para a classificação dos modelos. Isso pode, no entanto, simplesmente significar que os atributos escolhidos não foram as melhores opções no espaço de frequências, e não necessariamente que informações de frequência não são discriminantes. Novos atributos, baseados em técnicas mais refinadas como a análise de *wavelets*[15], podem ser testados no futuro.

As demais métricas (mediana, MAD, OS, LOW e ROW) apresentaram contribuições menores, o que também era esperado, por serem medidas puramente estatísticas e mais genéricas. Vale destacar que a LOW apresentou uma importância relativamente maior que as demais, e ela mede variações assimétricas nas regiões mais brilhantes das curvas. Isso pode significar, por exemplo, que essa métrica consiga capturar indiretamente a presença de eventos de ejeção de massa, que levam à formação dos discos em estrelas Be.

Em síntese, os resultados obtidos indicam que a chave para uma classificação eficaz de estrelas Be reside principalmente na definição de *features* fisicamente significativas. Direções futuras promissoras incluem a introdução de ainda mais atributos que capturem a dinâmica temporal, e de mais atributos no espaço de frequência, que foi pouco investigado.

4.2 Redes neurais convolucionais (CNNs)

4.2.1 CNN de classificação binária

Na tarefa binária, o objetivo era diferenciar estrelas candidatas a Be de não Be com base nas imagens das curvas de luz. Para analisar o comportamento do treinamento, as curvas da função de perda (*loss*) ao longo das épocas foram comparadas para diferentes taxas de aprendizado. Com isso, notou-se que (como pode ser visto na Figura 4.3), quando a taxa de aprendizado usada era muito alta, a curva de perda exibiu um comportamento irregular, oscilando muito e sem convergir. Isso aconteceu pois os ajustes dos pesos no modelo

[15] Análise de *wavelets*: fornece uma representação tempo-frequência dos sinais.

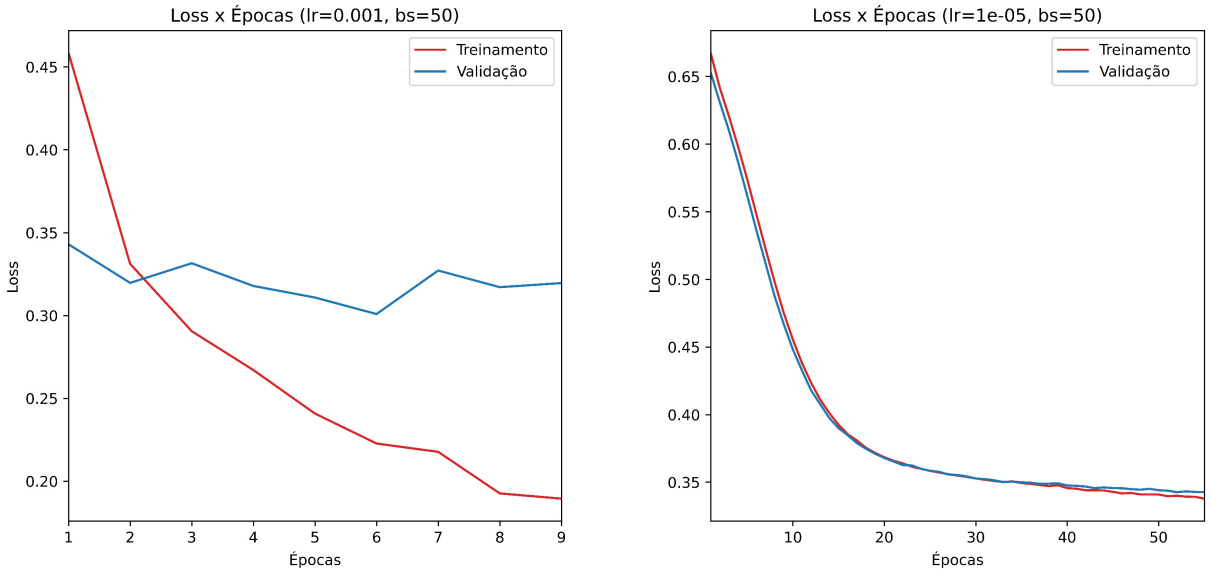

Figura 4.3: Curvas de perda por época para diferentes taxas de aprendizado para a CNN binária. À esquerda, o modelo com taxa de aprendizado = 0.001 apresentou comportamento muito errático e dificuldade de convergência na função de perda, especialmente na etapa de validação. À direita, com taxa de aprendizado = 0.00001, a curva é notavelmente mais suave e converge melhor. A diferença de épocas se deve pelo critério de convergência usado.

estavam sendo realizados em passos muito grandes, de forma que, a cada nova iteração de treinamento, o modelo sobrecorrigia os erros da iteração anterior, e isso impedia a rede de se estabilizar em um mínimo da função de custo. Em contrapartida, taxas de aprendizado mais baixas levaram a curvas de perda mais suaves e convergentes, que se estabilizaram em valores baixos de função de perda após algumas dezenas de épocas.

As métricas obtidas para o modelo binário final estão disponíveis na Tabela 4.2.

Tabela 4.2 - Desempenho da rede neural convolucional binária.

Algoritmo	**Acurácia**	**Precisão**	**Revocação**	***F1-score***
CNN binária	0.88	0.87	0.94	0.90

Os valores obtidos para a CNN binária são comparáveis, e até superiores, aos resultados obtidos com os modelos tradicionais, indicando que a CNN consegue bom desempenho mesmo sem a necessidade de criar *features* numéricas que descrevam as curvas de luz. Por outro lado, esse modelo demanda maior tempo de treinamento e recursos computacionais mais intensivos, o que deve ser levado em conta na escolha do algoritmo.

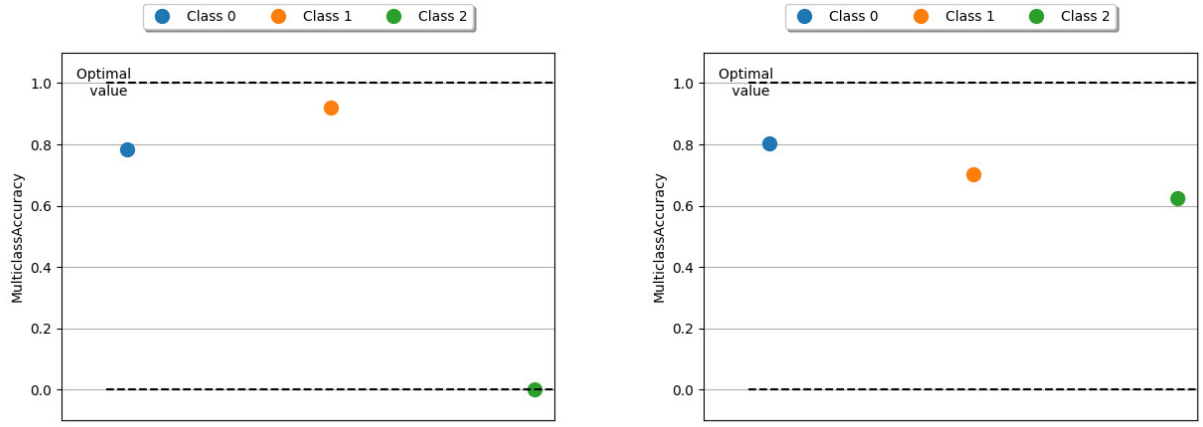

Figura 4.4: Acurácia por classe do modelo CNN multiclasse (classe 0: não-Be, classe 1: Be *pole-on*, classe 2: Be *edge-on*). À esquerda, o modelo sem pesos na função de perda gerou acurácia zero para classe 2 (Be *edge-on*) devido ao desbalanceamento do conjunto de dados. Já à direita, o modelo com pesos fez com que a acurácia da classe 2 aumentasse para aproximadamente 60%, com perda parcial de acurácia na classe 1.

4.2.2 CNN de classificação multiclasse

Após o bom desempenho no problema binário, a CNN foi adaptada para uma tarefa mais complexa: distinguir entre três classes — não Be (0), Be *pole-on* (1) e Be *edge-on* (2). Este foi um estudo exploratório para checar se o modelo seria capaz de inferir, além da classificação em Be ou não Be, a orientação das estrelas, já que esta muda a morfologia das curvas de luz.

Inicialmente, o modelo utilizava a função de perda *Cross Entropy*[16] padrão, como foi feito na outra rede, para avaliar sua predição. No entanto, devido ao forte desbalanceamento entre as classes (no conjunto de dados disponível, 1639 objetos eram classificados como *pole-on*, e apenas 112 como *edge-on*), a rede aprendeu a favorecer a classe mais numerosa, classificando todas as curvas Be, sem exceção, como *pole-on*. Nesse estágio, a acurácia da classe *pole-on* chegava a cerca de 90%, mas a classe *edge-on* obtinha acurácia nula (0%), como pode ser visto no painel esquerdo da Figura 4.4.

Para contornar esse problema, foram introduzidos pesos na função de perda, penalizando mais fortemente erros na classe com menos membros, a de estrelas Be com orientação *edge-on*. Essa modificação reduziu o desempenho do modelo na classe *pole-on*, mas aumentou consideravelmente o reconhecimento da classe *edge-on*, cuja acurácia passou de

[16] Cross Entropy: Função de perda que mede a dissimilaridade entre distribuições de probabilidade. Pode ser usada tanto em casos binários quanto multiclasse.

0% para acima de 60%. Como resultado, a acurácia média global ficou em torno de 71%, como pode ser visto no painel direito da Figura 4.4, e na Tabela 4.3 relativa às métricas de avaliação deste modelo.

Tabela 4.3 - Desempenho da rede neural convolucional multiclasse (que também estima a orientação).

Algoritmo	**Acurácia**	**Precisão**	**Revocação**	***F1-score***
CNN multiclasse	0.71	0.63	0.71	0.63

Embora os valores para este modelo seja inferiores aos obtidos nos demais modelos, esse resultado ainda mostra, de certa forma, que há informação suficiente nas curvas para distinguir diferentes orientações, mesmo que de forma limitada pelo tamanho pequeno e falta de equilíbrio do conjunto de dados. Para se obter melhores resultados, provavelmente será necessário utilizar um conjunto de dados maior e mais balanceado.

Assim, é observado que, mesmo com uma arquitetura simples e um conjunto de dados "não ideal", ambas redes neurais conseguiram identificar padrões morfológicos nas curvas de luz de estrelas Be, além de resultados medianos no problema de determinação da orientação do sistema. O desempenho obtido pode ser interpretado como um estudo exploratório promissor, e estudos futuros em conjuntos maiores, com distribuição mais equilibrada de classes e que eventualmente usem arquiteturas mais profundas ou otimizadas para o problema certamente trarão resultados ainda mais positivos.

4.3 Discussão sobre o conjunto de dados e limitações metodológicas

Os resultados obtidos devem ser interpretados levando em conta as características do conjunto de dados utilizado. A amostra original, derivada de Figueiredo et al. (2025), contém aproximadamente 3000 curvas de luz, das quais uma grande fração foi rotulada como candidata a estrela Be. Essa composição significa que o conjunto é fortemente super-representado em estrelas Be, quando comparado à frequência real desse tipo de estrela na natureza. Isso ocorre, em partes, por um viés de seleção: boa parte das estrelas selecionadas por Figueiredo et al. (2025) apresentava comportamento fotométrico compatível com estrelas desse tipo, justamente porque os autores estavam buscando candidatas a Be.

Esse viés de seleção é importante de se considerar, porque faz com que o conjunto de "não-Be" provavelmente não represente toda a diversidade de variáveis presentes em

catálogos reais. Em outras palavras, o modelo pode estar aprendendo a diferenciar estrelas Be de estrelas "não-Be, mas parecidas com Be" (ou talvez outra diferenciação mais simples do que o problema real), e não de um conjunto de fato heterogêneo de estrelas variáveis. Neste caso, o modelo atual provavelmente teria um desempenho inferior se aplicado a outro conjunto independente, com expressões de variabilidade mais diversas e complexas.

Por isso, para estudos futuros, seria importante analisar com mais cuidado as curvas não Be do conjunto atual: como é possível que elas incluam apenas alguns tipos de variáveis, ou muitas não variáveis, o ideal seria ampliar essa parte da amostra, incorporando mais tipos de variabilidade, tornando a classificação mais realista e os modelos mais generalizáveis.

Outra limitação está relacionada à natureza dos rótulos, que são definidos por inspeção visual e por isso sujeitos a vieses. Como algumas curvas podem ter interpretações ambíguas mesmo entre especialistas, é possível que o desempenho dos modelos esteja parcialmente limitado pela qualidade e subjetividade dos rótulos, e por isso nenhum modelo alcance acurácia na casa dos 90%. Para aplicações futuras, seria interessante aprimorar o modelo com classificações visuais feitas por diferentes avaliadores, para minimizar possíveis erros. Alternativamente, uma abordagem independente seria partir de uma amostra bem conhecida de estrelas Be (por exemplo, identificadas por espectroscopia), combinada de forma balanceada a outras classes de variáveis da literatura, minimizando os vieses mencionados. Uma fonte potencial é o catálogo de Be de Tan et al. (2025), que aplicaram redes neurais profundas a espectros do *Data Release* 11 do *Large Sky Area Multi-Object Fiber Spectroscopic Telescope* (LAMOST) para identificar milhares de candidatas a Be.

Outro ponto importante a se considerar é que, conforme mostrado na análise de importância por permutação, as *features* mais relevantes para a classificação foram justamente aquelas que carregam informação temporal. Isso reforça a importância da variabilidade no tempo para o desempenho dos modelos, mas também sugere que eles podem estar sensíveis à forma como o levantamento OGLE amostra o tempo — à sua cadência e distribuição de observações. Se o modelo estiver muito sensível a essas características, ele pode ter dificuldade para generalizar para outros levantamentos, como o LSST. Isso deve testado futuramente, aplicando o modelo a conjuntos independentes e de levantamentos distintos.

Por fim, como todas as avaliações foram feitas por validação cruzada dentro do próprio conjunto de dados original, um passo importante, em trabalhos futuros, é realizar validações externas, em bases de outros levantamentos ou em conjuntos independentes deste.

Capítulo 5

Conclusões

O presente trabalho apresentou o desenvolvimento e a aplicação de diferentes abordagens de aprendizado supervisionado para a identificação fotométrica de estrelas Be, utilizando curvas de luz do levantamento OGLE, classificadas manualmente entre candidatas a Be ou não por Figueiredo et al. (2025). Para além da comparação de modelos, o principal foco do trabalho foi na construção e validação de uma metodologia robusta, que pudesse ser aplicada futuramente em levantamentos fotométricos massivos, como o LSST.

Os modelos tradicionais (RF, XGBoost, KNN, SVM e MLP) apresentaram desempenhos semelhantes, com valores de *F1-score* entre 0.85 e 0.89. O melhor modelo tradicional obteve acurácia de 86%. Ademais, a pequena variação entre algoritmos indicou que o sucesso dos modelos no problema de classificação depende mais da qualidade e representatividade dos atributos numéricos usados, do que da escolha específica do classificador. As *features* mais relevantes, segundo análise de importância por permutação, foram os variogramas (que trazem informação sobre a maior variação em magnitude e sua escala de tempo) e o índice de Stetson J, que também traz informações sobre a dinâmica temporal das curvas. Esse resultado reforça que a variabilidade ao longo do tempo é um fator discriminante muito importante entre estrelas Be e não Be.

A rede neural convolucional (CNN), por sua vez, demonstrou que é possível que um modelo aprenda a diferenciar as estrelas diretamente das curvas de luz transformadas em imagens, ainda alcançando métricas comparáveis, e neste caso até levemente superiores, às dos modelos baseados em atributos. O modelo binário final, com hiperparâmetros ajustados por análise da curva de perda, obteve acurácia de 88%. Já na classificação multiclasse (que tinha objetivo de distinguir não só se uma curva de luz era Be ou não Be, mas também trazer informações sobre a orientação das estrelas Be), a inclusão de pesos na

função de perda permitiu à CNN distinguir parcialmente as orientações *pole-on* e *edge-on*, atingindo acurácia média de 71%. Esse desempenho, embora inferior ao da tarefa binária, já era esperado pelo desbalanceamento do conjunto de dados e pela maior complexidade do problema. Apesar disso, já representa um passo inicial em direção a modelos capazes de não só classificar as estrelas, mas possivelmente também estimar propriedades físicas.

De forma geral, o trabalho mostrou que, para modelos baseados em valores numéricos, a incorporação de informações temporais é crucial para a identificação de estrelas Be. Já para modelos baseados em imagens, a abordagem usando CNN confirmou que a morfologia visual das curvas de luz por si só contém padrões discriminativos suficientes para uma boa classificação, sem necessitar de uma etapa de seleção de *features*.

Como perspectivas futuras, pretende-se aplicar a metodologia desenvolvida a conjuntos de dados maiores e mais balanceados, incluindo possivelmente levantamentos como ASAS e KELT, mais dados do levantamento OGLE, e especialmente o *Legacy Survey of Space and Time* (LSST), do Observatório Vera C. Rubin. O grupo de pesquisa ao qual este trabalho está vinculado possui direitos de acesso aos dados do *Rubin Observatory*, o que possibilitará a utilização dos conjuntos de dados do LSST em fases futuras do projeto, quando houver cobertura temporal suficiente.

No âmbito da pós-graduação, pretende-se dar continuidade a essa linha de pesquisa, desenvolvendo novos projetos que apliquem técnicas de aprendizado de máquina e inteligência artificial a diferentes aspectos do estudo das estrelas Be, tanto em modelos de classificação (como este), quanto possivelmente em modelos de ML de regressão, para estimativa de parâmetros físicos dessas estrelas. Entre os objetivos principais futuros, tem-se a criação de um modelo automatizado de identificação de estrelas Be em dados do LSST, capaz de operar sobre o grande volume de dados produzidos por esse levantamento, como extensão do trabalho feito aqui.

Assim, este estudo marca o início de um programa de pesquisa de longo prazo, a ser desenvolvido ao longo dos próximos anos na pós-graduação. A metodologia estabelecida oferece uma base para o desenvolvimento de novas aplicações, especialmente para trabalhos de identificação de novas estrelas Be em levantamentos de larga escala.

"Embora a máquina aprenda, quem ensina
e interpreta continua sendo o humano."

Referências Bibliográficas

Breiman L., Random Forests, Machine Learning, 2001, vol. 45, p. 5

Chen T., Guestrin C., XGBoost: A Scalable Tree Boosting System. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining , KDD '16, Association for Computing Machinery, New York, NY, USA, 2016, p. 785–794

Cortes C., Vapnik V., Support-Vector Networks, Mach. Learn., 1995, vol. 20, p. 273–297

Cover T., Hart P., Nearest neighbor pattern classification, IEEE Transactions on Information Theory, 1967, vol. 13, p. 21

David E. Rumelhart Geoffrey E. Hinton R. J. W., Learning representations by back-propagating errors, EBSCOhost Academic Search Premier, 1986, vol. 323, p. 533

Debosscher J., Sarro L. M., Aerts C., Cuypers J., Vandenbussche B., Garrido R., Solano E., Automated supervised classification of variable stars: I. Methodology, Astronomy amp; Astrophysics, 2007, vol. 475, p. 1159–1183

Eyer L., Genton M. G., Characterization of variable stars by robust wave variograms: an application to HIPPARCOS mission, Astronomy and Astrophysics Supplement Series, 1999, vol. 136, p. 421

Figueiredo A. L., Carciofi A. C., Labadie-Bartz J., Pinho M. L., de Amorim T. H., dos Santos P. T., Soszynski I., Udalski A., , 2025 Be star demographics: a comprehensive study of thousands of lightcurves in the Magellanic Clouds

Haubois X., Carciofi A. C., Rivinius T., Okazaki A. T., Bjorkman J. E., Dynamical Evolution of Viscous Disks around Be Stars. I. Photometry, ApJ, 2012, vol. 756, p. 156

Hunter J. D., Matplotlib: A 2D graphics environment, Computing in Science & Engineering, 2007, vol. 9, p. 90

Ivezic Z., Kahn S. M., Tyson J. A., Abel B., Acosta E., et al. LSST: From Science Drivers to Reference Design and Anticipated Data Products, The Astrophysical Journal, 2019, vol. 873, p. 111

Lomb N. R., Least-Squares Frequency Analysis of Unequally Spaced Data, Ap&SS, 1976, vol. 39, p. 447

Matheron G., Traité de Géostatistique Appliquée. Tome I. No. 14 in Mémoires du BRGM, Technip Paris, 1962

Mennickent, R. E. Pietrzynski, G. Gieren, W. Szewczyk, O. On Be star candidates and possible blue pre-main sequence objects in the Small Magellanic Cloud***, A&A, 2002, vol. 393, p. 887

Mohammed S., Budach L., Feuerpfeil M., Ihde N., Nathansen A., Noack N., Patzlaff H., Naumann F., Harmouch H., The effects of data quality on machine learning performance on tabular data, Information Systems, 2025, vol. 132, p. 102549

Monsalves Jaque Arancibia, M. Bayo, A. Sánchez-Sáez, P. Angeloni, R. Damke, G. Segura Van de Perre, J. Application of Convolutional Neural Networks to time domain astrophysics. 2D image analysis of OGLE light curves, AA, 2024, vol. 691, p. A106

O'Shea K., Nash R., , 2015 An Introduction to Convolutional Neural Networks

Pedregosa F., Varoquaux G., Gramfort A., Michel V., Thirion B., Grisel O., Blondel M., Prettenhofer P., Weiss R., Dubourg V., Vanderplas J., Passos A., Cournapeau D., Brucher M., Perrot M., Duchesnay E., Scikit-learn: Machine Learning in Python, Journal of Machine Learning Research, 2011, vol. 12, p. 2825

Pepper J., Gould A., Depoy D. L., KELT: The Kilodegree Extremely Little Telescope. In The Search for Other Worlds: Fourteenth Astrophysics Conference , vol. 713, 2004, p. 185

Pojmanski G., The All Sky Automated Survey, Acta Astronomica, 1997, vol. 47, p. 467

Powers D. M. W., Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation, CoRR, 2020, vol. abs/2010.16061

Pérez-Ortiz M. F., García-Varela A., Quiroz A. J., Sabogal B. E., Hernández J., Machine learning techniques to select Be star candidates: An application in the OGLE-IVGaiasouth ecliptic pole field, Astronomy amp; Astrophysics, 2017, vol. 605, p. A123

Rímulo L. R., Carciofi A. C., Vieira R. G., Rivinius T., Faes D. M., Figueiredo A. L., Bjorkman J. E., Georgy C., Ghoreyshi M. R., Soszyński I., The life cycles of Be viscous decretion discs: fundamental disc parameters of 54 SMC Be stars, MNRAS, 2018, vol. 476, p. 3555

Sabogal B. E., Mennickent R. E., Pietrzynski G., Gieren W., Be star candidates in the Large Magellanic Cloud: the catalogue and comparison with the Small Magellanic Cloud sample, Monthly Notices of the Royal Astronomical Society, 2005, vol. 361, p. 1055

Samuel A. L., Some Studies in Machine Learning Using the Game of Checkers, IBM J. Res. Dev., 1995, vol. 44, p. 206

Scargle J. D., Studies in astronomical time series analysis. II. Statistical aspects of spectral analysis of unevenly spaced data., ApJ, 1982, vol. 263, p. 835

Tan L., Deng H., Mei Y., chi H., Chen Y., Liu T., Wang F., , 2025 A robust method for identifying Be stars in the LAMOST Data Release 11 based on Deep-learning approach

Udalski A., Szymanski M., Kaluzny J., Kubiak M., Mateo M., The Optical Gravitational Lensing Experiment, Acta Astronomica, 1992, vol. 42, p. 253