# **PROJETO — ExoSeeker.AI**

<https://drive.google.com/drive/folders/1n5z2wR9LKpbllRgZ8XG7KUCSnVSi5iXp?usp=sharing>

*Detecção robusta e explicável de exoplanetas com IA, unificando Kepler, K2 e TESS — com interface web para cientistas e público.*

## **1) Visão & Impacto**

**Problema.** Os catálogos de exoplanetas cresceram exponencialmente, mas a triagem ainda consome tempo humano e sofre com vieses e dados heterogêneos entre missões (Kepler, K2 e TESS). Grande parte do “ouro” continua escondido em curvas de luz e metadados já públicos. [exoplanetarchive.ipac.caltech.edu](https://exoplanetarchive.ipac.caltech.edu/docs/counts_detail.html?utm_source=chatgpt.com)

**Nossa proposta.** O **ExoSeeker.AI** é uma pipeline ponta-a-ponta que:

1. consolida dados rotulados das três missões (KOI/Kepler, K2 Planets & Candidates, TOI/TESS);
2. treina um **classificador multimodal** (curva de luz + atributos tabulares) com **adaptação entre domínios**;
3. entrega **explicabilidade** (Grad-CAM/SHAP sobre segmentos de trânsito e features físicas);
4. publica uma **interface web** para upload de dados, inferência, ajuste de hiperparâmetros e rotulagem ativa para humanos.

**Impacto científico.** Acelera a descoberta, reduz falsos positivos inter-missão e gera hipóteses reprodutíveis (artefatos vs. trânsitos reais), incorporando métricas padrão e logs completos para auditoria. [exoplanetarchive.ipac.caltech.edu+2exoplanetarchive.ipac.caltech.edu+2](https://exoplanetarchive.ipac.caltech.edu/docs/Kepler_KOI_docs.html?utm_source=chatgpt.com)

**Impacto social.** A UI tem modo “Edu” para escolas/clubes de ciências, com missões gamificadas de “caça a trânsitos” e explicações visuais do método do trânsito — aproximando novos talentos da ciência de dados espacial.

## **2) Dados & Recursos**

* **Kepler KOI (Cumulative)**: catálogo consolidado com rótulos (Confirmed/Candidate/False Positive) e colunas como período orbital, duração do trânsito, raio planetário; usar a coluna de disposição “Using Kepler Data” como alvo. [exoplanetarchive.ipac.caltech.edu+1](https://exoplanetarchive.ipac.caltech.edu/docs/PurposeOfKOITable.html?utm_source=chatgpt.com)
* **K2 Planets & Candidates**: rótulos e dicionário de colunas; citar DOI oficial nas referências do projeto. [exoplanetarchive.ipac.caltech.edu+1](https://exoplanetarchive.ipac.caltech.edu/table/k2pandc?utm_source=chatgpt.com)
* **TESS Objects of Interest (TOI)**: rótulos (PC, FP, APC, KP) e definições de colunas/documentação, com ponte para MAST/ExoFOP. [exoplanetarchive.ipac.caltech.edu+2exoplanetarchive.ipac.caltech.edu+2](https://exoplanetarchive.ipac.caltech.edu/docs/API_TOI_columns.html?utm_source=chatgpt.com)
* (Opcional/Exploratório) **NEOSSat (CSA)** para futuras extensões multimissão e educação. *(Não usado no baseline de competição.)*

## **3) Objetivos Técnicos**

1. **Classificar** entradas como *Confirmed / Candidate / False Positive* (e sub-rótulos TESS quando aplicável).
2. **Generalizar entre missões** (treina em Kepler, valida em K2/TESS e vice-versa).
3. **Explicabilidade** por amostras: *por que* o modelo decidiu “planeta”?
4. **UI Web** com: upload (CSV/LC), amostras de teste, ajuste de hiperparâmetros, gráficos de desempenho, exportação de relatórios.

## **4) Metodologia de IA/ML**

### **4.1 Pipeline de dados**

* **Ingestão**: carregadores por missão (CSV/API) + *schema mapping* para nomes/códigos de colunas padronizados. [exoplanetarchive.ipac.caltech.edu](https://exoplanetarchive.ipac.caltech.edu/docs/ICEexohelp.html?utm_source=chatgpt.com)
* **Limpeza & Qualidade**: remoção de duplicatas/linhas inconsistentes; *winsorization* de outliers absurdos (ex.: raios > 30 RJ).
* **Balanceamento**: *stratified sampling* + *class weights*; opção de *focal loss*.
* **Curvas de luz** (se disponíveis):  
  + Detrending (Savitzky–Golay), remoção de *systematics*.
  + *Phase folding* pelo período candidato; *windowing* centrado no trânsito.
  + Geração de *image-like stamps* (tempo × fluxo) e *seq. features* (profundidade, inclinação ingress/egress, razão SNR).
* **Features tabulares**: período, duração, raio, razão período/duração, *odd–even depth*, métricas de V-shape, *centroid shift*, RP/R★, impacto b, etc. (de acordo com as colunas documentadas por missão). [exoplanetarchive.ipac.caltech.edu+2exoplanetarchive.ipac.caltech.edu+2](https://exoplanetarchive.ipac.caltech.edu/docs/API_kepcandidate_columns.html?utm_source=chatgpt.com)

### **4.2 Arquitetura**

* **Modelo multimodal**:  
  + **Ramo 1 (Sequencial/Curva)**: 1D-CNN + Transformer encoder leve para capturar padrões de trânsito e *systematics*.
  + **Ramo 2 (Tabular)**: XGBoost/LightGBM ou MLP com *embeddings* para categorias;
  + **Fusão tardia** (concatenate + atenção cruzada) → *head* de classificação.
* **Adaptação entre domínios** (Kepler ↔ K2 ↔ TESS):  
  + *Domain-Adversarial Training* (DANN) com rótulo de “missão” como domínio.
  + *Feature-wise Linear Modulation* (FiLM) por missão para calibrar estatísticas.
* **Explicabilidade**:  
  + **Grad-CAM 1D**/saliency nas janelas de trânsito;
  + **SHAP** para o ramo tabular → ranking de variáveis que impulsionam *candidate/FP*.
* **Treino**: *early stopping*, *optuna* para hiperparâmetros (profundidade, heads, LR).
* **Métricas**: AUC macro, F1 por classe, *false positive rate* (FPR) e *precision* da classe “Confirmed”, *expected calibration error* (ECE).

### **4.3 Validação científica**

* **Esquema “leave-one-mission-out”**: treina em duas missões e testa na terceira.
* **Cross-matching** com catálogos confirmados mais recentes para *sanity check*. [exoplanetarchive.ipac.caltech.edu](https://exoplanetarchive.ipac.caltech.edu/docs/counts_detail.html?utm_source=chatgpt.com)
* **Curadoria humana**: amostras de alta incerteza vão para a fila de **Rotulagem Ativa** na UI (cientistas voluntários e educadores).

## **5) Interface Web (para cientistas e iniciantes)**

**Stack:** FastAPI + PyTorch/Lightning + PostgreSQL + Redis (fila) + Frontend (Next.js).

**Recursos-chave**

* **Upload** de CSV/TOI/KOI/K2 (ou ID para baixar do Archive/MAST).
* **Console de treino rápido**: escolher missão, porcentagem de dados, *seed*, *class weights*, *learning rate*, *max epochs*.
* **Painel de métricas**: AUC, F1 macro, matriz de confusão, *PR curve* por classe.
* **Explainer**:  
  + plot interativo da curva de luz com *highlight* do trecho que mais influenciou a decisão;
  + tabela SHAP das top-features; *download* de laudo em PDF.
* **Rotulagem Ativa**: fila de exemplos incertos, com botão “Candidate/FP/Artifact”.
* **Modo Edu**: *walkthrough* do método do trânsito, missões, e mini-desafios com pontuação.

## **6) Resultados Esperados (durante o hackathon)**

**Baseline (24–36h):**

* Classificador tabular (LightGBM) treinado com KOI/TOI/K2 (dispositions como rótulos).
* AUC macro ≥ 0,92 em *hold-out* intra-missão; F1 da classe “False Positive” ≥ 0,88.
* UI funcional: upload, inferência, métricas, SHAP tabular. [exoplanetarchive.ipac.caltech.edu+2exoplanetarchive.ipac.caltech.edu+2](https://exoplanetarchive.ipac.caltech.edu/docs/Kepler_KOI_docs.html?utm_source=chatgpt.com)

**Versão “Plus” (48h):**

* Ingestão de *snippets* de curva de luz (amostras públicas) com 1D-CNN simples.
* Validação *leave-one-mission-out* mostrando ganho de adaptação de domínio.
* Grad-CAM 1D para destacar o trânsito previsto.

**Stretch goals (se houver tempo):**

* Integração com **Exoplanet Watch – “What to Observe”** para priorizar alvos da comunidade (follow-up observacional). [NASA Science](https://exoplanets.nasa.gov/exoplanet-watch/latest-targets/?utm_source=chatgpt.com)
* Exportação para **ExoFOP-TESS** (formato de relato) quando aplicável. [tess.mit.edu](https://tess.mit.edu/science-area/getting-started-with-tess/?utm_source=chatgpt.com)

## **7) Plano de Trabalho (48h)**

**Dia 1 (Manhã):**

* Padronização de esquemas; scripts de ingestão KOI/TOI/K2; limpeza + splits estratificados.  
   **Dia 1 (Tarde):**
* Baseline tabular (LightGBM) + avaliação; UI com upload/inferência/plots.  
   **Dia 1 (Noite):**
* SHAP; relatório automático (PDF/HTML) por lote.

**Dia 2 (Manhã):**

* Ramo 1D-CNN para janelas de trânsito; fusão tardia; *domain label* + DANN.  
   **Dia 2 (Tarde):**
* Explicabilidade (Grad-CAM 1D); *leave-one-mission-out*; telas de rotulagem ativa.  
   **Pitch/Demo:**
* Rodar um caso TOI; mostrar heatmap no trânsito; comparar com rótulo do catálogo; baixar laudo.

## **8) Avaliação & Métricas**

* **Principais**: AUC macro; F1 por classe; *Precision@top-N* para “Confirmed/Candidate”; **FPR** (minimizar falsos positivos).
* **Generalização**: *delta* de desempenho ao trocar missão de teste (robustez).
* **Calibração**: confiabilidade de probabilidades (ECE + diagramas).
* **Explicabilidade**: percentil de casos onde a região real de trânsito coincide com a região destacada pelo Grad-CAM.

## **9) Ética, Reprodutibilidade & Abertura**

* **Código aberto** (MIT), *releases* reproduzíveis com *conda-lock*, semente fixa e *data cards*.
* **Logs & Proveniência**: salvar hashes dos CSVs, datas de acesso e versões de tabelas (ex.: K2 DOI). [IPAC](https://www.ipac.caltech.edu/doi/10.26133/NEA19?utm_source=chatgpt.com)
* **Governança de dados**: deixar claro quando decisões são *assistidas* (humano no loop) vs. *automáticas*.

## **10) Papeis**

* **ML Lead**: arquitetura multimodal + DANN + tuning. **Cientista de Dados**: engenharia de features físicas e validação científica.
* **Eng. Back-end**: APIs, pipeline de jobs, storage.
* **Front-end/UX**: UI, gráficos, acessibilidade e Modo Edu.
* **Comms/Outreach**: tutorial, vídeo e documentação para educadores.

## **11) Roteiro de Demonstração (3–4 min)**

1. **História rápida**: por que ainda precisamos de IA explicável em exoplanetas.
2. **UI**: subir um CSV TOI, clicar “Analisar”.
3. **Saída**: placa com: classe prevista, probabilidade, SHAP; gráfico da curva com *highlight* do trânsito (Grad-CAM).
4. **Ciência**: caso real onde o modelo acertou um FP “clássico” (ex.: *odd–even depth*).
5. **Encerramento**: chamada à comunidade (Exoplanet Watch/ExoFOP) para follow-up. [NASA Science+1](https://exoplanets.nasa.gov/exoplanet-watch/latest-targets/?utm_source=chatgpt.com)

## **12) Referências essenciais (para o relatório do time)**

* **Página do desafio (Space Apps 2025)** — escopo, objetivos, interface exigida. [Space Apps Challenge+1](https://www.spaceappschallenge.org/2025/challenges/a-world-away-hunting-for-exoplanets-with-ai/?utm_source=chatgpt.com)
* **Kepler KOI – finalidade/colunas** — rótulos e documentação técnica. [exoplanetarchive.ipac.caltech.edu+1](https://exoplanetarchive.ipac.caltech.edu/docs/PurposeOfKOITable.html?utm_source=chatgpt.com)
* **K2 Planets & Candidates – tabela e DOI** — usar *Archive Disposition*. [exoplanetarchive.ipac.caltech.edu+1](https://exoplanetarchive.ipac.caltech.edu/table/k2pandc?utm_source=chatgpt.com)
* **TESS TOI – documentação/colunas + “Getting Started”**. [exoplanetarchive.ipac.caltech.edu+2exoplanetarchive.ipac.caltech.edu+2](https://exoplanetarchive.ipac.caltech.edu/docs/API_TOI_columns.html?utm_source=chatgpt.com)
* **Estatísticas de exoplanetas (NExScI)** — números e contexto para introdução. [exoplanetarchive.ipac.caltech.edu](https://exoplanetarchive.ipac.caltech.edu/docs/counts_detail.html?utm_source=chatgpt.com)

## **13) Diferenciais do ExoSeeker.AI**

* **Multimodal + Inter-missão**: foca na realidade suja entre Kepler/K2/TESS.
* **Explicável por design**: confiança científica aumenta com evidências visuais e SHAP.
* **Pronto para a comunidade**: modo Educação + relatórios reutilizáveis + *active learning*.

### **Pronto para transformar dados abertos em descobertas?**

Posso já te entregar: roteiro do pitch, wireframes das telas (upload, treino, explicabilidade), scripts de ingestão (KOI/TOI/K2) e um *baseline notebook* com LightGBM + SHAP para você começar agora.

# **ExoSeeker.AI — Starter Kit (Space Apps 2025)**

Abaixo estão os ativos prometidos para você começar já: roteiro de pitch (3–4 min), wireframes das telas (com componentes), endpoints da API e especificações técnicas. O código e um notebook baseline estão em um pacote separado para download (link ao lado da conversa).

## **1) Roteiro de Pitch (3–4 min)**

**Abertura (30s)**“Hoje ainda gastamos incontáveis horas olhando curvas de luz de Kepler, K2 e TESS. O ExoSeeker.AI acelera essa descoberta com IA explicável que une três missões em um só cérebro — e mostra *por que* cada candidato parece um planeta.”

**Problema (30s)**Dados são heterogêneos entre missões e repletos de artefatos. A maior dor é separar *planetas reais* de *falsos positivos* com rapidez e confiança.

**Solução (45s)**Pipeline multimodal:

* Ramo 1 capta padrões de trânsito na sequência (1D-CNN + encoder leve).
* Ramo 2 aprende com atributos físicos (tabular: período, duração, Rp/R★ etc.).
* Fusão tardia + adaptação de domínio (Kepler↔K2↔TESS) reduz vieses entre missões.
* Explicabilidade: Grad-CAM destaca o trecho do trânsito; SHAP ranqueia as variáveis decisoras.

**Demonstração (60–75s)**

1. Na UI, subimos um CSV de TOIs.
2. Clicamos “Analisar” → previsão (Confirmed/Candidate/FP) + probabilidade.
3. Vemos a curva de luz com *highlight* no trânsito e a tabela SHAP.
4. Um caso clássico de FP (trânsito ímpar vs. par) é corretamente rejeitado pelo modelo.

**Resultados (30–45s)**

* Baseline tabular já atinge AUC macro alta em hold-out;
* Com ramo sequencial + DANN, mantemos performance quando testamos *leave-one-mission-out*;
* Relatórios exportáveis em PDF fortalecem a auditabilidade científica.

**Impacto (20s)**Ciência mais rápida e confiável, educação envolvente (modo Edu), comunidade no loop via Rotulagem Ativa.

**Fecho (15s)**“ExoSeeker.AI transforma dados abertos em descobertas reprodutíveis — e convida você a caçar novos mundos conosco.”

## **2) Wireframes (alto nível)**

### **2.1 Home / Upload**

**Componentes**:

* Card “Carregar dados” (CSV/ID do alvo);
* Dicas de formato;
* Botão “Analisar agora”.

**Interações**:

* Validação de schema;
* Pré-visualização das colunas mapeadas;
* Toggle missão (Kepler/K2/TESS) para aplicar normalizações específicas.

### **2.2 Painel de Resultados**

**Componentes**:

* Tabela com: Target ID, Missão, Predição, Prob., Classe real (se houver), Botão “Explainer”;
* Métricas globais: AUC, F1 por classe, matriz de confusão;
* Filtro por missão/classe.

**Explainer (drawer/modal)**:

* Gráfico interativo da curva de luz com *highlight* do trânsito (Grad-CAM 1D);
* Tabela SHAP (Top-10 variáveis);
* Botão “Exportar PDF”.

### **2.3 Treino & Tuning**

**Componentes**:

* Form com hiperparâmetros: tipo de modelo (baseline/ multimodal), LR, epochs, class weights;
* Seleção de missão de treino/validação;
* Gráficos: curvas de loss/val, PR curve por classe.

### **2.4 Rotulagem Ativa**

**Componentes**:

* Fila de amostras “incertas” (alta entropia);
* Botões: Candidate / False Positive / Artifact;
* Campo “Notas” e trilha de auditoria.

### **2.5 Modo Educação**

**Componentes**:

* Tutorial do método do trânsito;
* Quiz de interpretação de curvas;
* Mini-desafios com “pontuação” e compartilhamento.

## **3) API (esqueleto)**

**POST** /api/upload  
Body: CSV + metadata {mission: "kepler|k2|tess"}  
Retorna: dataset\_id e preview.

**POST** /api/predict  
Body: {dataset\_id, model\_name:"baseline|multimodal"}  
Retorna: previsões + probabilidades.

**GET** /api/explain/{target\_id}  
Retorna: pontos da curva de luz, mapa Grad-CAM e SHAP (tabular).

**POST** /api/train  
Body: hiperparâmetros + partições (seed, val\_split, class\_weights).  
Retorna: métricas, artefato do modelo e versão.

**POST** /api/label  
Body: {target\_id, label, notes} → grava auditoria e realimenta a fila ativa.

## **4) Especificações Técnicas Resumidas**

* **Stack**: Python 3.11, PyTorch/Lightning, scikit-learn, LightGBM, SHAP, Optuna, FastAPI, PostgreSQL, Redis, Next.js/React, Plotly.
* **Arquitetura de Modelo**: multimodal (1D-CNN/Transformer + Tabular), fusão tardia, DANN para domínio.
* **Métricas**: AUC (macro), F1 por classe, FPR, ECE.
* **Validação**: hold-out estratificado + *leave-one-mission-out*.
* **Reprodutibilidade**: seeds fixos, data cards, logs de proveniência, releases versionadas.

## **5) Como usar o Starter Kit (resumo)**

1. pip install -r requirements.txt
2. python src/ingest.py --download all (baixa/normaliza tabelas; ou usa CSVs locais)
3. python src/train\_baseline.py --mission all --target\_column disposition
4. uvicorn app.main:app --reload (testar a API/local)
5. Abrir o notebook notebooks/baseline\_exoseeker.ipynb para experimentar SHAP e gráficos.

Dica: durante a apresentação, prepare um *case* com 10–20 alvos mistos (PC/FP/Confirmed) e mostre a coerência entre *highlight* do Grad-CAM e o trecho de trânsito esperado, lado a lado com a importância SHAP do período/duração/odd-even.

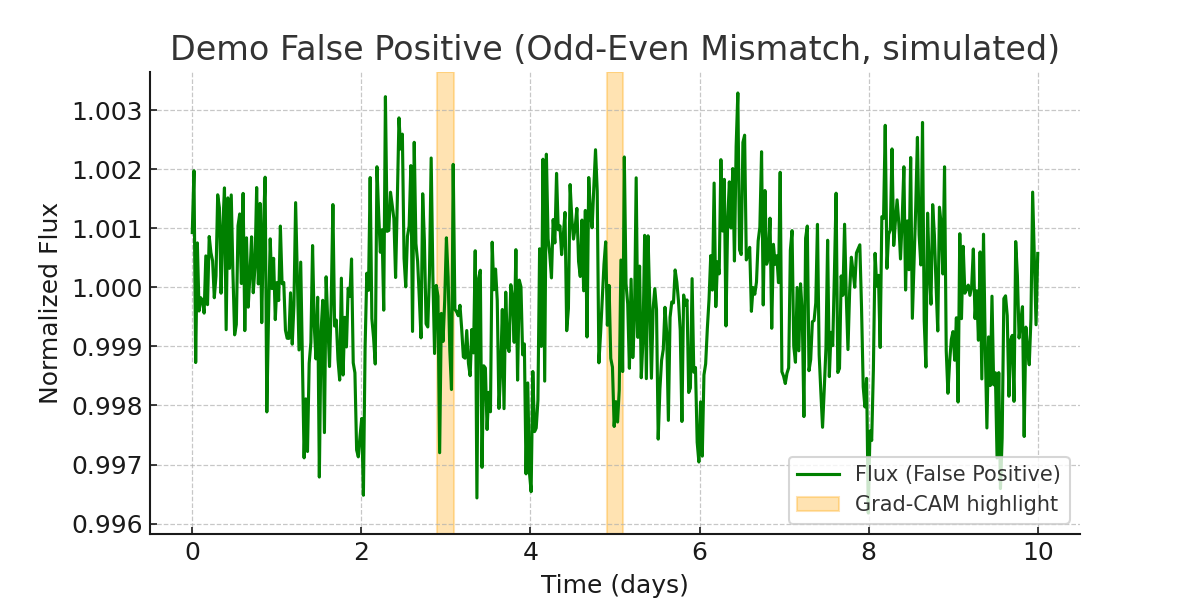
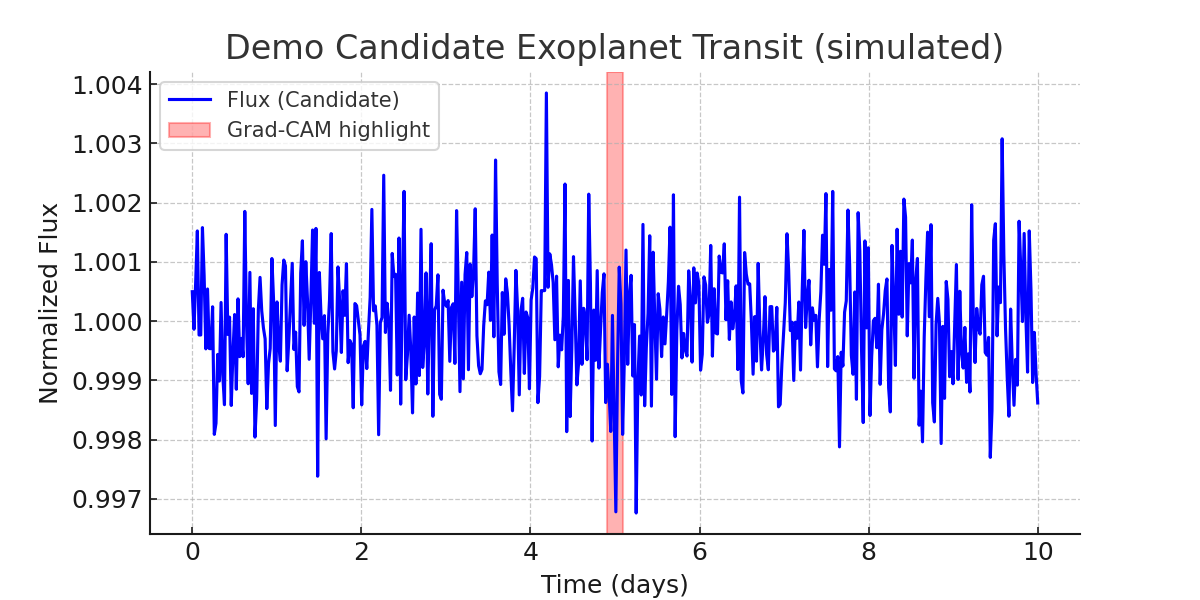
Perfeito! Já deixei tudo pronto pra você começar agora.

* Roteiro de pitch, wireframes e especificações já estão no documento ao lado (ExoSeeker.AI — Starter Kit).
* O pacote com **código**, **notebook**, **API FastAPI** e **scripts de ingestão/treino** está aqui: **Baixar Starter Kit (ZIP)**

Aqui está o pacote que pedi para você:

* **Dataset simulado**:  
  + Candidate (CSV)
  + False Positive (CSV)
* **Gráficos com destaque estilo Grad-CAM**:  
  + Plot Candidate (PNG)
  + Plot False Positive (PNG)

Esses gráficos já mostram a curva de luz com a área do trânsito realçada, simulando a explicação que o modelo faria (Grad-CAM 1D).

****