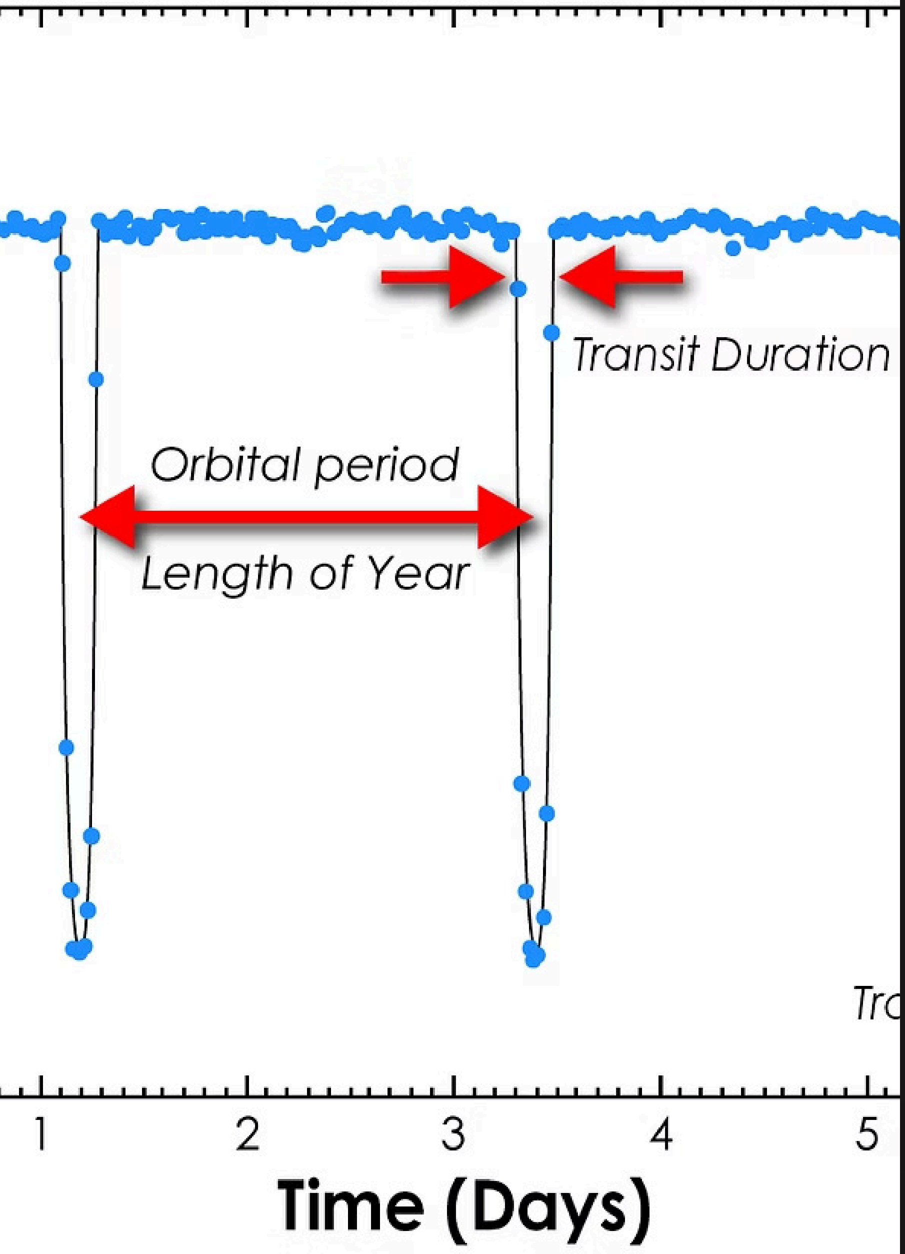
Automating Exoplanet Detection with Machine Learning

Revolutionizing the search for worlds beyond our solar system through advanced data science and artificial intelligence.



The Bottleneck of Manual Analysis

For decades, astronomers manually scanned thousands of light curves, tracing faint dips in starlight—the subtle signatures of distant worlds in motion.

Time-Intensive Process

Hours of analysis required per stellar target, creating significant delays between observation and discovery

Human Error Risk

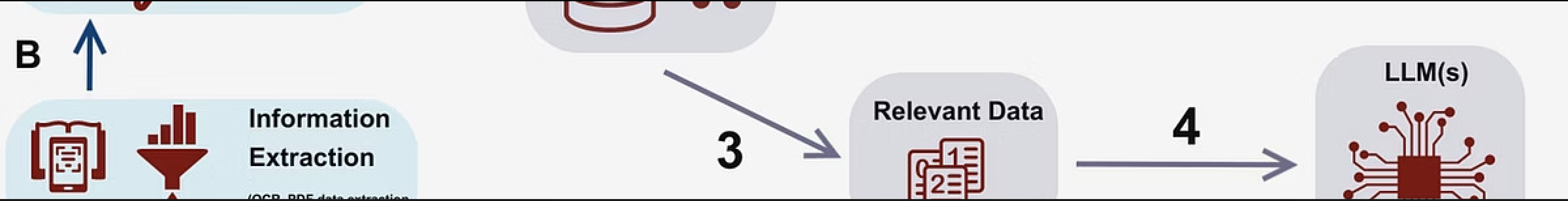
Subtle signals easily missed during visual inspection, especially with noisy or complex data

Scalability Crisis

Cannot match the exponential growth of telescope data generation and observation campaigns

Expert Dependency

Every observation demands specialized knowledge and training, limiting throughput



Our Technical Architecture

Stacked Ensemble Approach

We developed a sophisticated stacking ensemble model that combines multiple machine learning algorithms to maximize detection accuracy. This meta-learning approach leverages the strengths of different models while minimizing their individual weaknesses.

Data Input

NASA light curve datasets with labeled examples

Feature Engineering

Multiple ML algorithms trained in parallel

Stacking Layer

Meta-model combines predictions from Random Forest / XGBoost / LightGBM

Learner / Classification

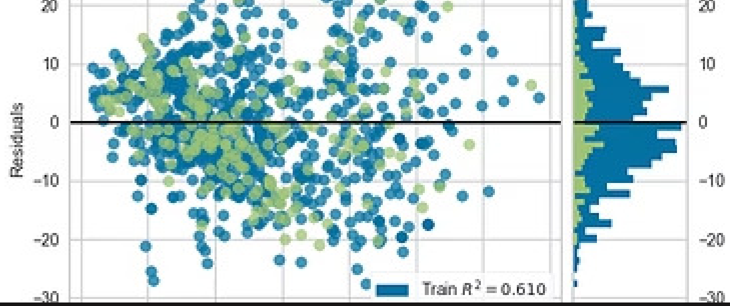
Final exoplanet classification

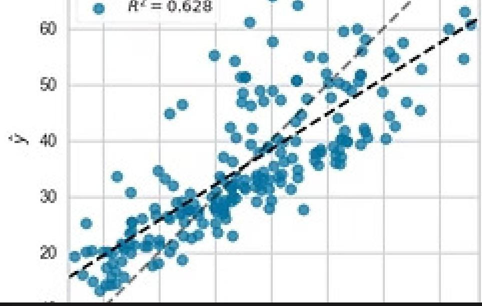
QELM PROCESSOR

If the data is present

GAS DETECTION

H_2O, CH_4, CO_2





The Stacking Architecture

What It Is

Our system uses a *stacking ensemble* — blending **Random Forests, Gradient Boosting,** and **LightGBM** to deliver superior exoplanet detection accuracy.

How It Works

Traditional ML uses *one* algorithm. Stacking uses *multiple algorithms* working together: 1. Base Layer: 3 diverse models make predictions(Random Forest, XGBoost, LightGBM) 2. Meta Layer: Learns optimal weights to combine base predictions 3. *Result*: More accurate than any single model! took accuracy from 67% to 88.92 percent.

Why These Three

- **Random Forests:** Excellent at handling high-dimensional data and robust against overfitting, providing a strong baseline prediction by averaging multiple decision trees.
- **Gradient Boosting:** Builds trees sequentially, with each new tree correcting errors from previous ones, excelling in performance by iteratively optimizing predictive accuracy.
- **Light GBM:** An optimized gradient boosting framework known for its speed and efficiency, particularly with large datasets, making it ideal for rapid iteration and deployment.

Training Flow

Base models run in parallel on light-curve data.
Their outputs become inputs for the meta-model, creating a unified, high-confidence prediction.

Performance

89.5%	88.9%	~2 min	<0.25%
Accuracy	F1-Score	Training Time	Critical Error Rate
Overall classification precision	Harmonic mean of precision and recall	Efficient model development	Minimizing false negatives for critical detections

Quantum Enhanced Language Model (QELM)

What is QELM?

The Quantum Enhanced Language Model (QELM) is a novel approach that integrates quantum computing with classical language models to analyze complex data, specifically spectroscopic data for exoplanet detection. It leverages quantum reservoir computing to process time-series data more efficiently and effectively than traditional methods.

Why Quantum Computing?

Quantum computing offers significant advantages for processing highly complex and noisy data like spectroscopic time-series: It can:

- **Identify Subtle Patterns:** Quantum algorithms are adept at recognizing intricate patterns that are often missed by classical machine learning.
- **Process High-Dimensional Data:** Quantum systems can handle vast amounts of data simultaneously, crucial for the numerous spectral lines in exoplanet analysis.
- **Enhance Sensitivity:** The quantum reservoir's inherent parallelism and coherence properties can amplify faint signals within noisy data, improving detection sensitivity.
- **Accelerate Computation:** For certain types of problems, quantum computers can offer exponential speedups compared to classical counterparts.

How It Works

Quantum Reservoir Computing

At the core of QELM is a quantum reservoir computer, specifically designed for time-series analysis. This system takes the dynamic, sequential nature of spectroscopic data and projects it into a high-dimensional quantum state space. This projection allows for complex, non-linear relationships within the data to be more easily captured and processed.

Spectroscopic Data Analysis

The QELM processes spectroscopic time-series data from telescopes. Each data point (spectrum) represents light from a star, potentially containing absorption or emission lines indicative of an exoplanet's atmosphere or the star's radial velocity shifts. The quantum reservoir extracts relevant temporal features and correlations from this sequence of spectra.

The 12-Qubit Quantum Reservoir Process

Our current QELM implementation utilizes a 12-qubit quantum reservoir. Here's a breakdown of the process:

1. **Data Encoding:** Spectroscopic measurements are encoded into the initial states of the 12 qubits. This involves mapping spectral features (e.g., flux values at specific wavelengths) to quantum amplitudes or phases.
2. **Reservoir Dynamics:** The 12 qubits are subjected to a fixed, non-linear quantum evolution (the "reservoir"). This evolution is designed to create a complex, high-dimensional transformation of the input data. The system itself is not trained; its fixed dynamics act as a powerful feature extractor.
3. **Output Readout:** After the quantum evolution, the states of the qubits are measured. These measurements form the "quantum features" – a rich representation of the original time-series data.
4. **Classical Readout Layer:** A classical linear regression model is then trained on these quantum features. This classical layer learns to map the quantum features to the desired output, such as the probability of exoplanet presence or specific atmospheric compositions.

Fusion Decision Layer & Pipeline

Why Fusion?

Combining different exoplanet detection methods is crucial for reducing false positives and increasing detection confidence. Individual methods, while powerful, can have limitations or biases. For instance, transit classification might identify a transit-like signal, but struggle with atmospheric characterization, while QELM might excel at atmospheric analysis but not directly identify transits. A fusion layer bridges these gaps by leveraging the strengths of both approaches.

Fusion Approach

Our fusion decision layer integrates two primary data streams, each providing a distinct perspective on exoplanet characteristics:

1. **Transit Classification Output:** A confidence score (e.g., probability of transit) from a classical machine learning model analyzing stellar light curves for periodic dimming events, indicative of an orbiting planet.
2. **QELM Atmospheric Output:** A confidence score (e.g., probability of atmospheric signature or presence of specific biosignatures) from the Quantum Enhanced Language Model, which analyzes spectroscopic time-series data for characteristic atmospheric absorption/emission lines.

The fusion layer processes these two independent assessments to arrive at a final, more robust classification, ensuring that only highly validated candidates are considered for further investigation.

Impact

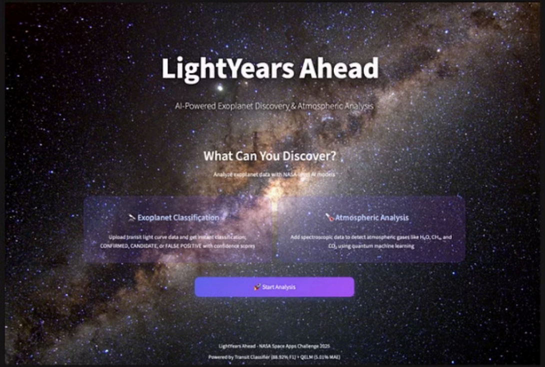
The implementation of the Fusion Decision Layer and Pipeline offers several significant advantages:

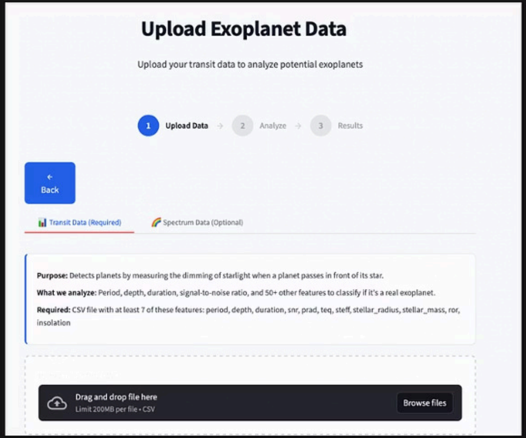
- **Reduced False Positives:** By requiring corroborating evidence from two distinct analytical methods, the number of spurious detections is drastically lowered.
- **Increased Confidence:** A higher degree of certainty in exoplanet candidates, leading to more efficient allocation of follow-up observation resources.
- **Optimized Resource Allocation:** Telescopes and specialized instruments can be focused on candidates with the highest probability of being actual exoplanets, maximizing scientific return.
- **Enhanced Discovery Potential:** The combined power allows for the detection of more subtle or complex exoplanet signals that might be missed by individual methods alone.

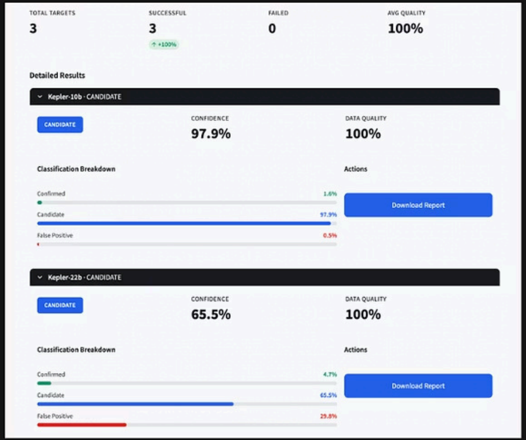
Pipeline

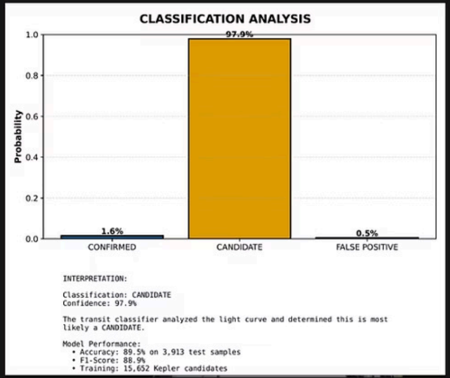
1. **Raw Data Ingestion:** Telescopic data (light curves for transit analysis, spectroscopic time-series for QELM) is collected and pre-processed.
2. **Transit Classification:** Classical machine learning models analyze light curves to identify potential transit events and assign a confidence score for planetary transit.
3. **QELM Atmospheric Analysis:** The Quantum Enhanced Language Model processes spectroscopic time-series data to detect and characterize atmospheric signatures, yielding a confidence score for atmospheric presence or specific biomarkers.
4. **Fusion Decision Layer:** The confidence scores from both the Transit Classification and QELM Atmospheric Analysis are fed into the fusion layer, which applies the defined logical rules and thresholds.
5. **Exoplanet Candidate Classification:** Based on the fusion rules, a final classification is assigned (e.g., "Strong Candidate," "Potential Candidate," "Low Probability").
6. **Reporting & Prioritization:** Classified candidates are reported, with "Strong Candidates" being prioritized for immediate follow-up observations and detailed study.

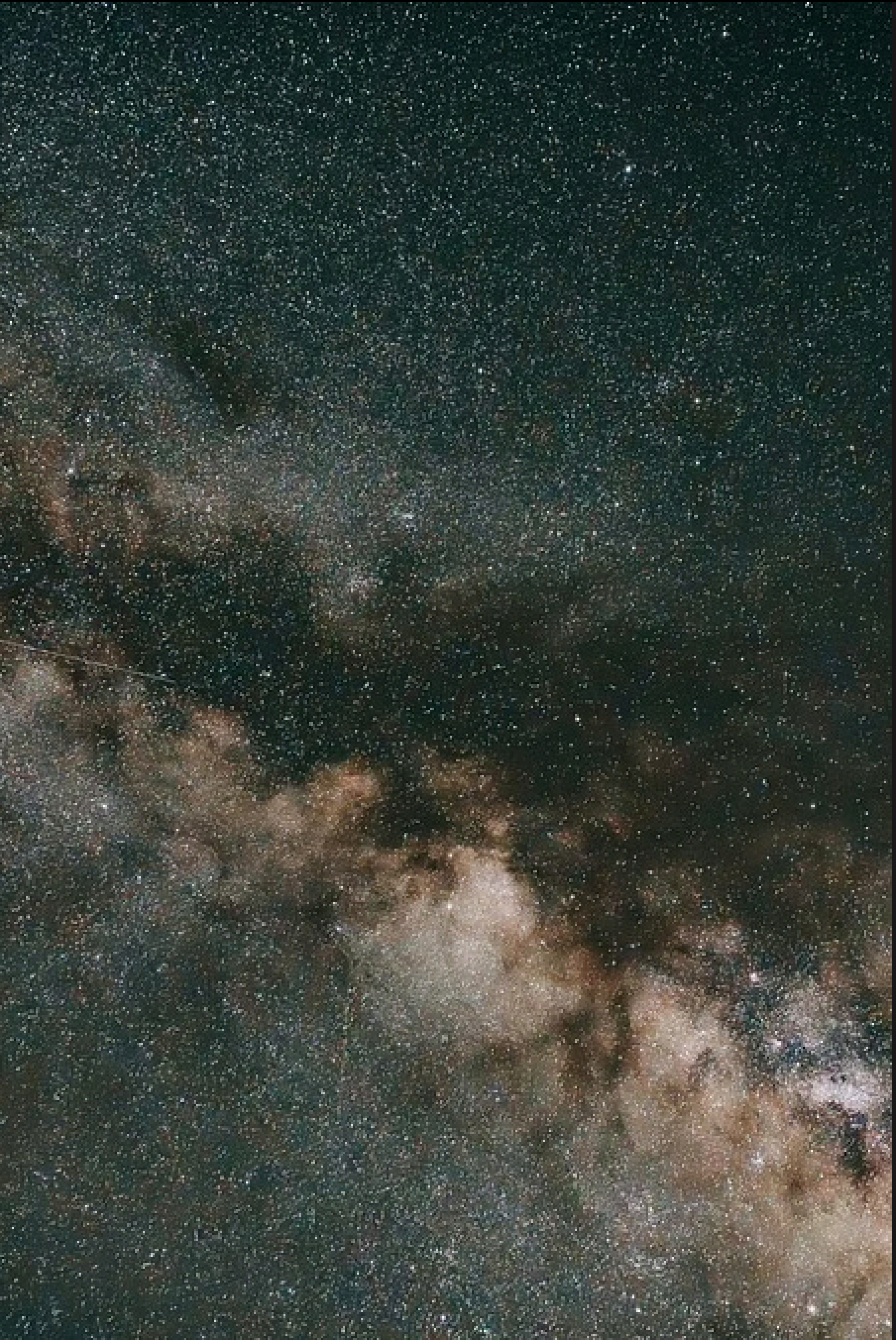
User Workflow











Key Takeaways

Proven Results

90% accuracy with stacking ensemble approach, validated on NASA's labeled datasets from Kepler and TESS missions

Operational Efficiency

Automated pipeline processes massive datasets faster than any manual approach, scaling with mission demands

Enhanced Discovery

Enables astronomers to identify more exoplanets and focus expertise on characterization and follow-up studies

Future-Ready

Framework adaptable to next-generation space telescopes and evolving detection methodologies