"Clutch or Cry" Team at TRACS @ WASP 2025:
A Hybrid Stacking Ensemble for Astrophysical Document Classification

Arshad Khatib and **Aayush Prasad** and **Rudra Trivedi**
Department of Artificial Intelligence
And
Shrikant Malviya
Department of Computer Science and Engineering
SVNIT Surat
{u24ai112, u24ai091, u24ai068}@aid.svnit.ac.in, shrikant@coed.svnit.ac.in

Abstract

Automatically identifying telescopes and their roles within astrophysical literature is crucial for large-scale scientific analysis and tracking instrument usage patterns. This paper describes the system developed by the "Clutch or Cry" team for the Telescope Reference and Astronomy Categorization Shared task (TRACS) at WASP 2025 (Grezes et al., 2025). The task involved multi-class telescope identification (Task 1) and multi-label role classification (Task 2) within scientific papers. For Task 1, we employed a feature-engineering approach centered on document identifiers (Id suffix) combined with metadata and textual features, utilizing a tuned Random Forest classifier to achieve high accuracy. For the more complex Task 2, we utilized a carefully designed two-level stacking ensemble. Level-0 combines a rule-based keyword classifier with the domain-adapted astroBERT transformer, effectively fusing symbolic and semantic information. Level-1 uses four independent XGBoost meta-learners for targeted per-role optimization. These architectures address the primary challenges: handling long documents and managing severe class imbalance in Task 2 (notably 1:91 for instrumentation). Systematic optimization focused on mitigating imbalance significantly improved Task 2 performance for minority classes. This work validates the effectiveness of tailored approaches for distinct subtasks and targeted optimization for imbalanced classification in specialized scientific domains.

1 Introduction

Automated classification of scientific literature is critical for knowledge discovery and resource management in large-scale research repositories. With millions of astrophysical papers archived in systems like the NASA Astrophysics Data System (ADS), manual annotation and categorization become infeasible (SAO/NASA Astrophysics Data System, 2025). Effective automated methods enable researchers to quickly identify relevant studies, track telescope usage patterns, understand instrumental capabilities, and trace scientific methodologies—ultimately accelerating scientific discovery and facilitating data-driven insights into observational astronomy practices (Wikipedia contributors, 2025). This capability extends beyond administrative utility, directly supporting evidence synthesis, reproducibility verification, and interdisciplinary research collaboration.

The Telescope Reference and Astronomy Categorization Shared task (TRACS) presents two intertwined classification challenges that together model real-world requirements faced by digital astronomy libraries (Kaggle, 2025). The task demands systems capable of identifying which telescopes are discussed as primary subjects versus peripheral mentions, and distinguishing the functional role of telescopes within scientific contexts—whether used for data acquisition, instrument characterization, or comparative analysis. These distinctions are semantically nuanced, often embedded in lengthy papers with inconsistent terminology, and severely imbalanced across class distributions. This shared task provides an ideal proving ground for advancing both fundamental NLP techniques and domain-specific adaptations needed for specialized scientific corpora, offering valuable insights into how machine learning systems can handle real-world complexity in domain-specific document understanding.

Addressing these challenges—long document context, nuanced semantic roles, and severe class imbalance—requires a robust and adaptable classification architecture. Simple models often struggle

Proceedings of the Third Workshop for Artificial Intelligence for Scientific Publications, pages 146–156
December 23, 2025 ©2025 Association for Computational Linguistics

with the sheer length of scientific papers and are overwhelmed by the majority of classes. We hypothesize that a hybrid stacking ensemble method offers a compelling solution. By combining a fast, symbolic keyword classifier (effective at broad categorization and handling explicit mentions across long texts) with a deep semantic model like astroBERT (capable of understanding nuanced context within specific text windows), we can leverage complementary strengths. Furthermore, employing a stacking architecture with independent, per-class meta-learners enables highly targeted optimization, allowing us to apply aggressive techniques such as class weighting and threshold tuning precisely where needed to combat the extreme class imbalance observed in the TRACS dataset. The specialized multi-level approach forms the core of our system design.

1.1 Shared Task: TRACS-2025

The shared task (Kaggle, 2025) comprises two classification objectives:

- **Task 1 (Telescope Identification):** Multi-class classification identifying the primary telescope discussed in a paper from the set {`CHANDRA`, `HST`, `JWST`, `None`}.
- **Task 2 (Role Classification):** Multi-label classification determining telescope roles with four binary labels: `science`, `instrumentation`, `mention`, and `not_telescope`.

1.2 Key Challenges

Two major challenges characterize this task:

Document Length and Context: Full-text scientific articles frequently exceed the input token limits of standard transformer models (typically 512 – 2048 tokens), requiring careful strategies for capturing relevant information from lengthy documents.

Severe Class Imbalance: Both tasks exhibit pronounced class imbalance. In **Task 1 (Telescope Identification)**, the distribution is extremely skewed. The **`NONE`** class represents a tiny fraction of the dataset (approximately 1 instance for every 273 samples), making it vastly outnumbered by majority classes like `HST` (which appears roughly 126 times more often than `NONE`). In **Task 2 (Document Role Classification)**, the **`instrumentation`** class appears with a positive-to-negative ratio of approximately 1:91, while **`not_telescope`** exhibits a ratio closer to 1:9. This extreme imbalance renders standard machine learning approaches ineffective, as models naturally bias toward majority classes.

We address these challenges through an ensemble-based methodology that combines symbolic and semantic models. Instead of optimizing a single model architecture, we leverage the complementary strengths of combined rule-based and neural approaches, enabling targeted optimization for each of the four output labels.

Our contribution includes:

1. A carefully designed two-level stacking architecture.
2. Systematic methodology for addressing extreme class imbalance through multiple complementary techniques.
3. Empirical validation that per-class optimization significantly improves performance on minority classes.

All code and trained models will be released publicly later to ensure reproducibility. Link: `https://github.com/Arshad-13/ClutchOrCry-TRACS-2025`

2 Related Work

Handling imbalanced classification is a well-studied problem. Common approaches include oversampling techniques such as SMOTE (Synthetic Minority Oversampling Technique) (Chawla et al., 2002), undersampling (Kubat and Matwin, 1997), cost-sensitive learning (Elkan, 2001), and ensemble methods (Galar et al., 2012). In NLP, handling imbalanced text classification has been addressed through various techniques, including threshold adjustment for optimal F1 scores (Zou et al., 2016; Hong et al., 2016) and cost-sensitive learning strategies (Elkan, 2001; Lee and Kim, 2020). Threshold adjustment helps by shifting the decision boundary away from the default 0.5 probability; it allows the model to correctly identify more instances of the rare class, often improving recall and the F1-score even if precision decreases slightly. Cost-sensitive learning directly tackles the imbalance during training by assigning a higher penalty for misclassifying minority class instances, forcing the model to learn features that better distinguish the rare class from the majority class. These techniques are pertinent to both tasks, given the severe class imbalance observed.

Rule-based systems and extensive feature engineering are often employed in scientific document classification, particularly when structured identifiers or metadata offer strong predictive signals. Approaches leveraging bibcodes or similar identifiers for categorization are common in bibliographic analysis and information retrieval within specific scientific domains. These methods excel at precision when identifier patterns are consistent but may lack robustness to variations or require supplementation with other features. Hybrid approaches, combining rule-based extraction with machine learning models trained on engineered features (including text statistics, keyword counts, and metadata), aim to balance the high precision of rules with the broader pattern recognition capabilities of models like Random Forests, a strategy reflected in our Task 1 architecture.

Domain adaptation of pre-trained transformers has also proven effective for specialized NLP tasks. Recent work in scientific document understanding has leveraged domain-specific models like SciBERT (Beltagy et al., 2019) and BioBERT, demonstrating that pre-training on domain-specific corpora improves downstream task performance. For astronomical text, `astroBERT` (Grezes et al., 2021) provides pre-training on 440,000 astrophysical abstracts from the NASA Astrophysics Data System, offering domain-specific vocabulary and patterns critical for astronomy-related classification tasks, which we utilize in our Task 2 system.

Ensemble methods that combine diverse classifiers have demonstrated strong performance on imbalanced problems (Galar et al., 2011; Khan et al., 2023). While simpler ensemble models, such as Random Forest (used in Task 1), inherently handle feature interactions, stacking ensembles, in particular, allow meta-learners to learn optimal combination strategies for integrating base model predictions (Nugroho et al., 2023). Our Task 2 approach extends this paradigm by using per-class meta-learners rather than a single global meta-learner, enabling fine-grained hyperparameter optimization tailored to each label's unique characteristics.

3 System Architecture

We employ distinct architectures tailored to the specific requirements of each task. Task 1 focuses on identifying the primary telescope using *rule-based features and a Random Forest*, while Task 2 uses a *stacking ensemble method* to classify the role of the document concerning telescopes.

3.1 Task 1: Telescope Identification Architecture

For identifying the primary telescope associated with an astrophysical document, our system employs a feature-engineering-centric approach, culminating in a *Random Forest classification model*. This architecture prioritizes extracting strong signals from the document identifier (`Id`), supplemented by metadata and textual features to enhance robustness and handle edge cases.

3.1.1 Rule-Based Feature Extraction (ID Suffix)

The cornerstone of this system is the extraction and encoding of information presumed to be embedded within the document's `Id` field, often structured similarly to astrophysical bibcodes.

Primary Rule: The system identifies the suffix following the last underscore (_) character in the `Id` string. **Mapping:** Recognized suffixes (e.g., `CHANDRA`, `HST`, `JWST`) are directly mapped to their corresponding telescope labels. `Id` strings without a recognized suffix or underscore are assigned a default category (e.g., `NONE` or `NO_UNDERSCORE`). **Feature Encoding:** The extracted suffix string is numerically encoded (e.g., using `LabelEncoder`) to be used as a categorical feature by the classification model. Additional binary features like `has_underscore` are also generated. This explicit encoding of the rule's output provides a high-precision signal to the classifier.

3.1.2 Comprehensive Feature Engineering

To complement the primary ID suffix feature and improve classification accuracy, especially for documents where the ID rule is insufficient, a wide array of supplementary features are engineered:

ID/Bibcode Characteristics: Features derived from the `Id` string itself, including its total length, the count of underscores, and the categorical prefix (often representing the year or journal, also label encoded).

Metadata Features: Utilizing the provided `year`, including derived features like the difference from a reference year and flags indicating publication eras (e.g., recent JWST era, pre-Chandra era).

Textual Content Features: *Length Features:* Character lengths of fields such as `title`, `abstract`, and `body`. Word counts are also included for key fields. *Keyword Mentions:* Binary flags and counts indicating the presence of specific telescope names (Chandra, JWST, Hubble/HST) within the `title`, `abstract`, body, and `acknowledgments`. *TF-IDF Representation:* Term Frequency-Inverse Document Frequency vectors generated from the combined text of the `title`, `abstract`, and body fields, using a constrained vocabulary (e.g., 150 features) and considering unigrams and bigrams.

Author & Grant Features: Simple features like author count and binary flags for the presence of grant or acknowledgment text.

3.1.3 Classification Model (Random Forest)

The final classification is performed by a `RandomForestClassifier` model. It takes a concatenated feature vector comprising all the engineered numeric/categorical features (including the encoded ID prefix) and the sparse TF-IDF text features.

Training and Hyperparameter Tuning: The model is trained on the full set of derived features. To optimize performance, hyperparameters were tuned using `RandomizedSearchCV` with 5-fold stratified cross-validation. The best parameters identified were:

- `n_estimators`: 200
- `max_depth`: 15
- `min_samples_split`: 2
- `min_samples_leaf`: 4
- `max_features`: 'sqrt'

This configuration achieved the best cross-validation accuracy of 0.7772. The final model used for prediction (`best_estimator_` from `RandomizedSearchCV`) is implicitly trained on the entire dataset using these optimal parameters. Class weighting (`class_weight='balanced'`) was also employed during the search process to mitigate the inherent imbalance in telescope label distribution. The model predicts a single categorical label representing the identified primary telescope (`CHANDRA`, `HST`, `JWST`, or `NONE`). Feature importance analysis consistently confirms that the ID suffix-derived features are the most dominant predictors, validating the hybrid rule-based and machine-learning strategy.

3.2 Task 2: Document Role Classification Architecture

As shown in Figure 1, our system for Task 2 utilizes a two-tier approach, comprising 'level 0' and 'level 1', within the stacking ensemble intended to merge quick symbolic classification with a slower yet more accurate semantic comprehension.

3.2.1 Level-0: Base Models

Rule-Based Keyword Classifier The keyword classifier provides high-recall signals through pattern matching. It utilizes a dictionary of over 1,000 domain-specific keywords (spanning telescope names, instruments, and scientific concepts), which we curated using a combination of large language models (LLMs) and established astrophysical references. Scores documents based on the presence, frequency, and contextual proximity of keywords. Outputs a 4-dimensional pseudo-probability vector, one value per output label, computed as normalized keyword match scores. While this approach cannot capture semantic nuance, it provides reliable signals for explicit references and demonstrates high recall for documents containing direct mentions of telescopes or scientific roles.

Fine-Tuning astroBERT The transformer component leverages `adsabs/astroBERT` (Grezes et al., 2021), a BERT variant pre-trained on 440,000 abstracts from astrophysical literature. The model provides domain-specific vocabulary and contextual understanding of astrophysical language. It is fine-tuned on the provided training data for 3 epochs using a learning rate of 2e-5. The model generates probabilities for three labels: `SCIENCE`, `INSTRUMENTATION`, and `MENTION`, excluding the `NOT_TELESCOPE` class, which is semantically distinct and handled exclusively by the keyword classifier and meta-learner. It outputs a 3-dimensional feature vector.

It reflects our hypothesis that `NOT_TELESCOPE` documents (discussing telescopes in non-primary contexts) require different signals than documents describing telescope roles in primary scientific contexts (see Section 4).

3.2.2 Level-1: Meta-Learner

The Level-1 meta-learner combines base model outputs into a unified classification:

Feature Construction: Outputs from both base models are concatenated into a 7-dimensional feature vector: $\mathbf{x}_{\text{meta}} = [\mathbf{x}_{\text{keyword}}, \mathbf{x}_{\text{astroBERT}}]$ where $\mathbf{x}_{\text{keyword}} \in \mathbb{R}^4$ and $\mathbf{x}_{\text{astroBERT}} \in \mathbb{R}^3$.

Per-Label Meta-Learners: Instead of training a single multi-label classifier, we train four independent XGBoost classifiers M_i (Chen and Guestrin,

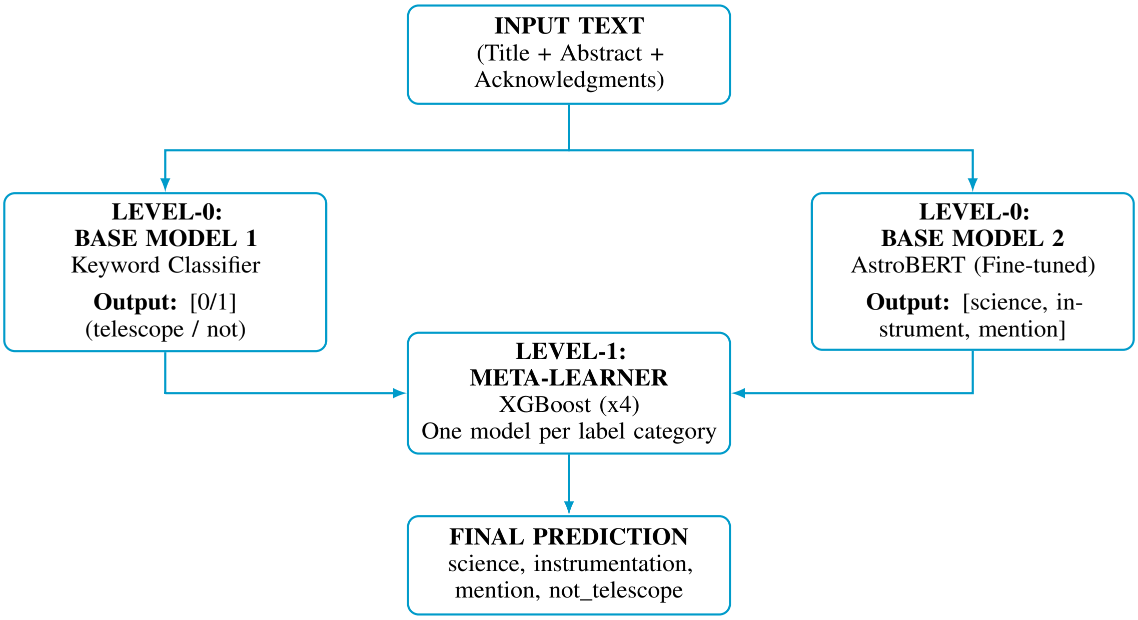
INPUT TEXT
(Title + Abstract + Acknowledgments)
LEVEL-0: BASE MODEL 1
Keyword Classifier
Output: [0/1] (telescope / not)
LEVEL-0: BASE MODEL 2
AstroBERT (Fine-tuned)
Output: [science, instrument, mention]
LEVEL-1: META-LEARNER
XGBoost (x4)
One model per label category
FINAL PREDICTION
science, instrumentation, mention, not_telescope

Figure 1: Two-level stacking ensemble architecture (Task 2). Level-0 base models process input text independently, producing 4 and 3-dimensional vectors, respectively. These are concatenated into a 7-dimensional meta-feature vector, which serves as input for four independent Level-1 meta-learners (one per label), each producing binary predictions.

2016), one per label. This one-vs-rest approach enables:

- independent hyperparameter optimization (particularly `scale_pos_weight`) tailored to each label's unique imbalance ratio.
- isolation of optimization strategies: SMOTE and calibration are applied only to models requiring them.
- flexibility to apply different decision thresholds for different labels.

Each meta-learner M_i produces a binary probability $p_i \in [0, 1]$ for label i, which is converted to a binary prediction using a label-specific threshold τ_i (see Section 5).

4 Subtask 1: Telescope Identification

4.1 Model 1: Stacked LSTM Network

Our initial approach used a stacked Long Short-Term Memory (LSTM) network to exploit the sequential structure of text in the title, abstract, and author fields. **Input:** Tokenized title, abstract, and author fields. **Architecture:** Two stacked LSTM layers (64 units each), followed by a Dense softmax classification layer. **Output:** Multi-class probabilities over four telescope classes.

4.2 Model 2: Domain-Specific Transformer (AstroBERT)

We transitioned to a more powerful, domain-adapted language model: `astroBERT`, pre-trained on astrophysics literature and fine-tuned it with a classification head on TRACS data. Deep language understanding alone was insufficient. Semantic signals were not strong enough to capture the presence or absence of telescope mentions.

4.3 Model 3: Hybrid (Logistic Regression + AstroBERT)

To better isolate the difficult `None` class, we decoupled its prediction into a binary subtask. First, a Logistic Regression model predicted whether a sample belonged to the `None` class. If not, astroBERT classified it into `CHANDRA`, `HST`, or `JWST`.

4.4 Model 4: Feature-Based Random Forest

We shifted focus from textual models to structured metadata features using a RandomForest classifier. The engineered features included field-specific keyword counts, publication year, and author-based patterns.

4.5 Model 5: Final Hybrid (RandomForest + Rule-Based Heuristic)

A comparative analysis of the previous models confirmed that the feature-engineered RandomForest was the most promising direction. However, a deep dive into its confusion matrix revealed a critical performance bottleneck: the vast majority of classification errors occurred because the model was consistently confusing two specific categories. To address this targeted issue, we sought a deterministic feature that could serve as a tie-breaker. We discovered a decisive cue in the bibcode field, where the suffix (the text after the last underscore) deterministically aligned with the true class label. This insight was used to create a rule-based override specifically for instances where the model was likely to be confused.

The final hybrid approach began with the output of the Random Forest model and then applied a rule-based correction to address its specific, known weakness. If the model's prediction was one of the two commonly confused fields, the system applied the rule-based override by extracting the final token from the bibcode. For all other predictions, the model's original output was trusted.

5 Subtask 2: Telescope Role Classification

We employed an iterative development methodology for Task 2, beginning with a baseline model and systematically addressing performance bottlenecks related to class imbalance.

5.1 Baseline System

Our baseline model employed standard stacking without specialized handling for imbalanced data. It used default XGBoost hyperparameters (scale_pos_weight=1, max_depth=6, learning_rate=0.1), a fixed decision threshold of 0.5 for all labels, and implemented no specific data augmentation or class weighting strategies.

This baseline achieved a Macro F1-score of **0.6191**, with severe degradation on minority classes (Table 1). The INSTRUMENTATION class achieved only 0.510 F1, while NOT_TELESCOPE reached 0.480 F1.

5.2 Optimization Strategies

To improve upon the baseline, we implemented five complementary techniques targeting different aspects of model training and prediction on imbalanced data. These strategies are summarized in Table 2.

Label	Precision	Recall	F1	Support
INSTRUMENTATION	0.650	0.420	0.510	132
MENTION	0.700	0.750	0.722	892
NOT_TELESCOPE	0.580	0.410	0.480	187
SCIENCE	0.780	0.750	0.765	2156
Macro Avg	0.678	0.582	0.619	—

Table 1: Baseline performance before optimization. Class imbalance creates severe bottlenecks, particularly for INSTRUMENTATION (1:91 ratio) and NOT_TELESCOPE (1:9 ratio).

Justification for Selective SMOTE Application We applied SMOTE exclusively to the INSTRUMENTATION meta-learner due to its extreme imbalance (1:91). Synthesizing data was deemed necessary to provide sufficient signal for the model to learn this rare class effectively. For the NOT_TELESCOPE class, with a more moderate imbalance (1:9), we found that aggressive class weighting (Strategy 3) alone was sufficient to manage the imbalance without the potential noise introduction or overfitting risks associated with synthetic data generation. The majority of classes required neither technique.

Threshold Tuning Procedure The custom decision thresholds (Strategy 4) were determined by performing a manual grid search over the probability outputs generated by the trained meta-learners on a held-out validation set (20% of the training data). For each minority class (INSTRUMENTATION and NOT_TELESCOPE), we evaluated thresholds ranging from 0.1 to 0.9 in steps of 0.01. The threshold that yielded the maximum F1-score on the validation set for that specific label was selected as the optimal threshold for generating final predictions on the test set.

Calibration Timing Probability calibration (Strategy 5) was applied **after** the XGBoost meta-learners were fully trained using the optimized hyperparameters (including aggressive class weights). The 'CalibratedClassifierCV' wrapper from scikit-learn was fitted using Isotonic Regression on the out-of-fold predictions from the same validation set used for threshold tuning. This post-hoc calibration step adjusts the output probabilities of the already trained models before the final optimized thresholds (determined in Strategy 4) are applied. This ensures the thresholds operate on more reliable probability estimates, improving reproducibility.

#	Strategy	Target Label(s)	Mechanism & Rationale
1	AstroBERT Fine-Tuning	All (via base model)	Unfreezing weights and training for 3 epochs adapts embeddings to the specific task, improving feature quality for the meta-learner.
2	SMOTE Oversampling (Chawla et al., 2002)	INSTRUMENTATION	Generates synthetic minority samples (k=5) to balance the training data to 1:1 for the meta-learner, providing more examples for this extremely rare class (1:91 ratio).
3	Aggressive Class Weighting	INSTRUMENTATION, NOT_TELESCOPE	Manually increases XGBoost's scale_pos_weight (180 for INSTR, 15 for NOT_TEL) beyond the theoretical ratio to heavily penalize misclassifications of minority classes, boosting recall.
4	Custom Prediction Thresholds	INSTRUMENTATION, NOT_TELESCOPE	Lowers the decision threshold (0.35 for INSTR, 0.40 for NOT_TEL) from the default 0.5 to optimize the F1-score by improving recall at an acceptable precision cost for imbalanced classes.
5	Probability Calibration	INSTRUMENTATION, NOT_TELESCOPE	Applies Isotonic Regression post-hoc to the meta-learner outputs to make predicted probabilities more reliable, enhancing the effectiveness of custom thresholds.

Table 2: Summary of optimization strategies applied to improve Task 2 performance.

6 Results

This section details the performance of our final systems for both subtasks, comparing final metrics against developmental stages and discussing the implications.

6.1 Subtask 1: Telescope Identification Results

Our iterative development process for Task 1 culminated in a hybrid model combining a feature-based RandomForest classifier with a rule-based heuristic leveraging the bibcode field (Model 5 in Section 4). As summarized in Table 3, this final approach achieved significantly higher performance than models relying solely on semantic or purely feature-based methods.

Model	Approach	Accuracy	F1	Recall
Model 1	Stacked LSTM	78%	75%	77%
Model 2	AstroBERT	79%	76%	78%
Model 3	Logistic Reg. + AstroBERT	82%	80%	81%
Model 4	Feature-based RandomForest	80%	78%	79%
Model 5	RandomForest + Rule-Based	**97%**	**96.8%**	**97.1%**

Table 3: Performance evolution across five model iterations for Subtask 1 (Telescope Identification).

The dramatic improvement from incorporating the rule-based correction underscores the importance of domain-specific structural features, which provided deterministic cues unavailable in the raw text or other metadata. Neural models struggled particularly with the None class, highlighting the limitations of purely semantic approaches for this specific task.

6.2 Subtask 2: Telescope Role Classification Results

Our final optimized stacking ensemble system, detailed in Section 5, achieved a locally validated **Macro F1-score of 0.683** for the Telescope Role Classification task, a notable enhancement from the baseline of 0.6191. This improvement was primarily driven by successfully mitigating the severe class imbalance affecting minority classes. Specifically, for the **INSTRUMENTATION** class, a combination of targeted strategies including SMOTE-based data augmentation, aggressive class weighting in XGBoost (e.g., scale_pos_weight=180), lowered custom decision thresholds (e.g., 0.35), and probability calibration via Isotonic Regression proved highly effective. These techniques collectively forced the model to better recognize the rare class instances by adjusting data representation, learning penalties, and decision boundaries, leading to a substantial increase in its F1-score from 0.510 to **0.782**. Importantly, these optimizations for minority classes maintained stable performance on the majority classes (MENTION and SCIENCE), demonstrating the robustness of our per-class approach.

Label	Precision	Recall	F1-Score	Support
NOT_TELESCOPE	0.788	0.344	0.479	187
MENTION	0.677	0.747	0.710	892
INSTRUMENTATION	0.901	0.690	**0.782**	132
SCIENCE	0.757	0.764	0.760	2156
Macro Avg (Local)	0.781	0.636	**0.683**	3367

Table 4: Final per-class performance for Task 2 based on local evaluation. The Macro F1-score computed locally is 0.683.

Leaderboard Results Our system achieved a combined Macro F1-score of **0.82** on the TRACS @ WASP 2025 competition leaderboard, securing **4th place**. This score represents the weighted combination of Task 1 (telescope identification, 97% accuracy) and Task 2 (role classification). For Task 2 specifically, our local validation achieved a Macro F1 of 0.683 across the four role labels. The per-class improvements (5) reported in the following analysis reflect the impact of our optimization strategies on each label performance during local evaluation. Model Performance and Analysis for both the share task is shown in Appendix A.

7 Discussion

7.1 Ensemble Synergy

Our results validate the complementary nature of symbolic and semantic models. The keyword classifier provides high-recall signals for explicit telescope mentions, excelling when documents contain direct references. Conversely, astroBERT captures nuanced semantic patterns, capturing context-dependent telescope roles. The meta-learner learns to weight these signals appropriately:

- For `INSTRUMENTATION`: Documents often lack explicit instrumentation keywords, making astroBERT's semantic understanding crucial.
- For `SCIENCE`: High keyword density provides strong signals, but astroBERT refinement reduces false positives.

A potential concern is whether relying on specific keywords might lead to overfitting, particularly if the lexicon is highly tuned to the training data. While the domain-adapted astroBERT component provides broader semantic understanding that can mitigate this, its performance might also degrade if future documents use entirely novel terminology not seen during pre-training or fine-tuning. Careful curation and potential expansion of the keyword list would be necessary for optimal generalization.

7.2 Why Per-Label Meta-Learners?

Our choice of four independent XGBoost meta-learners (rather than a single multi-label model) proved critical for handling extreme imbalance. This design enables: **(1) Fine-grained hyperparameter tuning**: Each label can employ `scale_pos_weight` values matched to its specific imbalance ratio. **(2) Selective data augmentation**: SMOTE is applied only to `INSTRUMENTATION`, avoiding artificial data generation for other classes. **(3) Flexible thresholding**: Different labels can employ different decision thresholds based on their precision-recall trade-off characteristics. **(4) Modular optimization**: New strategies can be tested for individual labels without affecting others.

7.3 Ensemble vs. End-to-End Transformers

While transformer models might seem like a simpler alternative, our ensemble approach offers advantages for this task. Firstly, **interpretability** is enhanced; we can analyze the relative contributions of the keyword (symbolic) and astroBERT (semantic) base models, providing insights into why a classification was made. Secondly, the **modularity** allows for easier updates—the keyword lexicon can be expanded or astroBERT replaced without retraining the entire system. Lastly, the per-label meta-learners provide **targeted robustness** against class imbalance, enabling specific, aggressive optimization strategies for minority classes that might be difficult to implement effectively within a single, monolithic transformer architecture.

8 Conclusion

We presented a hybrid stacking ensemble for the TRACS@WASP 2025 shared task on astrophysical document classification. Our system combines rule-based keyword detection with domain-adapted semantic modeling (astroBERT), using four independent XGBoost meta-learners—one per output label—to handle severe class imbalance through per-label optimization. The modular design enables targeted strategies, e.g., SMOTE augmentation, aggressive class weighting, calibrated probabilities, and custom decision thresholds, proving particularly effective for challenging minority classes.

We achieved a macro F1-score of 0.82 on the leaderboard, securing 4th place. The most significant improvements were realized in the extreme-minority classes: the F1-score for `INSTRUMENTATION` dramatically increased from 0.510 to 0.782 (+53.3%), and notable gains were also achieved for the difficult `NOT_TELESCOPE` label, showcasing the system's strength in high-imbalance scenarios without sacrificing majority class performance. We demonstrate that symbolic and neural approaches are complementary—their synergy is essential for specialized, imbalanced scientific corpora.

9 Limitations and Future Work

While our ensemble approach demonstrates strong performance, several key limitations warrant discussion and guide future research directions.

First, **handling long documents** remains a significant challenge. Our current reliance on astroBERT with a 512-token limit necessitates truncating lengthy astrophysical papers, potentially discarding crucial contextual information located later in the text. Future work should explore architectures specifically designed for long sequences, such as hierarchical attention models or transformers like Longformer (Beltagy et al., 2020), to capture document-wide context more effectively.

Second, the system's performance on the NOT_TELESCOPE class plateaued despite targeted optimization efforts. This suggests that the current feature representations derived from the keyword classifier and astroBERT lack sufficient discriminative power for this nuanced category. Addressing this could involve model-centric approaches like incorporating specialized external models or data-centric improvements such as creating finer-grained annotations for partial or non-primary telescope mentions, potentially leveraging weak supervision techniques to augment training data.

Third, **generalization beyond the TRACS dataset**, particularly to unseen telescopes, requires further investigation. Our system is optimized for the specific telescopes present in the training data (CHANDRA, HST, JWST). While astroBERT offers general domain knowledge, the keyword classifier's effectiveness heavily depends on its lexicon. Future efforts must focus on evaluating performance degradation on diverse astronomical corpora and developing robust strategies for rapid lexicon expansion and adaptation to ensure broader applicability.

While other limitations exist, such as the need for more detailed error analysis, addressing these three core areas—long document processing, minority class feature representation, and generalization—offers the most promising avenues for advancing the system's capabilities.

Future work will focus on addressing truncated context handling, building upon the significant gains achieved for the NOT_TELESCOPE class to further enhance its classification accuracy, and improving cross-domain generalization through hierarchical models and long-document transformers. Our framework provides a robust solution for scientific document classification in high-imbalance regimes, with applications extending beyond astrophysics.

Acknowledgments

We thank SVNIT for the resources (HPC-GPU access) provided to carry out our experiments.

We extend our sincere gratitude to Prof. Pruthwik Mishra for his invaluable guidance, constructive feedback, and mentorship throughout the development of this system.

References

Iz Beltagy, Kyle Lo, and Arman Cohan. 2019. SciBERT: A pretrained language model for scientific text. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3615–3620, Hong Kong, China. Association for Computational Linguistics.

Iz Beltagy, Matthew E Peters, and Arman Cohan. 2020. Longformer: The long-document transformer. *arXiv preprint arXiv:2004.05150*.

Nitesh V Chawla, Kevin W Bowyer, Lawrence O Hall, and W Philip Kegelmeyer. 2002. Smote: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, 16:321–357.

Tianqi Chen and Carlos Guestrin. 2016. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, pages 785–794. ACM.

Charles Elkan. 2001. The foundations of cost-sensitive learning. In *Proceedings of the Seventeenth International Joint Conference on Artificial Intelligence*, pages 973–978.

Mikel Galar, Alberto Fernandez, Edurne Barrenechea, Humberto Bustince, and Francisco Herrera. 2011. A review on ensembles for the class imbalance problem: Bagging-, boosting-, and hybrid-based approaches. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, 42(4):463–484.

Mikel Galar, Alberto Fernandez, Edurne Barrenechea, and Francisco Herrera. 2012. Ensemble methods for class imbalance learning. In Haibo He and Yunqian Ma, editors, *Imbalanced Learning: Foundations, Algorithms, and Applications*, pages 61–82. Wiley-IEEE Press.

Felix Grezes, Jennifer Lynn Bartlett, Kelly Lockhart, Alberto Accomazzi, Ethan Seefried, and Tirthankar Ghosal. 2025. Overview of TRACS: the telescope reference and astronomy categorization dataset & shared task. In *Proceedings of the Third Workshop* *for Artificial Intelligence for Scientific Publications*, Online. Association for Computational Linguistics.

Felix Grezes, Sergi Blanco-Cuaresma, Alberto Accomazzi, Michael J. Kurtz, Edwin Henneken, Mary Dempsey, Donna Thompson, Jonathan Luker, and Golnaz Shapurian. 2021. Building astroBERT, a language model for Astronomy & Astrophysics. In *Astronomical Data Analysis Software and Systems (ADASS) XXXI*.

Charmgil Hong, Rumi Ghosh, and Soundar Srinivasan. 2016. Dealing with class imbalance using thresholding. *arXiv preprint arXiv:1607.02705*.

Kaggle. 2025. Tracs @ wasp 2025: Telescope reference and astronomy categorization shared task. `https://www.kaggle.com/competitions/tracs-wasp-2025`. Accessed: 2025.

Azal Ahmad Khan, Omkar Chaudhari, and Rohitash Chandra. 2023. A review of ensemble learning and data augmentation models for class imbalanced problems: combination, implementation and evaluation. *arXiv preprint arXiv:2304.02858*.

Miroslav Kubat and Stan Matwin. 1997. Addressing the curse of imbalanced training sets: One-sided selection. In *Proceedings of the Fourteenth International Conference on Machine Learning*, pages 179–186.

Dong-Gyu Lee and Hwanjo Kim. 2020. Dynamic cost sensitive learning for imbalanced text classification. In *Proceedings of the KIISE Transactions on Computing Practices*, volume 26, pages 211–216.

Hanung Adi Nugroho, Endang Wiji Tias, Budi Widyakusumah, and Indra Waspada. 2023. A stacking ensemble model with smote for improved imbalanced classification on credit data. *TELKOMNIKA (Telecommunication Computing Electronics and Control)*, 21(4):873–881.

SAO/NASA Astrophysics Data System. 2025. The sao/nasa astrophysics data system. `http://ui.adsabs.harvard.edu/`. Accessed: 2025.

Wikipedia contributors. 2025. Observational astronomy — Wikipedia, the free encyclopedia. [Online; accessed 20-October-2025].

Qihua Zou, Jihua Yu, Yu Zhang, and Chang Liu. 2016. Finding the best classification threshold in imbalanced classification. In *2016 IEEE 16th International Conference on Data Mining Workshops (ICDMW)*, pages 217–224. IEEE.

A Model Performance and Analysis

A.1 Task 1 Performance (Telescope Identification)

The performance of the Task 1 Random Forest model was primarily assessed through 5-fold stratified cross-validation during the hyperparameter tuning phase using `RandomizedSearchCV`.

Cross-Validation Results: The optimized model configuration, selected based on the tuning process described in the System Architecture section, achieved a mean cross-validation accuracy of **0.7772**. The standard deviation across the folds was relatively small (around ±0.0004 according to similar runs shown in the notebook context), indicating consistent performance across different data subsets. This high accuracy suggests the model effectively minimizes confusion between the primary telescope classes (`CHANDRA`, `HST`, `JWST`) while mitigating the impact of the extreme imbalance posed by the `NONE` class, largely due to the strong predictive power of the ID-based features.

Feature Importance: As noted previously, feature importance analysis consistently highlighted the overwhelming predictive power of features derived directly from the `Id` string's suffix. This confirms that the rule-based extraction component, integrated as a feature, provides the primary signal for this classification task. Metadata features like `year` and certain TF-IDF terms offered minor contributions.

Final Prediction: The final model, trained implicitly on the full dataset using the best parameters from `RandomizedSearchCV`, was used to generate predictions for the submission file ('final_submission_task1.csv'). While detailed per-class metrics (precision, recall, F1) were not part of the hyperparameter search output, the strong cross-validation accuracy suggests effective classification, heavily driven by the identifier-based features.

A.2 Task 2 Performance Analysis (Document Role Classification)

Per-Class Performance Analysis Table 5 highlights the change in F1-score for each class from the baseline (Table 1) to the final optimized model (Table 4).

Label	**Baseline F1**	**Final F1**	**Improvement (△ F1)**
`NOT_TELESCOPE`	0.480	0.479	−0.001
`MENTION`	0.722	0.710	−0.012
`INSTRUMENTATION`	0.510	**0.782**	**+0.272**
`SCIENCE`	0.765	0.760	−0.005
Macro Avg	0.619	**0.683**	**+0.064**

Table 5: Comparison of F1-scores before and after optimization for Task 2, highlighting the substantial gain for the `INSTRUMENTATION` class.

The most dramatic success was in the **INSTRUMENTATION** class, which saw its F1-score jump from 0.510 to **0.782** (+0.272, a +53.3% relative improvement). This validates our targeted optimization strategy that combines SMOTE, aggressive class weighting (weight=180), a low decision threshold (0.35) and probability calibration. Precision improved to 0.901 while recall increased significantly from 0.420 to 0.690. Conversely, the **NOT_TELESCOPE** class proved resistant to optimization, with its F1 remaining static (0.480 → 0.479). Despite targeted weighting (weight=15) and thresholding (0.40), the model maintained high precision (0.788) but low recall (0.344), suggesting insufficient characteristic discrimination from the base models for this specific class.

The majority classes, **MENTION** and **SCIENCE**, showed minimal F1 change, indicating that optimizations targeting the minority classes did not negatively impact their performance.

Statistical Reliability The presented results are based on a single training run with a fixed random seed for reproducibility. Averaging results over multiple runs with different seeds could provide a more robust estimate of performance variance but was not performed due to computational constraints.