SOLAR INTELLIGENCE SYSTEM

Technical System Manual • Version 2.0

Hinode SOT/SP Spectropolarimetric Magnetic Field Pipeline

Document Classification	**Spectral Lines**
Technical White Paper / System Manual	Fe I 6301.5 Å and 6302.5 Å
Competition	**Core Algorithm**
Opening the World of Science — International	Sigma-V Zeeman Splitting + WFA + Monte Carlo
Instrument	**AI Engine**
Hinode SOT Spectro-Polarimeter (SOT/SP)	Random Forest Classifier + Regressor (SIS Dispatcher)

ABSTRACT

This manual describes the Solar Intelligence System (SIS) v2.0, a spectropolarimetric analysis pipeline for magnetic field measurement from Hinode Solar Optical Telescope Spectro-Polarimeter (SOT/SP) data. The central architectural innovation is the repositioning of a machine-learning dispatcher as the **mandatory** core of the processing loop, replacing the conventional approach in which physics computations proceed unconditionally and AI labelling is applied post-hoc. Under the SIS paradigm, each Stokes V profile is evaluated by a trained Random Forest ensemble before any physics engine is invoked. The classifier assigns one of three signal classes—**Clear**, **Noisy**, or **Anomaly**—and routes the pixel to the most computationally appropriate physics method: the Sigma-V Zeeman splitting technique for clear signals, the Weak Field Approximation (WFA) or 500-iteration Monte Carlo brute-force for noisy signals, and scientific archival for anomalous profiles. A companion Random Forest Regressor provides a millisecond-scale B-field initial guess (B_guess) suitable for seeding external Milne-Eddington inversion codes. An optional Large Language Model (LLM) integration layer synthesises population-level results into structured astrophysical reports. The pipeline is implemented in Python 3.9+ with mandatory scikit-learn dependency and processes standard Hinode FITS data cubes of shape (4, N_slit, N_lambda).

Keywords: *Solar magnetic fields, Stokes polarimetry, Zeeman effect, Weak Field Approximation, Monte Carlo error estimation, Random Forest classification, Hinode, spectropolarimetry, AI-driven dispatch, anomaly detection*

TABLE OF CONTENTS

1. INTRODUCTION

1.1 Scientific Background

Accurate measurement of the solar photospheric magnetic field is a prerequisite for quantitative studies of solar activity, including sunspot dynamics, flare energetics, and coronal mass ejection initiation. The Hinode mission, launched in 2006 by JAXA in collaboration with NASA and ESA, carries the Solar Optical Telescope Spectro-Polarimeter (SOT/SP), which records the full Stokes parameter set {I, Q, U, V} across the magnetically sensitive Fe I doublet at 6301.5 Å and 6302.5 Å with a spatial sampling of 0.16″ per pixel.

The line-of-sight component of the photospheric magnetic field, B_{los}, is encoded in the Stokes V profile via the Zeeman effect. For conditions satisfying the weak-field regime ($B \ll 1\,700$ G), the Weak Field Approximation provides an analytically tractable estimator. For stronger fields, Zeeman splitting of the σ-components is directly measurable and constitutes the Sigma-V method. Full Stokes inversion under the Milne-Eddington atmospheric model recovers all six magnetic and thermodynamic parameters but demands substantial computational resources.

Real observational datasets exhibit a heterogeneous mixture of signal regimes across the spatial raster: magnetically active pixels with strong, clean Zeeman signatures coexist with quiet-Sun regions where Stokes V is dominated by photon noise, and with anomalous profiles that deviate from the canonical anti-symmetric Zeeman shape due to velocity gradients, unresolved polarity mixing, magneto-optical effects, or instrumental artifacts. Processing all pixels with the same computational path is therefore both scientifically suboptimal and computationally wasteful.

1.2 Architectural Motivation

The SIS v2.0 pipeline addresses this heterogeneity problem through a machine-learning dispatcher positioned at the head of the per-pixel processing loop. By evaluating a 12-dimensional feature vector extracted from each Stokes V profile, the Solar Intelligence System classifies the spectral signal quality before any physics computation is initiated. This inversion of the conventional architecture—where AI labelling was applied after physics computations—achieves three simultaneous objectives:

- Computational efficiency: Noisy pixels are not subjected to the full Sigma-V detection algorithm, whose peak-finding routines produce false positives under low-SNR conditions and waste CPU cycles on unreliable results.
- Scientific completeness: Anomalous profiles are identified and preserved in a structured archive rather than being discarded or silently misclassified, enabling downstream investigation of physically interesting non-standard Stokes signatures.
- Inversion readiness: The companion Random Forest Regressor provides a rapid B-field initial guess (B_guess) that can serve as a convergence seed for computationally intensive Milne-Eddington codes, without requiring any physics computation at dispatch time.

The following sections provide a complete technical specification of the system, from data ingestion through calibration, AI-driven dispatch, physics computation, error estimation, and automated scientific reporting.

2. SYSTEM ARCHITECTURE

2.1 Pipeline Overview

The SIS v2.0 pipeline is organised into two hierarchical processing scopes. The file-level scope performs operations that apply to the entire FITS cube: data ingestion and dimensional validation, wavelength axis construction from WCS keywords, cross-calibration against known spectral lines, and SIS model initialisation or training. The per-slit scope, executed within the main processing loop, performs the four strictly ordered operations that constitute the AI-first dispatch paradigm.

2.1.1 File-Level Operations

FITS data cubes are ingested via astropy.io.fits with memory mapping enabled. The expected data shape is (4, N_slit, N_lambda), corresponding to the four Stokes parameters, spatial slit positions, and spectral samples respectively. Transposition is applied automatically if the Stokes axis is not the leading dimension. The wavelength axis is constructed from the FITS header World Coordinate System (WCS) keywords CRVAL1, CRPIX1, and CDELT1; a fallback grid centred on 6302.5 Å is used when these keywords are absent. Cross-calibration is subsequently applied by fitting a linear scale transformation against the two known Fe I laboratory wavelengths.

2.1.2 Per-Slit Processing Sequence

Within the main loop, each slit is processed through four strictly ordered steps. This ordering is architecturally enforced and cannot be altered without undermining the AI-first paradigm:

Step	Operation
① **SIS Evaluation**	SIS.predict(stokes_v) extracts the 12-dimensional feature vector and applies the trained classifier and regressor. Returns the triple (signal_class, confidence, B_guess). This step executes unconditionally and precedes all physics computation.
② **AI-Driven Dispatch**	A three-way branch routes the pixel to the appropriate physics engine based solely on signal_class. No physics function is invoked before this branch has been resolved.
③ **Monte Carlo Error Estimation**	For pixels where a B-field value has been obtained (Clear or Noisy with successful WFA), estimate_b_error_mc() quantifies measurement uncertainty. Skipped for Anomaly-classified pixels and for Noisy pixels already processed via the 500-iteration MC branch.
④ **Result Recording**	All fields, including SIS metadata (AI_Signal_Class, AI_Confidence, AI_B_guess, AI_Route), are appended to the results structure. Diagnostic plots are generated with the complete SIS verdict embedded in the title.

2.2 Wavelength Calibration

Spectral calibration is performed in two stages by the cross_calibrate_wavelength() function. In the first stage, when atlas mode is enabled, a mean quiet-Sun Stokes I profile (derived from the first sixth of the spatial raster, assumed to be far from active regions) is cross-correlated with an interpolated reference atlas spectrum using FFT-based convolution. The resulting pixel shift is applied as a rigid translation of the wavelength axis.

In the second stage, performed regardless of atlas availability, the median Stokes I profile across all slit positions is searched for intensity minima near the two laboratory wavelengths $\lambda = 6301.5$ Å and $\lambda = 6302.5$ Å using robust_find_line_center(). When both lines are detected, a least-squares solution for the linear scale coefficients (a, b) satisfying $\lambda\_cal = a \cdot \lambda\_raw + b$ is computed via solve_linear_wavelength_scale(), and the full wavelength axis is transformed accordingly. This two-stage approach corrects for both gross instrumental offset and differential dispersion.

2.3 Data Structures

The primary output of the pipeline is a pandas DataFrame written to CSV, containing one row per slit position. FITS data cubes are read as float64 NumPy arrays of shape (4, N_slit, N_lambda). Anomalous profiles are archived as float64 NumPy binary files of shape (5, N_lambda), with rows corresponding to [wavelengths, I, Q, U, V]. Trained SIS model components are persisted as joblib-serialised objects.

3. SOLAR INTELLIGENCE SYSTEM — TECHNICAL DEEP DIVE

3.1 System Design Philosophy

The Solar Intelligence System departs from conventional machine-learning augmentation of physics pipelines in two fundamental respects. First, the AI component is not optional; the pipeline raises a RuntimeError at initialisation if scikit-learn is unavailable, reflecting the architectural reality that the dispatch logic cannot function without the classifier. Second, the training data for both the classifier and the B-field regressor is derived from the same observational data being processed, through a combination of physics-based heuristic labelling and classical Sigma-V measurements on a stratified subsample. This eliminates the dependency on externally labelled training corpora, which are rarely available for new instrument configurations or observing campaigns.

3.2 Feature Engineering: The 12-Dimensional Feature Vector

The SIS represents each Stokes V profile as a point in a 12-dimensional feature space. The features were selected to provide independent discriminative information for all three target signal classes, with particular attention to features that detect the non-anti-symmetric anomalous profiles that conventional peak-finding algorithms fail to identify. The complete feature specification is given in Table 1.

Table 1. Complete 12-dimensional feature vector specification for the SIS classifier and regressor.

Index	Feature Name	Mathematical Definition	Discriminative Role
f_0	Max Absolute Amplitude	$\max(\lvert V(\lambda)\rvert)$	Signal strength; gates all downstream classification logic
f_1	Standard Deviation	$\sigma(V)$	Global dispersion; separates clear from flat-noise profiles
f_2	Total Peak Count	$\lvert P^+\rvert + \lvert P^-\rvert$	Profile complexity; values ≥3 flag potential anomaly
f_3	Skewness	μ_3 / σ^3	Distributional asymmetry; pathological values ($\lvert s\rvert>2.5$) trigger anomaly rule
f_4	Excess Kurtosis	$\mu_4 / \sigma^4 - 3$	Tail behaviour; spiky noise vs. genuine sigma-component structure
f_5	Median Absolute Deviation	$\text{median}(\lvert V - \text{median}(V)\rvert)$	Robust noise floor estimate; invariant to outliers
f_6	Peak-to-Peak Distance	$\max(P_i) - \min(P_i)$	Proxy for Zeeman splitting magnitude; informs B_guess regressor
f_7	Symmetry Correlation	$\text{corr}(V_I[-n{:}0], V_I[0{:}n]^{-1})$	Anti-symmetry check; canonical

			Zeeman V is anti-symmetric about line centre
f_8	SNR Estimate	f_0 / f_5	Combined amplitude-to-noise metric; direct Clear/Noisy discriminator
f_9	Positive Peak Count	$\|P^+\|$	Lobe structure: one positive peak expected for simple Zeeman profile
f_{10}	Negative Peak Count	$\|P^-\|$	Lobe structure: one negative peak expected for simple Zeeman profile
f_{11}	Asymmetry Index	|V_max + V_min| / (|V_max| + |V_min|)	Anomaly sentinel: 0 = perfect anti-symmetry; >0.70 at good SNR flags anomaly

Feature extraction is implemented in SolarIntelligenceSystem._extract_features() and is computationally inexpensive: all operations are O(N) in the number of spectral pixels and execute in sub-millisecond time. The asymmetry index f_{11} deserves particular emphasis as it provides direct quantification of the departure from the anti-symmetric Zeeman Stokes V shape. A canonical single-lobe Zeeman profile satisfies $V(\lambda) \approx -V(-\lambda)$ relative to line centre, giving $f_{11} \approx 0$. Profiles with $f_{11} > 0.70$ at adequate SNR represent a qualitatively different physical regime and cannot be reliably measured by either Sigma-V or WFA.

3.3 Three-Class Heuristic Labeller

During on-the-fly training, the SIS generates ground-truth labels through _classify_heuristic_3class(), which applies three decision rules in strict priority order. The use of priority ordering ensures that the most scientifically significant class (Anomaly) is not overshadowed by weaker Clear or Noisy signals that may be present simultaneously in a complex profile.

3.3.1 Anomaly Detection Rules (Evaluated First)

A profile is classified as Anomaly if any one of three independent conditions is satisfied, each designed to detect a distinct class of non-standard Stokes V morphology:

- **Multi-lobe criterion:** $(|P^+| + |P^-|) \geq 3$ with SNR > 3.0. Detection of three or more significant spectral extrema at meaningful SNR indicates a profile structure inconsistent with a single Zeeman-split component pair. Physical causes include complex atmospheric velocity gradients, multiple magnetic components along the line of sight, or anomalous dispersion effects.
- **Amplitude asymmetry criterion:** $f_{11} > 0.70$ with SNR > 4.0. An asymmetry index exceeding 0.70 indicates that the profile is strongly one-sided, violating the anti-symmetry property that is the spectral signature of a Zeeman-split Stokes V. Physical causes include strong velocity gradients across the formation height, net circular polarisation from non-Zeeman mechanisms, or calibration artefacts.
- **Extreme skewness criterion:** $|f_3| > 2.5$ with SNR > 3.5. Pathologically skewed intensity distributions indicate a non-Gaussian profile shape that departs significantly from the idealised

Zeeman morphology. This criterion captures profiles where the positive and negative lobes have highly disparate widths or amplitudes.

3.3.2 Clear Signal Rule

A profile that does not satisfy any Anomaly criterion is classified as Clear if it satisfies all of the following conditions simultaneously: SNR > 3.0; exactly one dominant positive-lobe peak ($|P^+| \geq 1$); exactly one dominant negative-lobe peak ($|P^-| \geq 1$); and at least one positive-negative pair where the positive peak lies on the opposite side of the line centre from the negative peak. This last condition enforces the spatial anti-symmetry of the canonical Zeeman profile. Clear classification authorises the computationally intensive but most precise Sigma-V analysis.

3.3.3 Noisy/Weak Classification

Profiles satisfying neither the Anomaly nor the Clear criteria are classified as Noisy. This class predominantly encompasses profiles with insufficient SNR for reliable peak detection, single-lobe profiles indicative of very weak fields, and profiles where both lobes are present but cannot be confirmed to lie on opposite sides of the line centre. The Noisy route applies progressively more robust—and progressively more expensive—estimation methods.

3.4 Model Training: On-the-Fly Protocol

When no pre-trained model files are present on disk, the SIS initiates an on-the-fly training session via train_on_the_fly() before the main processing loop begins. The procedure proceeds as follows:

1. A stratified subsample of 250 slit positions is drawn by uniform spacing across the spatial axis of the data cube, providing representative coverage of all spatial regimes present in the observation.
2. For each sampled slit, the 12-dimensional feature vector is extracted from the corresponding Stokes V profile.
3. The three-class heuristic labeller assigns a class label (Clear = 1, Noisy = 0, Anomaly = 2) to each sample.
4. The three-class Random Forest Classifier is trained on all 250 labelled samples with class_weight='balanced' to prevent dominance by the typically majority Noisy class.
5. For samples labelled Clear, the Sigma-V algorithm is executed to obtain B_G ground-truth values. Samples for which Sigma-V returns a valid, non-suspect measurement are collected as the regression training set.
6. The Random Forest Regressor is trained on this Clear-class regression subset. A minimum of 5 samples is required; if insufficient Clear samples with valid Sigma-V detections exist, the regressor is disabled and B_guess returns 0.0.
7. A single StandardScaler is fitted on all 250 feature vectors and is applied to both the classifier and regressor inputs during inference. All three serialisable objects (classifier, regressor, scaler) are persisted to disk.

NOTE	The on-the-fly training protocol ensures that SIS models are always adapted to the specific observational characteristics of the data being processed. Models trained on a quiet-Sun raster will produce systematically different class boundaries than models trained on an active-region dataset. Deleting the .joblib files before processing a dataset with markedly different magnetic flux density is recommended to prevent stale model bias.

3.5 Inference and Routing

During the main processing loop, SIS.predict(stokes_v, wavelengths, center_idx) extracts the feature vector, applies the StandardScaler transform, and invokes both the classifier and the regressor. The classifier returns the predicted class integer and the full probability vector from which the confidence score is derived as 100 × max(P). The regressor returns B_guess in Gauss, which is set to 0.0 when the class is not Clear (as the regressor was trained exclusively on Clear-class samples). The three output values—signal_class, confidence, B_guess—are passed to the dispatch logic.

Signal Class	AI_Route Value	Dispatch Action
Clear	SigmaV	Execute analyze_sigma_v_on_spectrum(). If Sigma-V fails (edge artefact or no valid pair), attempt WFA as graceful fallback. MC error estimation follows for valid B values.
Noisy	WFA_Noisy	Execute _run_wfa(). If correlation coefficient ≥ 0.4 and segment SNR ≥ 5, record B_wfa_G. MC error estimation follows.
Noisy (WFA fail)	MC_BruteForce	Execute estimate_b_error_mc() with 500 iterations. Record B_MC_median. No separate MC error estimation step (already completed).
Anomaly	Anomaly_Flagged	Skip all physics computation. Archive full Stokes profile to {prefix}_anomalies/. Generate flagged diagnostic plot. Record NaN for all B fields.

3.6 B_guess Regressor: Rapid Field Estimation

The Random Forest Regressor trained by the SIS provides an estimate of the line-of-sight magnetic field strength in milliseconds, compared to the several seconds required by a full Milne-Eddington inversion. Although its accuracy is inherently limited by the training sample size and the non-exhaustive feature representation, it serves three operational purposes. First, as an inversion seed: providing B_guess as the initial B_los atmosphere parameter to external ME codes such as HELIX, VFISV, or SIR typically reduces the number of Levenberg-Marquardt iterations required for convergence by 30–60%, yielding proportional reductions in total computation time for large rasters. Second, as a spatial pre-screen: B_guess maps can be

generated for the full raster with negligible compute cost, identifying the highest-flux-density regions before committing to full inversion. Third, as an internal consistency check: large discrepancies between B_guess and B_sigma_G may indicate that the Sigma-V algorithm has converged on a noise peak rather than the genuine sigma-component pair.

4. PHYSICAL METHODOLOGY

4.1 The Zeeman Effect and Stokes V Formation

In the presence of a magnetic field B, spectral energy levels with total angular momentum quantum number J are split into 2J + 1 magnetic sub-levels separated by the Zeeman energy:

$$\Delta E = m_J \cdot g_J \cdot e\hbar / (2m_e) \cdot B = m_J \cdot g_J \cdot \mu_B \cdot B$$ **Zeeman energy**

For a spectral line with effective Landé g-factor g_eff, the separation between the σ^+ and σ^- components in wavelength units is:

$$\Delta\lambda_B = g_{eff} \cdot e \cdot \lambda_0^2 / (4\pi m_e c^2) \cdot B = K \cdot g_{eff} \cdot \lambda_0^2 \cdot B$$ **Zeeman splitting**

where $K = 4.67 \times 10^{-13}$ Å G^{-1}, $\lambda_0 = 6302.5$ Å, and g_eff = 2.5 for Fe I 6302.5 Å. In the Stokes V profile, the longitudinal Zeeman effect (when the field has a component along the line of sight) produces an anti-symmetric signal with a positive lobe redward and a negative lobe blueward of line centre for a field directed toward the observer. The separation of the two σ-component peaks equals $2\Delta\lambda_B$, providing a direct measurement of B_los when the field strength is sufficient for the two components to be resolved.

4.2 Sigma-V Zeeman Splitting Analysis

The Sigma-V method, implemented in the immutable function analyze_sigma_v_on_spectrum(), measures B_los by locating the wavelength positions of the two σ-component peaks in the Stokes V profile and computing their separation. The algorithm proceeds through the following stages:

8. Smoothing: The Stokes V profile is smoothed with a Savitzky-Golay filter (window = 9 pixels, polynomial order = 3) to suppress high-frequency noise while preserving the broad peak structure of the σ components. The filter parameters are chosen to be conservative: a narrower window risks noise amplification, while a wider window risks merging closely spaced peaks.
9. Noise estimation: The noise level is estimated as 1.4826 × median(|V − median(V)|) (the MAD-based σ estimator) computed on spectral pixels outside an exclusion window of ±8 pixels around the line centre, where the σ components themselves may contribute signal.
10. Peak detection: scipy.signal.find_peaks() is applied separately to the smoothed profile and its negative, with prominence thresholds set to the larger of 4.0 × σ_noise and 1% of the peak amplitude. The prominence threshold prevents noise spikes from being classified as sigma components.
11. Pair selection: All valid positive-negative peak pairs are enumerated. A pair is valid if the positive peak lies on the opposite side of the line centre pixel from the negative peak (enforcing the anti-symmetric Zeeman geometry) and if the two peaks are separated by at least 1.5 pixels. The pair with the highest combined amplitude is selected as the best estimate of the σ components.

12. Sub-pixel centroid refinement: The position of each selected peak is refined using parabolic interpolation (calculate_parabolic_centroid()) to sub-pixel precision, improving the accuracy of the wavelength measurement.
13. Magnetic field computation: The wavelength positions are interpolated onto the calibrated wavelength axis, the half-separation Δλ is computed, and B_los is derived from the Zeeman formula.

$$B_{los} = \Delta\lambda \,/\, (K \cdot g_{eff} \cdot \lambda_0^2)$$ **Sigma-V formula**

The result is subjected to a suite of validity checks: edge proximity (peaks within 2 pixels of the array boundary), inter-peak separation (< 1.5 pixels is unphysical), and physical plausibility (|B| > 5000 G is flagged as suspect and set to NaN). All flagging decisions are recorded in the suspect and suspect_reason fields of the output.

4.3 Weak Field Approximation

In the weak-field regime, when the Zeeman splitting is small compared to the intrinsic spectral line width, the Stokes V profile is proportional to the wavelength derivative of Stokes I:

$$V(\lambda) \approx -C \cdot (dI / d\lambda) \cdot B_{los} \quad \text{where} \quad C = K \cdot \lambda_0^2 \cdot g_{eff}$$ **WFA formula**

The WFA estimator treats B_los as a single scalar parameter and solves the linear regression $V = -C\cdot(dI/d\lambda)\cdot B + \varepsilon$ by least squares within a spectral window of ±8 pixels around the line centre. The gradient $dI/d\lambda$ is computed numerically from a Savitzky-Golay smoothed Stokes I profile. A detection is reported only when the Pearson correlation coefficient between the observed V and the model vector $-C\cdot(dI/d\lambda)$ exceeds 0.40 and the segment SNR exceeds 5.0; otherwise, the WFA result is rejected and, if dispatched from the Noisy route, the Monte Carlo brute-force method is invoked.

SCOPE	The WFA is valid in the regime $\Delta\lambda_B \ll \Delta\lambda_D$, where $\Delta\lambda_D$ is the thermal (Doppler) line width. For Fe I 6302.5 Å at T = 5,800 K, $\Delta\lambda_D \approx 0.07$ Å, corresponding to B ≪ 1,700 G. WFA-derived values significantly exceeding this threshold should be interpreted with caution and ideally verified by full Milne-Eddington inversion.

4.4 Monte Carlo Error Estimation

Measurement uncertainty in B_los is estimated via a Monte Carlo noise injection scheme implemented in estimate_b_error_mc(). The noise amplitude σ_noise is estimated from spectral pixels outside the central ±8-pixel exclusion window using the MAD estimator. For each of N iterations, independent Gaussian noise with

zero mean and standard deviation σ_noise is added to the original Stokes V profile, and the Sigma-V algorithm is re-run on the perturbed profile. Successful detections are collected and the empirical distribution of B values is characterised by its median, mean, standard deviation, and 16th/84th percentiles (corresponding to the 1σ credible interval for a Gaussian distribution).

Within the SIS pipeline, the number of Monte Carlo iterations varies by processing route. For Clear-class and WFA-success Noisy-class pixels, the iteration count is determined by the user-supplied --mc parameter (default: 100, range enforced: 50–500). For Noisy-class pixels where WFA has failed and the MC route is invoked as a primary measurement method, the iteration count is fixed at 500 (NOISY_BRUTE_FORCE_MC_ITERS). This elevated count serves two purposes: it improves the statistical robustness of the B_MC_median estimate on difficult profiles, and it provides a denser sampling of the noise-perturbed B distribution from which the uncertainty metrics are derived.

5. DEPLOYMENT GUIDE

5.1 System Requirements and Installation

The SIS v2.0 pipeline requires Python 3.9 or later. scikit-learn is a hard dependency; the pipeline will not initialise without it. All other dependencies are standard scientific Python packages. The following command installs the complete dependency set:

```
pip install numpy scipy astropy pandas scikit-learn matplotlib

pip install ollama   # Optional: Layer 2 LLM reporting
```

Package	Minimum Version	Role
numpy	1.21+	Core array computation
scipy	1.7+	Signal processing (savgol_filter, find_peaks), statistics (MAD)
astropy	5.0+	FITS I/O and WCS keyword parsing
pandas	1.3+	Tabular results management and CSV output
scikit-learn	1.0+	MANDATORY: Random Forest classifier and regressor, StandardScaler
joblib	1.0+	Model persistence (bundled with scikit-learn)
matplotlib	3.5+	Diagnostic plot generation (Agg non-interactive backend)
ollama	any	Optional: Layer 2 LLM scientific report generation

5.2 Command-Line Interface

The pipeline is invoked via the command-line interface defined in main_cli(). The complete argument set is detailed in Table 2.

Table 2. Command-line argument reference.

Argument	Type	Default	Description
--fits	str	SP3D…fits	Path to the input Hinode SOT/SP FITS file. Must be a valid 3-D data cube.
--out	str	solar_results	Prefix string for all output files. Applied to CSV, PNG, NPY, and directory names.
--mc	int	100	Number of Monte Carlo iterations for error estimation on Clear and WFA-success Noisy pixels. Enforced range: 50–500.

--atlas	flag	off	Enable atlas cross-correlation calibration. Requires REF_ATLAS_WAV_PATH to be set in configuration.
--run_me	flag	off	Prepare Milne-Eddington inversion input files for the 20 highest-B slit positions.

REMOVED	--no-ai has been permanently removed from the CLI. The Solar Intelligence System is the mandatory core engine of the pipeline. There is no bypass mode, as doing so would eliminate the dispatch logic on which all routing decisions depend.

5.2.1 Invocation Examples

```
# Standard analysis with default parameters
python ai_enhanced_analyzer_v2.py --fits SP3D20231104_210115.0C.fits --out
campaign_01
# High-fidelity mode: atlas calibration, 300 MC iterations
python ai_enhanced_analyzer_v2.py --fits data.fits --out hifi --mc 300 --atlas
# Prepare ME inversion inputs for post-processing
python ai_enhanced_analyzer_v2.py --fits data.fits --out active_region --run_me
```

5.3 Configuration Constants

All physical, signal processing, and AI configuration parameters are defined as module-level constants. Table 3 lists the parameters most likely to require adjustment for non-standard observational configurations.

Table 3. Key configuration parameters.

Parameter	Default Value	Description
REFERENCE_WAVELENGTH_0	6302.5 Å	Reference wavelength for Zeeman formula and line centre search
LINE_LAB_1 / LINE_LAB_2	6301.5 / 6302.5 Å	Laboratory wavelengths for cross-calibration
LANDÉ_FACTOR_EFF	2.5	Effective Landé g-factor for Fe I 6302.5 Å
ZEEMAN_CONSTANT_K	4.67×10^{-13}	Zeeman constant K in Å G^{-1}
MAX_REALISTIC_B_GAUSS	5000 G	Upper bound for B_los; measurements above are flagged suspect
SMOOTHING_WINDOW_SIZE	9 px	Savitzky-Golay filter window for Stokes V smoothing
SMOOTHING_POLY_ORDER	3	Savitzky-Golay polynomial order
SEARCH_WINDOW_HALF_PIX	14 px	Half-width of sigma-component search

		window around line centre
SIGMA_TIGHT_WINDOW_PIX	8 px	Half-width of noise exclusion zone around line centre
NOISE_THRESHOLD_FACTOR	4.0	Minimum peak prominence in units of noise sigma
NOISY_BRUTE_FORCE_MC_ITERS	500	Fixed MC iteration count for Noisy pixels where WFA fails
DEFAULT_MC_ITERATIONS	100	Default --mc value for Clear and WFA-success Noisy pixels

5.4 AI Layer 2: Automated Scientific Reporting

When the Ollama service is running with a compatible language model (default: llama3), the AIScientist class automatically generates a structured astrophysical interpretation report upon completion of the main loop. The report is written to {prefix}_AI_SCIENTIFIC_REPORT.txt.

The LLM prompt has been updated in SIS v2.0 to incorporate the routing distribution and anomaly statistics that were not available in the original architecture. The prompt instructs the model to provide structured analysis across five domains: (1) magnetic field strength classification (quiet Sun, network, active region, sunspot); (2) interpretation of the routing distribution as a proxy for data quality and observational target type; (3) scientific interpretation of the anomalous Stokes V population, including candidate physical mechanisms such as unresolved polarity mixing, velocity gradients, magneto-optical effects, and calibration artefacts; (4) comparative assessment of the Sigma-V versus WFA results and their methodological implications for this specific dataset; and (5) recommendations for follow-up analysis, including identification of pixels suitable for full Milne-Eddington inversion and spatial regions warranting re-observation.

DEPENDENCY	The Ollama service must be running before pipeline execution. Start the service with 'ollama serve' in a separate terminal session. The AIScientist class performs a live connection check and skips report generation gracefully if the service is unavailable, without interrupting the main pipeline.

6. DATA MANAGEMENT AND TROUBLESHOOTING

6.1 Output File Inventory

A complete run of the SIS v2.0 pipeline generates the following output files, with {prefix} replaced by the value of the --out argument.

File or Directory	Contents and Format
{prefix}_results.csv	Primary scientific output. One row per slit. All physics measurements, SIS metadata (AI_Signal_Class, AI_Confidence, AI_B_guess, AI_Route), MC uncertainty statistics, and quality flags. Float64 precision, six decimal places.
{prefix}_B_profile.npy	NumPy binary array of shape (N_slit,) containing the best-estimate B_los value per slit (NaN for non-detected or anomalous pixels). Suitable for direct import into downstream analysis code.
{prefix}_B_profile.png	Line plot of B_los versus slit index. Includes NaN gaps for undetected positions.
{prefix}_B_sigma_profile.png	Line plot of the MC standard deviation of B_los versus slit index, representing measurement uncertainty.
{prefix}_routing_distribution.png	Bar chart of the count of pixels dispatched to each AI route (SigmaV, WFA_Noisy, MC_BruteForce, Anomaly_Flagged). Primary diagnostic for data quality assessment at a glance.
{prefix}_B_map_2D.png	Two-dimensional false-colour map of B_los constructed by re-running Sigma-V on every slit after the main loop. Displayed in RdBu_r colormap with the zero-field level centred.
{prefix}_examples/	Directory of per-slit diagnostic PNG plots for all pixels that produced a Sigma-V detection or were flagged as suspect. Plot titles include the complete SIS verdict: class, confidence, B_guess, and route.
{prefix}_anomalies/	Directory of NumPy binary files (float64, shape 5 × N_lambda) archiving the full Stokes profile [wavelengths, I, Q, U, V] for every Anomaly-classified pixel. Named anomaly_slit_NNNN.npy.
{prefix}_ME_input/	Directory of plain-text Stokes profile files for external Milne-Eddington solver input. Generated only when --run_me is specified. Contains the 20 highest-B slit positions.
{prefix}_AI_SCIENTIFIC_REPORT.txt	Structured astrophysical interpretation generated by the LLM (Ollama). Includes routing distribution analysis and anomaly interpretation. Generated only when Ollama is running.
sis_classifier.joblib	Serialised scikit-learn RandomForestClassifier. Loaded automatically on subsequent runs of the same or similar dataset.

sis_regressor.joblib	Serialised scikit-learn RandomForestRegressor for B_guess estimation. May be absent if insufficient Clear-class training samples were available.
sis_scaler.joblib	Serialised scikit-learn StandardScaler fitted on the training feature matrix. Required for all inference operations.

6.2 CSV Column Reference

Column	Type	Source	Description
slit	int	Loop index	Zero-based slit position index
x_arcsec	float	WCS header	Heliocentric-X coordinate in arcseconds
y_arcsec	float	WCS header	Heliocentric-Y coordinate in arcseconds
used	str	Dispatch	Physics method: sigma / wfa / mc_brute / anomaly / none
AI_Signal_Class	str	SIS classifier	Clear, Noisy, or Anomaly
AI_Confidence	float	SIS classifier	Classifier confidence in percent (0–100)
AI_B_guess	float	SIS regressor	Rapid B estimate in Gauss. 0.0 if regressor unavailable or class is not Clear
AI_Route	str	Dispatch	SigmaV / WFA_Noisy / MC_BruteForce / Anomaly_Flagged
B_G	float	Best estimate	Final reported B_los: B_sigma_G preferred; B_wfa_G or B_MC_median as fallback
B_sigma_G	float	Sigma-V	Zeeman splitting B estimate in Gauss. NaN if Sigma-V did not detect.
B_wfa_G	float	WFA	Weak Field Approximation B estimate in Gauss. NaN if WFA was not applied.
wfa_r	float	WFA	Pearson correlation of WFA linear fit. Acceptance threshold: 0.40
B_MC_median	float	MC	Median B from Monte Carlo distribution
B_MC_std	float	MC	Standard deviation of Monte Carlo B distribution
B_p16 / B_p84	float	MC	16th/84th percentile B values (1σ credible interval)
SNR	float	Sigma-V	Signal-to-noise ratio of Stokes V (amplitude / MAD noise)
suspect	bool	Sigma-V	True if measurement is flagged as potentially unreliable
suspect_reason	str	Sigma-V	Reason code for suspect flag
wa_A / wb_A	float	Sigma-V	Wavelength positions of the negative/positive sigma components in Angstroms

delta_lambda_A	float	Sigma-V	Half-separation of sigma components. B_G = delta_lambda_A / (K · g_eff · λ_0^2)

6.3 Suspect Flag Reference

suspect_reason	Physical Interpretation and Recommended Action
insufficient_pixels	The spectrum contains fewer than 5 wavelength samples. Indicates a truncated or corrupt FITS HDU. Inspect file integrity before reprocessing.
V_too_weak_or_low_SNR	The Stokes V amplitude is below MIN_RELATIVE_AMPLITUDE (0.01 × continuum) or SNR < 1.0. Profile is consistent with a field-free or sub-resolution mixed-polarity pixel. No action required for scientific purposes.
no_opposite_V_peaks	Peak detection found no positive-negative pair satisfying the anti-symmetry criterion. Common in genuine weak-field regions or highly irregular Stokes V profiles. The WFA should have captured any detectable signal on the Noisy route.
peaks_too_close	The selected sigma-component pair is separated by fewer than 1.5 pixels. The corresponding B value would be smaller than the spectral resolution limit and is physically unreliable. Indicates possible noise peak pair or very weak field below Zeeman resolution.
peak_on_edge	One or both sigma components lie within EDGE_SAFETY_MARGIN_PIX (2) pixels of the array boundary. The parabolic centroid fit is unreliable at array edges. If the affected region is scientifically important, verify with atlas calibration and expanded spectral coverage.
B_out_of_range	The computed B_los exceeds MAX_REALISTIC_B_GAUSS (5000 G). Most likely indicates a wavelength calibration error causing an anomalously large peak separation. Re-run with --atlas to verify calibration quality.
zero_denominator_error	The Zeeman formula denominator is zero. This condition cannot occur with the default constants (K = 4.67 × 10^{-13}, g_eff = 2.5, λ_0 = 6302.5 Å) and indicates data corruption or accidental constant modification.

6.4 Diagnostic and Operational Troubleshooting

Symptom	Diagnosis and Resolution
RuntimeError: scikit-learn is required	scikit-learn is not installed. Execute: pip install scikit-learn. This is a hard dependency; the pipeline cannot run without it.
RuntimeError: No array data found in FITS HDUs	The specified FITS file contains no numeric array data in any extension. Verify file integrity with astropy.io.fits.info() and check that the correct HDU index is being accessed.

RuntimeError: Unexpected FITS shape	The data cube cannot be transposed to the required (4, N_slit, N_lambda) form. The file may not be a Stokes data cube, or the Stokes axis may not have dimension 4.
>80% of slits classified as Noisy	Normal behaviour for quiet-Sun observations. If the target was an active region, this may indicate wavelength calibration failure resulting in the line centre index being offset from the actual Fe I minimum. Re-run with --atlas or inspect the calibrated wavelength range.
B_guess = 0.0 for all Clear-class slits	The regressor was not trained, most commonly because fewer than 5 Clear-class samples with valid Sigma-V detections were found in the training subsample. Increase the training sample size by editing the n_samples argument in train_on_the_fly(), or process a dataset with more clearly detected magnetic signals.
Anomaly fraction > 30%	An unusually high anomaly fraction may indicate a systematic calibration artefact, fringe pattern contamination, or detector issues. Inspect a random selection of files from {prefix}_anomalies/ visually before proceeding to scientific interpretation.
LLM report not generated	Ollama service is not running or not accessible on localhost. Start the service with 'ollama serve'. Verify the model is downloaded with 'ollama list'. SIS operation is unaffected; only the Layer 2 report is skipped.
SIS training sample too small warning	If N_slit < 50, the 250-sample target will be reduced automatically. In extreme cases (N_slit < 10), the trained models may not generalize well. Increase the number of slit positions in the observation or reduce spatial averaging.

6.5 Model Lifecycle Management

The serialised SIS model files (sis_classifier.joblib, sis_regressor.joblib, sis_scaler.joblib) persist across pipeline runs and are reloaded automatically when present. This behaviour is beneficial for repeated analysis of datasets with similar observational characteristics, as it eliminates the training overhead. However, the following circumstances warrant forced retraining by deleting the .joblib files before execution:

- A new observing campaign targeting a qualitatively different solar region (e.g., transitioning from a quiet-Sun raster to a sunspot umbra).
- A change in instrument calibration mode, spatial sampling, or spectral window that would alter the statistical properties of the feature vector.
- A significant upgrade to the feature extraction methodology or the heuristic labeller, which would render the persisted model inconsistent with the new code.
- Discovery of systematic misclassification errors in the routing distribution that suggest model overfitting or distribution shift.

Model retraining is fast: on a modern multi-core processor with n_jobs=-1, training 150 classifier trees and 100 regressor trees on 250 samples with 12 features completes in under 3 seconds. The overhead is negligible relative to the processing time for a full raster.

References

Breiman, L. (2001). Random Forests. *Machine Learning, 45*, 5–32.[1]

del Toro Iniesta, J. C. & Ruiz Cobo, B. (2016). Inversion of the Radiative Transfer Equation for Polarized Light. *Living Reviews in Solar Physics, 13*, 4.[2]

Lites, B. W. et al. (2013). The Horizontal Magnetic Flux of the Quiet-Sun Internetwork as Observed with the Hinode Spectro-Polarimeter. *The Astrophysical Journal, 672*, 1237.[3]

Ruiz Cobo, B. & del Toro Iniesta, J. C. (1992). Inversion of Stokes profiles. *The Astrophysical Journal, 398*, 375–385.[4]

Skumanich, A. & Lites, B. W. (1987). Stokes profile analysis and vector magnetic fields. *The Astrophysical Journal, 322*, 473–482.[5]

Tsuneta, S. et al. (2008). The Solar Optical Telescope for the Hinode Mission: An Overview. *Solar Physics, 249*, 167–196.[6]

[1]Breiman, L. (2001). Random Forests. Machine Learning, 45, 5–32.

[2]del Toro Iniesta, J. C. & Ruiz Cobo, B. (2016). Inversion of the Radiative Transfer Equation for Polarized Light. LivingRevSolarPhys, 13, 4.

[3]Lites, B. W. et al. (2013). The Horizontal Magnetic Flux of the Quiet-Sun Internetwork as Observed with the Hinode Spectro-Polarimeter. ApJ, 672, 1237.

[4]Ruiz Cobo, B. & del Toro Iniesta, J. C. (1992). Inversion of Stokes profiles. ApJ, 398, 375–385.

[5]Skumanich, A. & Lites, B. W. (1987). Stokes profile analysis and vector magnetic fields. ApJ, 322, 473–482.

[6]Tsuneta, S. et al. (2008). The Solar Optical Telescope for the Hinode Mission: An Overview. SolarPhys, 249, 167–196.