Peer-Review

Pinto, Aaron, Juliana Caulkins, and Anita Shaw. 2025. "A Machine Learning Approach to Enhance the Classification of Dark Matter Signals." *Journal of High School Science* 9 (4): 192–208. https://doi.org/10.64336/001c.147389

The paper requires significant revisions before it can be considered for publication. The following issues should be addressed:

Use of simulated data:

The study relies exclusively on toy simulation data from XENON1T rather than full detector simulations and calibration data. While toy simulations are useful for method development and feasibility studies, they are not necessarily representative of real WIMP data. Machine learning models trained only on toy WIMP/background events risk learning non-physical features, leading to poor generalization, false positives, or missed signals. This limitation makes performance metrics such as F1 score less meaningful in terms of physical interpretation. Although the author acknowledges this in the limitations section, further effort should be made to validate the approach with more realistic simulation or calibration data.

Insufficient references:

The manuscript lacks sufficient citations to relevant prior work. The author should include references to existing literature, both on WIMP detection strategies and on machine learning applications in this domain.

XENON1T context:

In the XENON1T experiment, the ratio of scintillation to ionization is already used to distinguish nuclear recoils (expected from WIMPs) from electronic recoils (backgrounds). The paper should clarify why additional machine learning methods are needed. Is the signal-to-noise ratio so low that conventional techniques are insufficient?

Application of ML methods:

The current version of the paper emphasizes comparing machine learning methods rather than demonstrating their actual utility for WIMP identification. The application of ML should be motivated by clear physics goals, not just methodological exploration.

Choice of models:

The paper should provide stronger justification for the choice of models. For instance, if XGBoost produces comparable results to ensembles, why not rely on the simpler model?

Novelty:

Since other studies have already explored the use of ML in this context, the author should clarify what is novel about the present contribution.

Background events:

The focus on radiogenic neutrons as the sole background is a major limitation. WIMP searches must contend with multiple background sources, and limiting the study to a single one reduces its practical impact.

Theoretical and experimental context:

The paper would benefit from a short background on WIMPs, noting that none have yet been detected, and a discussion of what kind of signals are expected from WIMP interactions. Based on this, the author should simulate events in the presence of realistic background noise to test whether ML models can truly distinguish between signal and background.

Research background score:

The justification for a research background score of 0.8 is unclear and should be elaborated.

“Use of simulated data:
The study relies exclusively on toy simulation data from XENON1T rather than full detector simulations and calibration data. While toy simulations are useful for method development and feasibility studies, they are not necessarily representative of real WIMP data. Machine learning models trained only on toy WIMP/background events risk learning non-physical features, leading to poor generalization, false positives, or missed signals. This limitation makes performance metrics such as F1 score less meaningful in terms of physical interpretation. Although the author acknowledges this in the limitations section, further effort should be made to validate the approach with more realistic simulation or calibration data.”

Addressed in:
- Abstract — “The dataset … is generated based on the XENON1T findings.”
- 3.1 Data generation — cloning/using Blueice/Laidbax and SearchForDarkMatter; explicit validation strategy (cross-backend comparison, folding geometry, light-collection maps, electron lifetime).
- 3.3 Algorithms / calibration — calibration diagnostics, CalibratedClassifierCV, reliability plots and Brier scores.
- 4 Results — cross-backend performance and domain-shift quantification (FastSim↔Blueice numbers; Figure 3).
- 5.3.1 Limitations — explicit statement of the simulation-to-reality gap and suggestion for more backends / domain-adaptation.
- 6 Conclusion — “measured simulation-to-reality gap of roughly 4–5% F1” and “further calibration … necessary.”

“Insufficient references:
The manuscript lacks sufficient citations to relevant prior work. The author should include references to existing literature, both on WIMP detection strategies and on machine learning applications in this domain.”

Addressed in:
- Inline citations throughout Introduction (e.g., (2,3,6,13,16,18,24) etc.).
- 3.1 Data generation — cites XENON1T/Blueice/Laidbax (2,3,19).
- Intro ML discussion / novelty paragraph — cites prior ML work (5,12,14,20,23).
- References section (7.) — full bibliography listing the cited works (Akerib et al., Aprile et al., Balázs et al., Guest et al., Renner et al., Priel et al., etc.).

“XENON1T context:
In the XENON1T experiment, the ratio of scintillation to ionization is already used to distinguish nuclear recoils (expected from WIMPs) from electronic recoils (backgrounds). The paper should clarify why additional machine learning methods are needed. Is the signal-to-noise ratio so low that conventional techniques are insufficient?”

Addressed in:
- Introduction — explicit explanation: “The statistical methods of the XENON1T experiment rely on … S1/S2 … However, this method performs well only for bulk background rejection and struggles with specific cases.” (follows with examples: neutron-induced NRs, surface-background distortions).
- 3.1 / 3.2 — use of S1, S2, corrected S1/S2 and engineered features (log_s1, log_energy) shows ML is built on those observables.
- 5.1 Discussion & Conclusion — argues ML complements traditional methods and targets subtle event classes (neutrons, surface effects).

“Application of ML methods:
The current version of the paper emphasizes comparing machine learning methods rather than demonstrating their actual utility for WIMP identification. The application of ML should be motivated by clear physics goals, not just methodological exploration.”

Addressed in:
- Introduction — motivation that ML can capture multi-dimensional correlations and identify subtle patterns missed by cut-based S1/S2.
- 2.1 Hypothesis — physics-centered goals and performance baselines tied to utility (robust event classification, controlled degradation under systematics).
- 3.4 Ensemble & 4 Results — operating points, per-class metrics, and calibrated probabilities reported (practical outputs for event selection).
- 5.1 Discussion & 6 Conclusion — interprets results in physics terms (signal recovery, detector-sensitivity identification, probability-based event selection).

“Choice of models:
The paper should provide stronger justification for the choice of models. For instance, if XGBoost produces comparable results to ensembles, why not rely on the simpler model?”

Addressed in:
- Introduction (ML paragraphs) — role of Random Forest (feature ranking/interpretability), XGBoost (nonlinear discrimination, sample efficiency), SVM (margin-based bias) and rationale for combining.
- 3.3 Algorithms — hyperparameter grids and intended roles (RF as diagnostic; XGBoost for high performance; SVM for complementary bias).
- 3.4 Ensemble — soft-voting, comparison to hard-voting and stacking described.
- 4 Results & 5.2 Comparison — concrete performance comparison (Table 1, Figure 5) and statement that XGBoost is best single model but ensemble yields top aggregate metrics.

“Novelty:
Since other studies have already explored the use of ML in this context, the author should clarify what is novel about the present contribution.”

Addressed in:
- Introduction — explicit novelty paragraph — “novel in its multi-observable focus and interpretability pipeline” with two concrete claims: (1) analyses across four background types vs typical ER/NR binary focus, (2) emphasis on Random Forest feature ranking + SHAP to mitigate simulation-specific artifacts.
- 3.1–3.4 and Conclusion — repeated emphasis on transferability, cross-backend tests, and interpretability pipeline supporting the novelty claim.

“Background events:
The focus on radiogenic neutrons as the sole background is a major limitation. WIMP searches must contend with multiple background sources, and limiting the study to a single one reduces its practical impact.”

Addressed in:
- Abstract — lists cosmogenic neutrons, surface events, electronic recoils, radiogenic neutrons.
- Introduction — explains each background type and why they mimic WIMPs.
- 3.1 Data generation — explicit event list and dataset composition (50k WIMP, 50k background; 12.5k each background type).

- 4 Results — 5×5 confusion matrix and per-class metrics (Table 2) showing treatment of multiple backgrounds.
- 5.3.2 Class-Specific Weaknesses — notes surface events are hardest and proposes future remedies.

“Theoretical and experimental context:
The paper would benefit from a short background on WIMPs, noting that none have yet been detected, and a discussion of what kind of signals are expected from WIMP interactions. Based on this, the author should simulate events in the presence of realistic background noise to test whether ML models can truly distinguish between signal and background.”

Addressed in:
- Introduction (opening paragraphs) — SUSY/MSSM, neutralino as candidate, relic abundance argument, reasons WIMPs are hard to detect (citations: Jungman et al., Martin, etc.).
- Introduction (TPC paragraph) — explanation of TPC detection, S1/S2 signals, and how nuclear recoils relate to WIMP ID.
- 3.2 Preprocessing / feature list — lists physically meaningful observables used for classification (ties to expected signals).

“Research background score:
The justification for a research background score of 0.8 is unclear and should be elaborated.”

Addressed in:
- 2.1 Hypothesis — references past research to form baselines “obtained from previous related research … metrics within the range of 0.80 to 0.85.”
- 4 Results / 5.1 Discussion — shows achieved metrics (~0.843 F1, 0.847 accuracy) relative to that baseline.

The reviewer thanks the author for the time and effort invested in addressing the previous comments. However, several important issues remain that need to be resolved before the manuscript can be considered for publication.

1. Terminology clarification:
The manuscript refers to both Supersymmetry and the Standard Model using the abbreviation “SM.” This is incorrect and potentially confusing. Supersymmetry should be abbreviated as SUSY, while the Standard Model should remain SM.

2. Simulation–data consistency:
The current analysis does not account for potential differences between the simulated detector and real experimental detectors. In real detectors, event detection probability varies with spatial position and energy. The simulation should therefore be reweighted to reproduce the energy and position distributions observed in experimental data (e.g., XENON1T).
Without such corrections, the machine-learning model may show artificially strong performance on synthetic data but fail to generalize to real detector conditions.

3. Feature correlation and multicollinearity:
Several input features appear to be mathematically dependent (for example, S2/S1 is directly derived from S1 and S2). This introduces high correlation among variables.
While tree-based models such as Random Forest and XGBoost are robust to multicollinearity, SVMs are sensitive to correlated inputs, which may bias results. It is recommended that the author examine feature correlations (e.g., via a correlation matrix or Variance Inflation Factor analysis) and document how such dependencies were handled.

4. Reproducibility and data-handling transparency:
To avoid data leakage and ensure fair evaluation, please clarify the calibration procedure — specifically, when and on which subsets (training/validation/test) the CalibratedClassifierCV or

equivalent methods were applied. Calibration should be performed strictly on training/validation data, not on the final test set.
In addition, please include reproducibility details such as random seed, Python version, library versions, and hardware configuration to enable others to replicate the reported results.

"Terminology clarification: … Supersymmetry should be abbreviated as SUSY"
- Modifications: Abstract; Introduction.
- Change: Supersymmetry labeled "Supersymmetry (SUSY)", Standard Model kept as "Standard Model (SM)".

- "Reweight to reproduce energy and position distributions …"
  - Modifications: Section 3.1 (Data generation); Section 4 (Results).
  - Change: Added 2D energy–position reweighting (50×50 bins, Gaussian smoothing), explained how sample_weight/resampling were applied during training and reported improved cross-backend F1 (FastSim→Blueice 0.801→0.822; Blueice→FastSim 0.794→0.818).
- "Examine feature correlations (e.g., S2/S1) and document handling (VIF, PCA, etc.)"
  - Modifications: Section 3.2 (Preprocessing).
  - Change: Added Pearson/VIF analysis and documented two SVM mitigations
- "Clarify calibration procedure and provide reproducibility details (seeds, versions, hardware)"
  - Modifications: Section 3.3 (Algorithms); Section 3.1 (Data generation).
  - Change: Documented specifications and materials in more detail

---

The reviewer thanks the author for addressing all of the reviewer's concerns. I recommend this paper for publication.
