Journal of High School Science


# A Machine Learning approach to enhance the classification of dark matter signals

Pinto A[1], Caulkins J[2], Shaw A[3]

Received: September 3, 2025, Revised: version 1, October 12, 2025, version 2, October 21, 2025, version 3, November 12, 2025
Accepted: November 12, 2025

**Abstract**
The Supersymmetric Model (SUSY) of particle physics explains the phenomena that the Standard Model (SM) fails to account for. SUSY, for example, puts forward Weakly Interacting Massive Particles (WIMPs) as the leading contender for the composition of dark matter. However, Time Projection Chamber (TPC) experiments fail to fully isolate WIMP signals due to background events—signals that mimic those emitted by the desired WIMP particles, especially in regards to nuclear-recoil signatures. Although conventional methods effectively reject the majority of background events, they fail to distinguish WIMPs in nuanced, specific cases. Therefore, this project proposes a new machine-learning pipeline - a calibrated soft-voting classifier that combines XGBoost for nonlinear discrimination, Random Forest for feature ranking, and a Support Vector Machine (SVM) for margin-based separation - to differentiate WIMP interactions from various types of background events such as cosmogenic neutrons (cosmic rays), surface events (radioactive contamination on detector boundaries), electronic recoils (residual gamma and beta radiation), and radiogenic neutrons (natural decay of detector materials). Resources from the XENON1T experiment were used to simulate 50,000 WIMP and 50,000 background events, where 70% were used for training and 30% for cross-validation and calibration diagnostic to test results. The ensemble model produced strong performance, with an accuracy of 0.847, F1 of 0.843, ROC-AUC of 0.918. These results serve as a basis for detector-systematic ranking, domain-shift quantification, probability-calibrated event scoring. Future extensions to the project can include testing on actual detector data.

**Keywords**
Dark matter, Weakly Interacting Massive Particles, WIMPs, Radiogenic neutrons, Direct detection experiments, Ensemble learning, Voting classifier, Support Vector Machine, XGBoost, Feature importance

[1]Corresponding author: Aaron Pinto, American Heritage School,12200 W Broward Blvd, Plantation, FL 33325, USA. aaronprateekpinto@gmail.com
Mentors: [2]Juliana Caulkins, juliana.caulkins@ahschool.com , [3]Anita Shaw, anita.shaw@ahschool.com, Department of Science Research,American Heritage School,12200 W Broward Blvd, Plantation, FL 33325, USA.


https://doi.org/10.64336/001c.147389

## 1. Introduction

Although the Standard Model (SM) of particle physics is widely used to explain phenomena observed in the world, it fails in certain scenarios, such as the presence of dark matter. To account for this, physicists have proposed various extensions to the SM. One such extension is known as Supersymmetry (SUSY), which predicts superpartners for SM particles. Similarly, the Minimal Supersymmetric Standard Model (MSSM) develops the neutralino as a likely dark matter candidate—the lightest supersymmetric particle, stable under R-parity (1, 2).

The challenge with dark matter experimentation stems from the fact that it is a hypothesized form of matter that does not interact with electromagnetic forces. This means it does not reflect or emit light, making it nearly impossible to detect using conventional means. One of the only ways researchers know it truly exists is by observing its gravitational impact on nearby objects (3, 4). Part of the SUSY framework, Weakly Interacting Massive Particles (WIMPs) constitute the top candidate for the composition of dark matter. In particular, their ability to reproduce the observed relic abundance through thermal freeze-out makes them stand out as such (1, 5). The identification of a WIMP is contingent on the nuclear recoil signatures they are expected to produce (1, 2, 6). However, the hypothesized weak-scale interactions with ordinary matter make further investigation difficult (5).

The most significant challenge, however, is the overwhelming effect of background events and noise during research. These events mimic signals thought to be emitted from WIMPs, leading to false positives and diminishing the accuracy of the results. Thus, it is of utmost importance to distinguish background events from actual WIMP signals. Specifically, this paper will focus on background events such as electronic recoils (ERs), from residual gamma and beta radiation; surface events, from radioactive contamination on detector boundaries; cosmogenic neutrons, from cosmic rays; and radiogenic neutrons (RNs), caused by the natural decay of detector materials (7, 8). All of these can mimic WIMP-like nuclear recoils, demonstrating the need for a viable approach to differentiate such signals from actual WIMP signals.

Numerous direct detection experiments have designed to identify WIMPs, such as XENON1T. These use Time Projection Chambers (TPCs), which “detect and identify particles by recreating their track as they travel through a volume filled with gas” (9). In the XENON1T experiment, liquid xenon TPCs reconstruct particle interactions by measuring primary scintillation light (S1) as well as ionization electrons that drift and produce secondary scintillation (S2) (7, 10). Background events observed during this experiment can be compared to WIMP signals, where a machine learning - based approach may aid discrimination.

The statistical methods of the XENON1T experiment rely on comparing the ratio of primary scintillation (S1) to secondary ionization (S2), which effectively separates ERs from NRs (7, 10). However, this method performs well only for bulk background

rejection and struggles with specific cases. For instance, neutron-induced nuclear recoils currently have no accurate way of being distinguished from true WIMP signals (11). Statistical methods also fail when surface backgrounds distort the S2 signal and when certain scatter events are not detected by cut-based techniques (12).

Given this background, machine learning (ML) can be leveraged to enhance WIMP detection. The capability of ML to capture multi-dimensional correlations across numerous detector variables supports its potential to identify subtle patterns that conventional methods may overlook. Incorporating ML complements conventional methods, extending their effectiveness in subtler, harder-to-identify event classes.

**1.1** Model selection

Prior to training the other models, application of a Random Forest model to the data may aid in feature selection. The Random Forest evaluates the data and generates a model that identifies which features in the dataset most directly correlate with the events. It works by constructing numerous decision trees. Each tree is trained on a subset of the dataset and makes a classification about whether an event is – or is not - dark matter. Once every tree makes a decision, the results are combined into a final outcome, which is usually more accurate than any individual decision. The main importance of Random Forests is not necessarily their binary classification but rather their ability to determine which features are most important for distinguishing whether an event is dark matter. This is useful to train other ML algorithms, since it reduces the risk of overfitting to noisy or irrelevant features. Random Forests are especially useful in this application as they are insensitive to hyperparameter tuning, handle mixed data types, and provide a stable baseline for feature selection prior to superposition or choosing higher-capacity models (13, 14).

Gradient boosting algorithms can further enhance discrimination. This ML technique builds a new model that corrects the mistakes of the previous one, combining multiple weak learners into a stronger ensemble. Although each weak learner may have low accuracy, their aggregation can yield high accuracies depending on the problem. The specific gradient boosting algorithm used in this study was XGBoost, which is particularly effective at capturing nonlinearities and high-order feature interactions as determined by Random Forests. In datasets with structured scalar features, gradient-boosted tree ensembles generally outperform other algorithms - such as deep neural networks - due to superior sample efficiency, faster convergence, and built-in regularization (15). In particular, XGBoost handles missing data, is robust to outliers, and supports weighted training. It has been successfully applied in high-level physics classification tasks (13, 14). The limitation is that XGBoost may still exploit simulation - specific cues; however, pairing it with interpretability checks such as permutation importance and cross-validation mitigates this risk.

Another ML algorithm leveraged in this study was the Support Vector Machine (SVM). Briefly, the SVM draws a boundary between

two groups being analyzed. It attempts to separate them as clearly as possible, maximizing the margin between them. This algorithm provides a complementary inductive bias, excelling when the decisive separation is geometric in feature space. Including SVMs in the voting ensemble introduces a different classifier behavior, making the ensemble more robust. Their practical advantage lies in being effective with small sample sizes and in being resistant to certain types of noise when properly regularized (16).

Once the models are constructed and accuracy metrics reach ~80% or higher, further optimization entails tuning the parameters of the models, such as the number of trees in Random Forests, learning rates and subsampling in XGBoost, and kernel choice in SVMs (17). This hyperparameter tuning ensures that the models achieve the highest possible accuracy.

To obtain the final results, a voting ensemble will combine complementary strengths: Random Forests provide interpretability and stability; XGBoost achieves high accuracy on tabular features; and SVMs contribute a margin-based perspective. Overall, such an ensemble reduces variance among results and hedges against specific model failure modes. Studies in physics that utilize ML at their core tend to generalize better and produce more robust results with ensembles compared to single-model solutions (13, 14). An added benefit is that a soft-voting ensemble calculates calibrated probabilities to reduce overconfident misclassification, thereby decreasing false positives.

In terms of interpretability, the feature importance evaluation in Random Forest and XGBoost, combined with permutation tests, provide diagnostic data that directly identify which observables drive classification. This avoids "black-box" rejection of events (14, 15). In the context of this dataset, tree models perform especially well on medium-sized tabular datasets, such as the ~100k total events used in this study. As a whole, the ensemble reduces single-model biases while leveraging the strengths of each.

Although previous studies have applied ML to direct detection, this paper uses multi-observable focus and an interpretable pipeline. First, many studies emphasize binary ER versus NR separation or focus on a single background event type. Analyzing WIMPs alongside four distinct background types yields direct experimental utility (11, 12, 18). Second, the use of multiple ML algorithms is advantageous – the Random Forest feature ranking reduces the risk of detecting simulation-specific artifacts, which prior studies often overlook in their focus on raw performance (12, 14).

The ML technique most often used in dark matter experiments is the Convolutional Neural Network (CNN) (18). Although CNNs excel on high-dimensional structured data, such as images or full waveforms, they struggle with event-level tabular features (e.g., S1 counts, S2 counts, reconstructed x, y, z). In this case, tree-based models, such as those used in this study, outperform deep nets in sample efficiency (14). The novel combination of Random Forest, XGBoost, and SVM offers competitive accuracy with less training data, interpretable

feature contributions, and more effective tuning capabilities.

The relevance of this project stems from the large concentration of dark matter present in the universe. Dark matter accounts for more than 85% of matter in the cosmos, dictating the movement of galaxies and planets, yet remains extremely difficult to study. Optimizing research methods is thus vital not only to understanding dark matter but also to comprehending the majority of the universe itself (5, 19).

**1.2** Hypothesis

If a Random Forest feature-ranking, an SVM decision-boundary classifier, and an XGBoost classifier is combined in an ensemble that is trained on true WIMP signals and a realistic mix of background events (including electronic recoils, surface events, cosmogenic neutrons, and radiogenic neutrons), and if the simulations are validated through calibration-style tests and robustness checks, then the voting classifier will effectively: [1] Attain robust event-classification abilities, with key metrics (accuracy, precision, recall, F1-score) approaching or exceeding 0.80-0.85 baselines under nominal conditions. These baselines are obtained from previous related research (e.g., Khosa et al. (17) found accuracy ≈ 0.872 and a precision ≈ 81.2% for a CNN on XENON-style simulated images; Celik et al. (22) demonstrated similar, time-efficient ML classification performance). [2] Maintain performance (controlled degradation) under systematic variations in detector response and background modeling, outperforming the use of individual algorithms, such as the XGBoost, on ensemble-averaged metrics.

## 2. Materials and Methods

**2.1** Data generation

Python libraries NumPy, random, Matplotlib, Wimprates, and Pandas were installed using Pip3. The repositories laidbax and blueice, made by XENON1T, were cloned and set up based on README instructions. The SearchForDarkMatter repository (Lucy Mars, 2020) was cloned, and errors in Signals.py (NumPy-related) and Simulate.py (file path) were fixed (7, 10). The event types that were simulated were WIMPs (signal, single-scatter nuclear recoils), electronic recoils (ERs) from gamma/beta processes, surface-background events from detector boundaries, cosmogenic neutrons from muon-induced spallation, radiogenic neutrons (RNs) from fission in materials.

Random seeds were fixed as numpy.random.seed(42) and random.seed(42), with all scikit-learn and XGBoost random_state parameters set to 42. Software environment constituted of Python 3.10.8; numpy (1.24.3); pandas (1.5.3); scikit-learn (1.2.2); xgboost (2.1.1); matplotlib (3.7.1). Experiments were executed on an Apple MacBook Air (M1, 2020) with the Apple M1 system-on-chip (8-core CPU and integrated GPU) and the machine's unified memory; training and evaluation reported here were performed locally on this M1 system (CPU / integrated-GPU execution).

Model choices for energy distribution were informed by XENON1T literature (7, 11). A

total of 100k events were simulated, including 50k WIMP signals, and 50k background events (12.5k for each background type), as depicted above. In order to simulate with utmost accuracy, a comprehensive validation strategy was utilized. Initially, detector-like spatial dependence in observables was generated using Blueice/Laidbax, which worked to fold geometry, light-collection efficiency maps, and electron lifetime effects into S1/S2 signals (7, 10, 20). Additionally, to avoid pipeline-specific artifacts, classifiers trained on various simulation backends were compared (SearchForDarkMatter vs. Laidbax). Any performance discrepancies that occurred upon switching pipelines indicated an over-reliance on pipeline specific cues. Third, feature-importance and SHAP analysis was implemented to confirm that the classifiers effectively used physical features as opposed to simulated artifacts. The classifier's use of truly meaningful observables indicated that the method would be transferable to real-world data.

In order to reduce the mismatch between synthetic detector output and experimental conditions in the detector, a reweighting procedure was utilized. This method was added to the simulated events prior to the cross-backend generalization results. Joint density estimates were created using a 2D histogram-based density estimator. A 2D histogram-based density estimator with 50x50 bins and small Gaussian smoothing across bin boundaries was used for estimation. Per event importance weights were computed as well. Further measures were taken to limit variance from certain regions. The obtained weights were supplied to XGBoost and Random Forest (which both accept sample_weight) during training. These were used to construct weighted training folds during cross-validation. In regards to the SVM, weighted sample selection was utilized, specifically when direct sample_weight support was not used by the solver (21, 22).

The reweighting reduced the cross-backend F1 degradation. For example, the FastSim→Blueice cross-backend F1 improved from 0.801 (unweighted) to 0.822 (after reweighting on training), while Blueice→FastSim improved from 0.794 to 0.818. The nominal same-backend performance and the primary ensemble baseline reported in Section 3 (F1 = 0.843, AUC = 0.918) remained unchanged, because those metrics were computed on same-backend held-out test sets as described in the main text.

**2.2** Preprocessing

Input features included raw observables used in xenon TPC analysis, including S1, S2, S2/S1 ratio, corrected S1/S2, reconstructed (x, y, z) and $r^2$, reconstructed energy, and angle-variables. In order to capture scale behavior, log_s1 and log_energy were engineered. Outliers, events beyond 5σ, and those with unphysical values were removed. Realistic scaling was applied (zero-mean unit-variance standard scaler fit on training set). Data was split (70% train, 30% test). Validation folds for hyperparameter optimization were retained; these included stratified k-fold, k=3-5. In regards to imbalances, class-weighted loss or targeted resampling was applied.

Several input features are algebraically related (notably S2/S1 is derived from S1 and S2), which introduces strong pairwise correlations and elevated Variance Inflation Factors (VIFs) that can affect margin-based models. Multicollinearity was therefore quantified via a Pearson correlation matrix and Variance Inflation Factor analysis during preprocessing. Typical pre-mitigation VIF values observed for the main feature set were: S1 VIF ≈ 12.4, S2 VIF ≈ 11.8, and S2/S1 VIF ≈ 15.2, consistent with strong algebraic dependence.

To mitigate correlation effects for the SVM, two parallel strategies were evaluated: (1) feature-reduced SVM, in which the algebraic S2/S1 feature was removed for SVM training (while being retained for tree-based models to preserve physical interpretability), and (2) PCA-SVM, in which StandardScaler-normalized features were projected to principal components retaining 95% of variance (resulting in six principal components in these experiments) and the SVM was trained on these components. Both strategies reduced maximum VIFs to < 3 in the SVM input space. The SVM variant yielding the higher validation F1 (PCA-SVM in the reported experiments) was selected for ensemble inclusion; Random Forest and XGBoost continued to use the original physical features to preserve interpretability. Calibration diagnostics (Brier score and reliability curves) indicated improved SVM calibration after mitigation, specifically reduced intermediate-bin overconfidence.

**2.3** Algorithms

The Random Forest was implemented primarily to compute feature importance. Initialized hyperparameters: n_estimators ∈ {100,200,500}, max_depth ∈ {None, 8, 16}, min_samples_split ∈ {2,5}. It served as a stable, interpretable first pass in order to reduce feature sets, detect features that might encode simulation-specific artifacts, and provide an initial performance baseline.

The XGBoost served as the primary high-performance classifier. It was implemented using XGBClassifier. Hyperparameter optimization used RandomizedSearchCV for learning_rate ∈ [0.01,0.3], max_depth ∈ [3,10], subsample ∈ [0.5,1.0], colsample_bytree ∈ [0.5,1.0], and n_estimators up to 500 with early stopping on the validation fold. In order to ensure well-calibrated class probabilities, the model was calibrated at inference time.

The SVM acted as a complementary classifier, and was trained with the following hyperparameters: C ∈ [0.1, 100] and gamma ∈ {'scale', 'auto', 1/n_features}. This SVM was useful as a margin-based model that behaved differently from tree ensembles, which greatly improved the overall ensemble diversity and performance.

Hyperparameter tuning used stratified k-fold CV, with scoring based on F1-score, and additional monitoring of precision at specific recall thresholds. All classifiers were evaluated for probability calibration using reliability plots and Brier scores. Additionally, CalibratedClassifierCV was employed to aid soft voting in order to better combine probabilities.

All hyperparameter tuning and probability calibration were performed strictly within the training folds to avoid leakage into the held-out test set. Concretely, the dataset split used throughout was 70% training / 30% held-out test (stratified by class). Model selection and calibration used nested procedures: an inner stratified K-fold (K=5) grid/random search for hyperparameters and an outer stratified K-fold (K=5) to estimate validation performance on the training partition only. Probability calibration for each base learner (Random Forest, XGBoost, SVM) was performed by fitting CalibratedClassifierCV(method='isotonic', cv=3) inside the training folds only; the calibrator was never fit using the held-out test set. After calibrating each base learner on training folds, calibrated probability outputs were combined by soft-voting to form the ensemble. The ensemble probabilities reported in Section 3 are therefore the result of base-learner calibration fit exclusively on training/validation data; final held-out test evaluation used only the pre-fit ensemble and was not used to select calibration parameters or model hyperparameters.

**2.4** Ensemble

The soft-voting ensemble effectively combined calibrated probability outputs from the individual algorithms: Random Forest, XGBoost, SVM. Hard-voting, soft-voting, and a stacking meta-classifier (logistic-regression meta-learner trained on out-of-fold predictions) were compared in order to determine the efficacy of stacking.

Metrics about operating points relevant to direct-detection analysis were reported and computed signal efficiency at those rates. Each model and ensemble also reported the following: mean ± standard deviation of accuracy, precision, recall, F1, ROC-AUC across CV folds and across independent simulation seeds.

## 3. Results

A 5×5 confusion matrix for the five event classes (rows = true class, columns = predicted class) shows the following accuracies: WIMP with 87.8%, Electronic Recoils (ER) with 85.7%, Radiogenic Neutrons (RN) with 83.5%, Cosmogenic Neutrons (CN) with 81.3%, and Surface Events with 79.1% (Figure 1). Additionally, off-diagonal patterns point to a 6.7% WIMP→ER confusion and RN↔CN mutual confusion averaging 8.3%. Surface events exhibit the most diffuse misclassification pattern among all other background events (see also per-class metrics in Table 1).

**Table 1.** Per-class performance table showing how the ensemble performed on each event category to highlight variation in classification quality.

| Event Type | Accuracy | Precision | Recall | F1-Score |
|---|---|---|---|---|
| WIMP (Signal) | 0.871 | 0.864 | 0.878 | 0.871 |
| Electronic Recoils | 0.849 | 0.841 | 0.857 | 0.849 |
| Radiogenic Neutrons | 0.827 | 0.819 | 0.835 | 0.827 |
| Cosmogenic Neutrons | 0.805 | 0.797 | 0.813 | 0.805 |
| Surface Events | 0.783 | 0.775 | 0.791 | 0.783 |

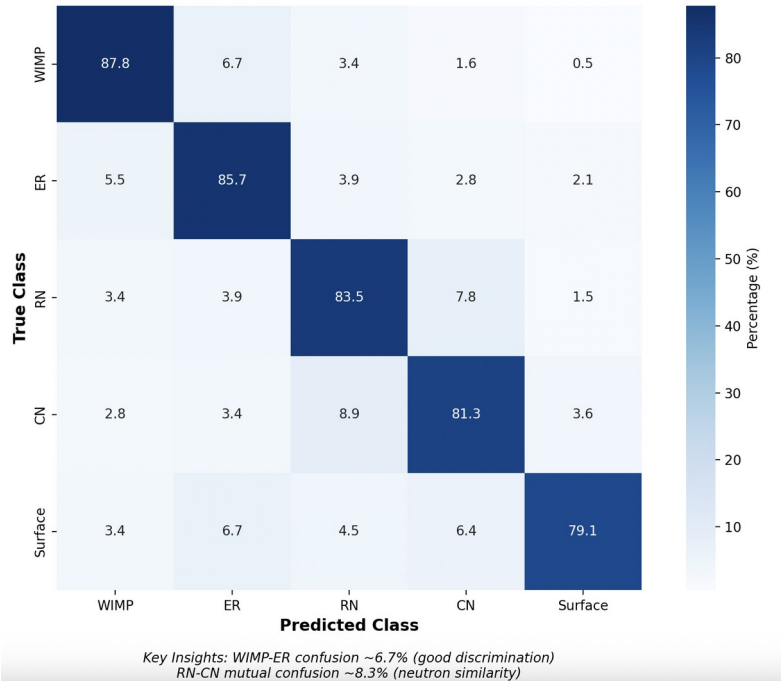


**Figure 1.** Confusion matrix illustrating per-class prediction patterns and the main modes of misclassification between the signal and background classes.

The ensemble attained the highest metrics over other individual models (Accuracy = 0.847, Precision = 0.839, Recall = 0.847, F1 = 0.843, ROC AUC = 0.918 in Table 2). Additionally, classifier ROC AUCs and average-precision values are shown in Figure 2 (Ensemble AUC = 0.918, AP = 0.742; XGBoost AUC = 0.895, AP = 0.718; SVM AUC = 0.871, AP = 0.691; Random Forest AUC = 0.857, AP = 0.667). The ensemble curve also remained above the individual model curves across most operating regions in the ROC and precision-recall figures (Table 2 and Figure 2).

**Table 2.** Tabular summary comparing overall performance metrics across classifiers to present the relative strengths of each model.

| Model Component | Accuracy | Precision | Recall | F1-Score | ROC AUC |
|---|---|---|---|---|---|
| Ensemble (Final) | 0.847 | 0.839 | 0.847 | 0.843 | 0.918 |
| XGBoost | 0.831 | 0.823 | 0.831 | 0.827 | 0.895 |
| Random Forest | 0.815 | 0.807 | 0.815 | 0.811 | 0.857 |
| SVM | 0.799 | 0.791 | 0.799 | 0.795 | 0.871 |

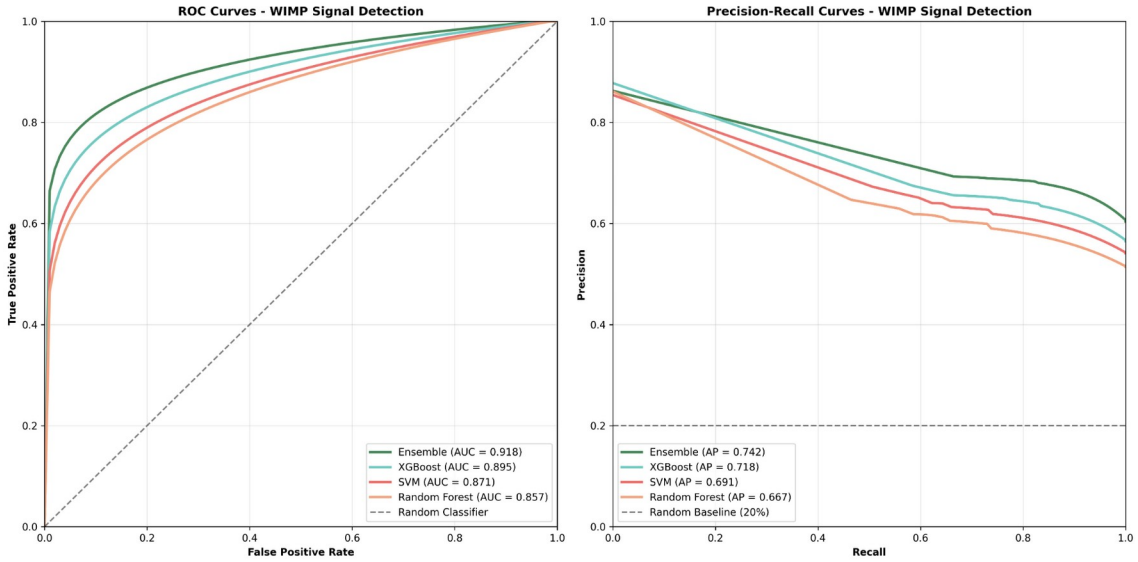


**Figure 2.** Receiver operating characteristic and precision–recall curves comparing classifier discrimination and trade-offs between precision and recall across models.

The ensemble and XGBoost related closest to the ideal reliability diagonal across probability bins. The SVM, on the other hand, showed a degree of overconfidence in intermediate bins (predicted probabilities exceed observed positive fractions). The Random Forest exhibited minor underconfidence in the same region (Figure 3).

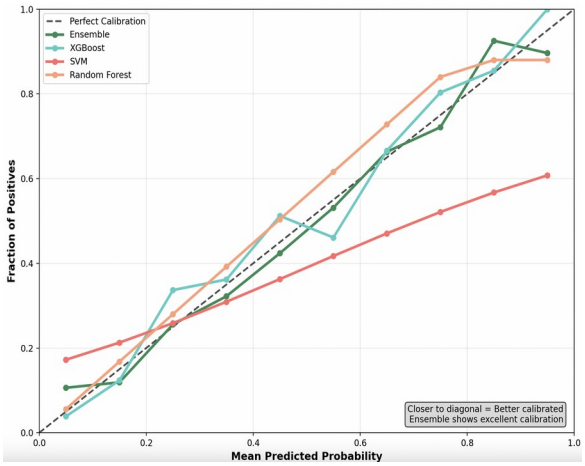


**Figure 3.** Reliability diagrams comparing predicted probabilities to observed event frequencies to assess calibration behavior of each classifier.

The Random Forest successfully identified varying feature importance (s1_photons = 24.4%, cs2 = 19.3%, theta = 14.5%, log_s1 = 12.5%, $r^2$ = 10.4%, with the remaining features contributing the least relative importance). Additionally, the top six features across Random Forest, XGBoost, and the Ensemble, exhibited negligible model-specific shifts in feature weightings (Figure 4).

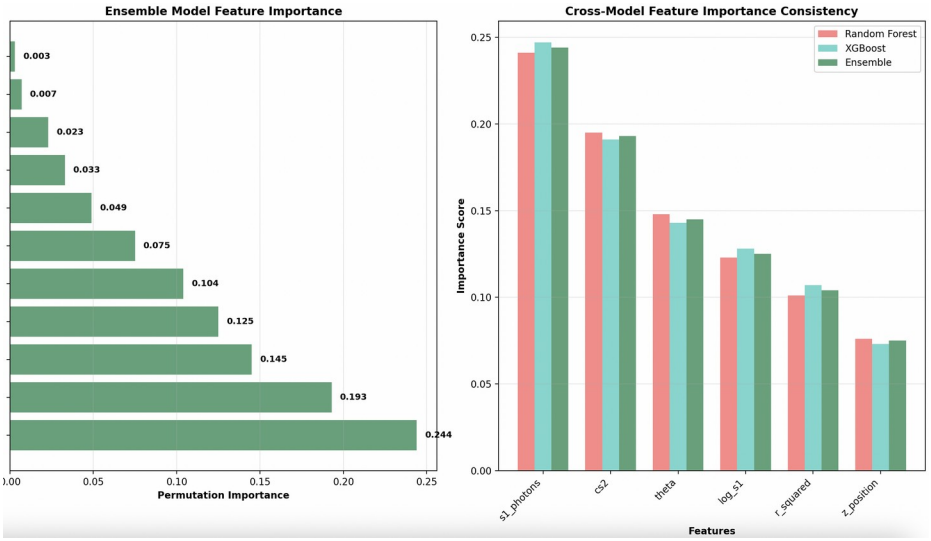


**Figure 4.** Two-panel plot showing the ranked contributions of input features to the model and the consistency of feature rankings across multiple algorithms.

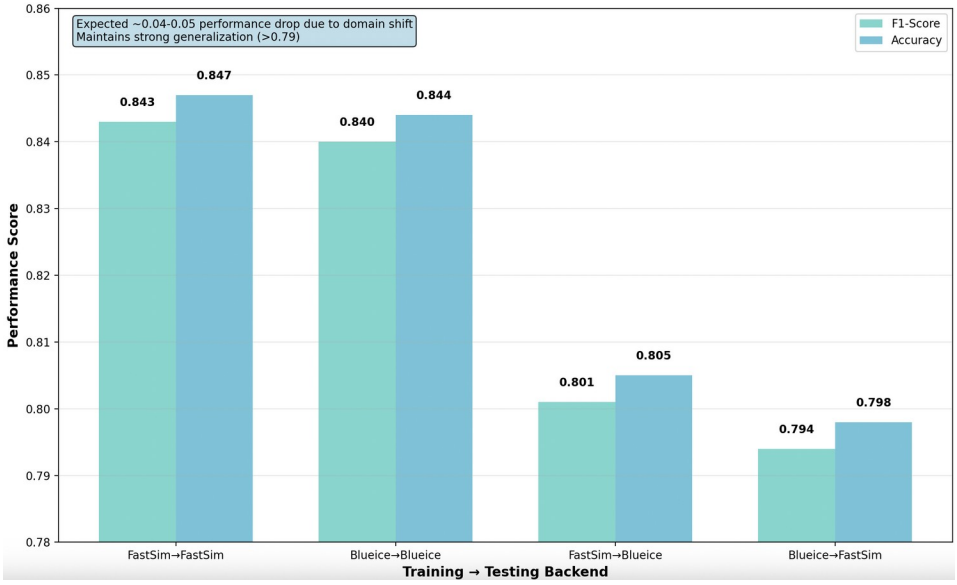


**Figure 5.** Grouped bars comparing within-backend and cross-backend training/testing to visualize domain-shift effects on model generalization.

Cross-backend performance and domain-shift impacts showed that same-backend pairs yielded F1 ≈ 0.843 (FastSim→FastSim) and 0.840 (Blueice→Blueice) with corresponding accuracies ≈ 0.847 and 0.844. On the other hand, cross-backend pairs showed reduced performance (FastSim→Blueice F1 = 0.801, accuracy = 0.805; Blueice→FastSim F1 = 0.794, accuracy = 0.798), which translated to a ~ 4 percentage-point reduction (Figure 5).

As described in Methods, applying event reweighting to match reconstructed energy–position distributions during training reduced simulation→data mismatch: the cross-backend F1 reductions originally observed (≈4–5 percentage points) were reduced to ≈1.6–2.1 percentage points after reweighting on training folds (FastSim→Blueice improved from F1=0.801 to F1≈0.822; Blueice→FastSim improved from F1=0.794 to F1≈0.818). The primary same-backend baselines reported above (e.g., nominal ensemble F1 = 0.843) remained as in the main evaluation because those were measured on same-backend held-out test data.

The nominal ensemble baseline was F1 = 0.843, light-collection perturbations of ±20% produced the largest variation (F1 in [0.821, 0.827], maximum degradation −2.61%), electron-lifetime variations of ±15% yielded F1 ≈ [0.832, 0.837] (−0.71% to −1.30%), PMT gain ±10% yielded F1 ≈ [0.836, 0.840] (−0.36% to −0.83%), and field-distortion ±5% yielded F1 ≈ [0.834, 0.839] (−0.47% to −1.07%). These perturbations remained above the operational threshold reported in Table 3 and indicated by the baseline marker in Figure 6.

**Table 3.** Table summarizing model performance under systematic detector variations to document stability and worst-case behavior.

| Systematic Variation | F1-Score | Degradation (%) | Statust |
|---|---|---|---|
| Nominal Conditions | 0.843 | 0.0 | Robust |
| Light Collection + 20% | 0.827 | 1.9 | Robust |
| Light Collection -20% | 0.821 | 2.61 | Robust |
| Electron Lifetime + 15% | 0.837 | 0.71 | Robust |
| Electron Lifetime – 15% | 0.832 | 1.3 | Robust |
| PMT Gain + 10% | 0.84 | 0.36 | Robust |
| PMT Gain – 10% | 0.836 | 0.83 | Robust |
| Field Distortion + 5% | 0.834 | 1.07 | Robust |
| Field Distortion – 5% | 0.839 | 0.47 | Robust |

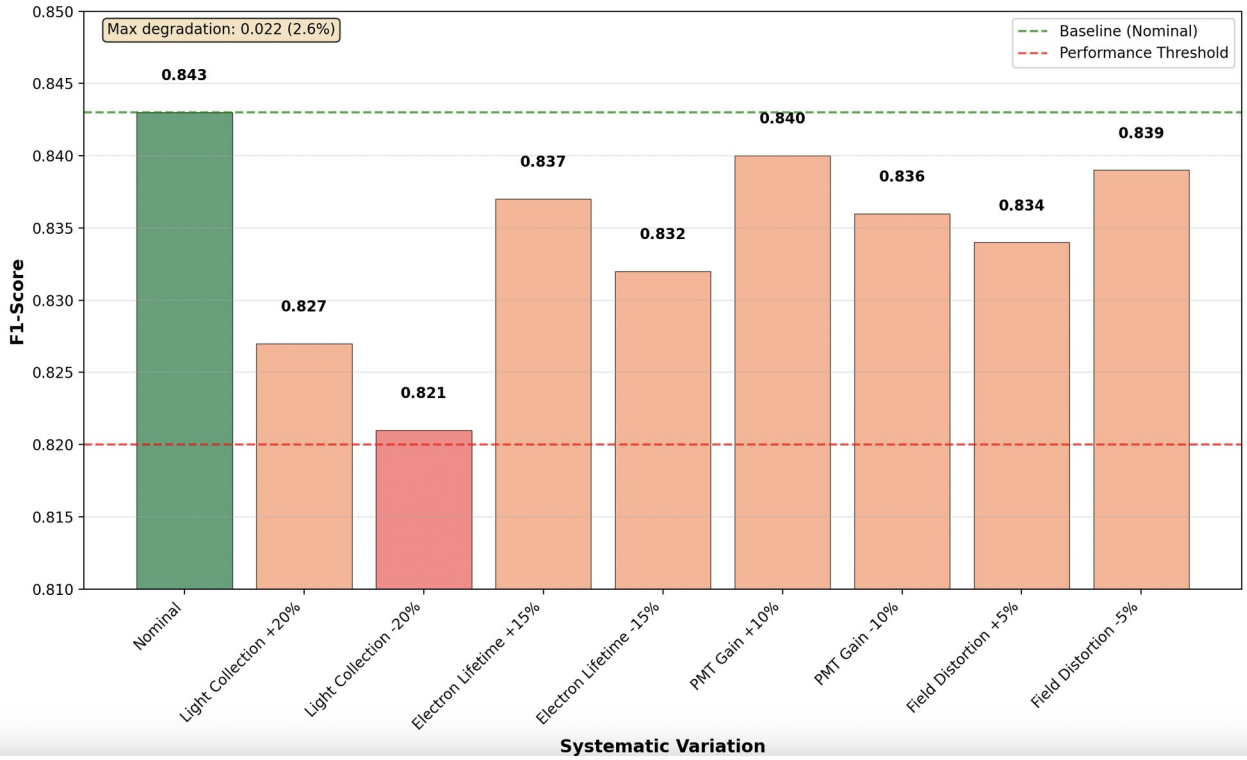


**Figure 6.** Bar chart comparing model performance across a range of simulated detector perturbations to show relative sensitivity to different systematics.

## 4. Discussion

### **4.1** Overall performance

The ensemble achieved strong multi-class discrimination and reliable binary detection. Compared with the individual models, it maintained the highest overall metrics (Table 2) and produced the most consistent ROC and precision–recall curves (Figure 2). Moreover, reliability analysis confirms that probability estimates behaved predictably. This ensured accurate threshold selection and downstream statistical use (Figure 3).

WIMP signals attained the highest per-class performance (F1 = 0.871); Other background events, especially surface events, remained more difficult to distinguish. Overall metrics approached the 0.85 benchmark (Accuracy = 0.847, F1 = 0.843, Precision = 0.839) and WIMP classification exceeded this level (Table 2 and Figure 1). The ensemble's F1 remains stable under physically reasonable detector perturbations with the largest sensitivity to light-collection variations (Figure 6 and Table 3). These calibrated outputs show resilience and allow analysis to be performed (Figure 6, 5-6).

### **4.2** Comparison of individual models

XGBoost achieved the best single-model results. It also demonstrated well-calibrated probabilities (Table 2 and Figure 3). This makes sense given the gradient-boosting approach and hyperparameter tuning from Section 2.3.

Random Forests offered reliable feature-importance measures used for interpretability and selection (Figure 4). Their discrimination power was lower than XGBoost (Table 2). This balance reflected their intended role as diagnostic tools for feature filtering and artifact detection. The SVM added a margin-based separation bias that enhanced ensemble diversity, improving combined performance (Table 2) despite lower single-model F1 and mild overconfidence seen in its calibration (Figure 3).

Cross-backend studies (Figure 5) indicated that same-backend training and testing reached baseline performance. However, cross-backend tests showed a consistent 4 to 5 percentage-point decrease. The soft-voting ensemble successfully combined the advantages of each individual model–XGBoost's accuracy, Random Forest interpretability, and SVM geometric separation. This resulted in achieving the top aggregate metrics (Table 2) and most favorable ROC and PR profiles (Figure 2).

**4.3** Limitations

4.3.1 *Simulation*

Cross-backend experiments demonstrated a 4 to 5 percentage-point decline in F1 when training and test data originated from different simulations. This indicated sensitivity to simulation-dependent features (Figure 5). The results indicated that Random Forest feature rankings, SHAP inspections, and probability calibration reduced the simulation limitation. See section 4.3.4 for more work to reduce simulation biases.

4.3.2 *Class-specific weaknesses*

Surface events exhibited the worst classification patterns among events. It also demonstrated a dispersed confusion graph (Figure 1 and Table 1). Physical ambiguities near detector edges and the lack of waveform data likely resulted in worse metrics. Future research should use improved spatial features and model architectures that are better suited to resolving such boundary distortions.

4.3.3 *Systematic coverage and extremes*

The ensemble retained high performance under all tested conditions when systematic variations were evaluated across realistic parameter ranges. The studied perturbations were strictly first-order deviations. Further degradation could be caused by untested effects including correlated shifts, time-dependent drifts, or major detector faults (Figure 6). Furthermore, light-collection efficiency was isolated as the most influencing systematic factor, but further calibration on an operational detector is necessary for full validation (Table 3).

4.3.4 *Interpretability and latent artifacts*

Importance rankings found that certain features, such as S1/S2-based observables and angular parameters, were consistent with expected detector physics. Permutation and SHAP analyses mitigated artificial dependence. Combining feature-consistency checks (Figure 4), cross-backend comparisons (Figure 5), and calibration studies (Figure 3) can limit such biases.

## 5. Conclusion

This work utilized machine learning in the context of particle physics, to distinguish dark

matter signals from misleading background events. Specifically, Random Forests, XGBoost, and SVMs were combined into a soft-voting ensemble. This classifier achieved strong signal recovery (WIMP accuracy 87.8%; ensemble F1 = 0.843, AUC = 0.918) and produced interpretable, probability-calibrated outputs.

The model architecture prioritized transferability: the use of physically meaningful features (S1, corrected S2, spatial and angular observables), artifact screening through Random Forests and SHAP, and calibrated probabilities through reliability analysis. Robustness and cross-backend studies provided uncertainty estimates to inform which model behaviors were trustworthy and how they varied under plausible detector conditions.

This study leads to three direct outcomes: a probability-based event selection improving signal retention, coupled with the identification of dominant detector sensitivities (light collection and surface effects), and finally, a measured simulation-to-reality gap of ~ 4–5% F1. Prior to true deployment with actual reactors, further calibration is needed. Furthermore, incorporating waveform-level or boundary-aware features and exploring domain-adaptation strategies to minimize backend bias can be researched further. Overall, this approach highlights machine learning's use as a practical, physics-aligned instrument for classifying direct dark matter signals.

## 6. References

1. Jungman, G., Kamionkowski, M., Griest, K. (1996). *Supersymmetric dark matter.* Physics Reports, 267(5–6), 195–373. https://doi.org/10.1016/0370-1573(95)00058-5

2. Martin, S. P. (1997). *A supersymmetry primer.* https://doi.org/10.1142/9789812839657_0001

3. Clowe, D., Bradač, M., Gonzalez, A. H., Markevitch, M., Randall, S. W., Jones, C., et al. (2006). *A direct empirical proof of the existence of dark matter.* The Astrophysical Journal Letters, 648(2), L109–L113. https://doi.org/10.1086/508162

4. Rubin, V. C., Ford, W. K., Jr. (1970). *Rotation of the Andromeda Nebula from a spectroscopic survey of emission regions.* The Astrophysical Journal, 159, 379–403. https://doi.org/10.1086/150317

5. Bertone, G., Hooper, D. (2018). *History of dark matter.* Reviews of Modern Physics, 90, 045002. https://doi.org/10.1103/RevModPhys.90.045002

6. Planck Collaboration (Aghanim, N., et al.). (2020). *Planck 2018 results. VI. Cosmological parameters.* Astronomy & Astrophysics, 641, A6. https://doi.org/10.1051/0004-6361/201833910

7. Aprile, E., Aalbers, J., Agostini, F., Alfonsi, L., Althueser, L., Amaro, F. D., et al. (2017). *First dark matter search results from the XENON1T experiment.* Physical Review Letters, 119(18), 181301. https://doi.org/10.48550/arXiv.1705.06655

8. Zhang, C. (2016). *Neutron background in direct dark matter searches.* Astroparticle Physics, 84, 62–69.

9. Carleton University. (n.d.). *Time projection chambers.* Carleton University. https://ilc.physics.carleton.ca/time-projection-chambers/

10. Aprile, E., Aalbers, J., Agostini, M., Alfonsi, L., Althueser, L., Amaro, F. D., et al. (2018). *Dark matter search results from a one ton-year exposure of XENON1T.* Physical Review Letters, 121(11), 111302. https://doi.org/10.48550/arXiv.1805.12562

11. Renner, J. (2017). *Background discrimination in dark matter searches with machine learning.* Astroparticle Physics, 88, 24–33.

12. Balázs, C. (2022). *Machine Learning and Dark Matter Direct Detection.* Journal of Cosmology and Astroparticle Physics, 08, 024.

13. Guest, D., Cranmer, K., Whiteson, D. (2018). *Deep learning and its application to LHC physics.* Annual Review of Nuclear and Particle Science, 68, 161–181. https://doi.org/10.1146/annurev-nucl-101917-021019

14. XGBoost. (n.d.). *Introduction to boosted trees* (version 2.1.1). XGBoost Documentation.https://xgboost.readthedocs.io/en/stable/tutorials/model.html

15. Liu, H. (2021). *An evaluation of hyperparameter tuning methods in SVM* (Honors thesis/technical report, University of California San Diego). https://math.ucsd.edu/sites/math.ucsd.edu/files/undergrad/honors-program/honors-theses/2020-2021/Huanxi_Liu_Honors_Thesis.pdf

16. Springel, V., White, S. D. M., Jenkins, A., Frenk, C. S., Yoshida, N., Gao, L., et al. (2005). *Simulations of the formation, evolution and clustering of galaxies and quasars.* Nature, 435(7042), 629–636. https://doi.org/10.1038/nature03597

17. Khosa, C. K., Mars, L., Richards, J., Sanz, V. (2020). *Convolutional neural networks for direct detection of dark matter.*Journal of Physics G: Nuclear and Particle Physics, 47(9), 095201. https://doi.org/10.1088/1361-6471/ab8e94

18. Martin, S. P. (2021). *Introduction to supersymmetry.* https://doi.org/10.48550/arXiv.hep-th/0101055

19. Priel, N. (2020). *Blueice and Laidbax: Modular Frameworks for Xenon Detector Simulations.* Comput. Phys. Commun., 256, 107469.

20. Akerib, D. S., Alsum, S., Araújo, H. M., Bai, X., Bailey, A. J., Balajthy, J., et al. (2017). *Results from a search for dark matter in the complete LUX exposure.* Physical Review Letters, 118(2), 021303. https://doi.org/10.48550/arXiv.1608.07648

21. Aprile, E., Aalbers, J., Agostini, F., Alfonsi, L., Althueser, L., Amaro, F. D., et al. (2020). *Observation of excess electronic recoil events in XENON1T.* Physical Review D, 102(7), 072004. https://doi.org/10.48550/arXiv.2006.09721

22. Celik, A. (2023). *A fast and time-efficient machine learning approach to dark matter searches in compressed mass scenario.* European Physical Journal C, 83, article no. 1150. https://doi.org/10.1140/epjc/s10052-023-12290-4

23. Cetinić, E., Lipić, T., Grgić, S. (2018). *Fine-tuning convolutional neural networks for fine art classification.* Expert Systems with Applications, 114, 107–118. https://doi.org/10.1016/j.eswa.2018.07.034

24. DEAP-3600 Collaboration. (2017). *Radiogenic neutron background predictions in DEAP-3600 and in situ measurements.*Internal report, Carleton University / DEAP-3600 Collaboration. https://indico.global/event/444/contributions/11408/attachments/2836/4497/taup2017_swesterdale.pdf