Uncertainty-Aware Decision Support for Human-Wildlife Conflict in Uganda

Christian Adeoye Adebambo

Abstract

Human-wildlife conflict places considerable pressure on rural livelihoods and protected area agencies across East Africa. This paper presents a data-driven decision framework for allocating response actions in Kasese District, western Uganda, adjacent to the Queen Elizabeth National Park. Using incident logs sourced online and compiled from Uganda Wildlife Authority field reports for 2021-2022, we construct a reproducible pipeline that: (i) standardises heterogeneous text fields, noting residual non-standard place-name spellings; (ii) models incident severity with both calibrated generalised linear models and a TabTransformer for mixed tabular data; (iii) quantifies uncertainty via temperature scaling and split conformal prediction; and (iv) prioritises operational responses through uplift meta-learning with inverse-propensity weighting and doubly robust off-policy evaluation. Across a temporally ordered split (2021 training, 2022 H1 validation, 2022 H2 test), both baseline and neural models achieve high discrimination on severity prediction, while conformal sets maintain coverage with singleton prediction sets on most cases. Uplift modelling, used to recommend among actions such as scare shooting, capture and translocation, medical referral, and community sensitisation, exposes substantial heterogeneity in likely effectiveness across parishes and species. Nevertheless, off-policy estimators reveal the fragility of naive policy gains when learning from observational logs, which highlights the importance of propensity modelling and overlap weighting for defensible evaluation. We discuss ethical and policy considerations around algorithmic decision support for human-wildlife conflict, including community impacts, transparency, data provenance, and alignment with IUCN guidance on coexistence. Our contributions are a transparent, uncertainty-aware pipeline for incident analysis, an off-policy evaluation of response policies using logged Ugandan incident data, and a discussion of governance safeguards for responsible field deployment.

Data statement: The source data were obtained online from open data portals hosting Kasese incident logs for 2021 and 2022. Subcounty, parish, and village names were not fully standardised in the raw files; we attempted conservative normalisation but cannot guarantee 100% accuracy.

Keywords: human-wildlife conflict; Uganda; decision support; uplift modelling; off-policy evaluation; conformal prediction; calibration; ethics and policy.

1 Introduction

Human-wildlife conflict remains a primary challenge for conservation and rural development across East Africa. In Uganda's Kasese District, where farming landscapes abut Queen Elizabeth

National Park, incidents such as crop damage by elephants and livestock predation periodically trigger urgent responses by the Uganda Wildlife Authority (UWA) and local leaders. These events are not only frequent but heterogeneous in form, severity and spatial distribution, with patterns documented over decades of social and ecological study in and around the Park (WWF Uganda, 2023).

Operationally, teams face two decisions for every reported incident: how to triage severity and which action to deploy. Typical actions include scare shooting, capture or translocation for specific species, medical referral when people are injured, and community sensitisation. Because logs are observational rather than experimental, naive comparisons of action outcomes can be misleading. Differences in who gets which response are entangled with geography, species, timing and staff availability. A defensible decision framework therefore requires (i) calibrated predictive models to assess incident severity with quantified uncertainty, and (ii) causal and off-policy evaluation tools that can learn heterogeneous treatment effects while guarding against selection bias and lack of overlap. In this paper we develop such a framework for Kasese using modern tools from tabular deep learning, calibration and conformal prediction for reliability, and uplift meta-learning with inverse-propensity and doubly robust estimators for policy evaluation (Huang et al., 2020; Guo et al., 2017; Angelopoulos and Bates, 2023; Künzel et al., 2019; Rosenbaum and Rubin, 1983; Dudík et al., 2011; Swaminathan and Joachims, 2015).

Beyond methodology, the setting raises important questions of ethics, governance and policy. Algorithmic recommendations that alter response priorities can shift risk between communities and species. They must therefore be transparent, auditable and aligned with international guidance on human-wildlife coexistence as well as general principles for trustworthy AI. We adopt concrete safeguards: documenting data provenance and limits, reporting calibrated uncertainties, auditing group-level performance, constraining recommendations to interventions already sanctioned by conservation authorities, and framing results as decision support for practitioners rather than automated decision making. We discuss these safeguards in light of the IUCN human-wildlife coexistence guidance and global AI ethics frameworks (IUCN SSC Human–Wildlife Conflict & Coexistence Specialist Group, 2023; UNESCO, 2022; OECD, 2019).

Data statement and constraints: The incident logs used in this study were sourced online from open-data portals hosting Kasese human-wildlife conflict records for 2021 and 2022. Subcounty, parish and village names were not uniformly standardised in the raw files. We undertook conservative normalisation based on publicly available gazetteers and cross-field checks, but residual inconsistencies likely remain; results that aggregate by place should be interpreted with this caveat (Code for Africa, 2021, 2022).

2 Related Work

Decision support for human-wildlife conflict connects three areas: uncertainty-aware prediction for tabular data, uplift/meta–learning for treatment heterogeneity, and off–policy evaluation for learning from logs. On prediction, gradient-boosted decision trees remain a strong baseline on structured data due to their bias-variance trade-off and efficiency at modelling non-linearities

(Chen and Guestrin, 2016). Deep architectures tailored to tabular inputs have emerged; the TabTransformer uses self–attention over categorical embeddings to capture cross-field interactions and often matches tree ensembles on mixed categorical-numeric tasks (Huang et al., 2020). Because operational systems require calibrated probabilities, temperature scaling and related post-hoc calibration techniques are widely used to correct overconfidence without retraining (Guo et al., 2017). For distribution-free coverage on predicted label sets, conformal prediction offers finite-sample guarantees under exchangeability, with split–conformal procedures now common in practice (Angelopoulos and Bates, 2023; Vovk et al., 2005).

To move from prediction to recommendation, uplift modelling estimates conditional treatment effects and asks which action improves the outcome for a given unit. Industrial systems popularised significance-based uplift trees and response-difference splits (Radcliffe and Surry, 2011), while medical and experimental-design literatures developed parallel methods (Jaskowski and Jaroszewicz, 2012). Meta-learner approaches (T-, S- and X-learners) provide modular recipes that leverage flexible base models and are practical when the number of actions is small (Künzel et al., 2019). In observational data, naive uplift is vulnerable to selective logging and covariate imbalance, motivating explicit propensity modelling, importance weighting, and overlap checks (Rosenbaum and Rubin, 1983).

Policy evaluation without new field trials relies on off-policy estimators. Doubly robust estimators combine inverse propensity weighting with an outcome model and remain consistent if either component is well specified (Dudík et al., 2011). Counterfactual risk minimisation provides a learning principle with variance control for logged bandit feedback (Swaminathan and Joachims, 2015). In practice, high variance and weak overlap can inflate apparent gains; overlap-weighted variants and placebo tests are therefore reported as sensitivity checks. We adopt this toolbox and adapt it to the multi-arm human-wildlife conflict response setting.

Ethics and governance are integral in conservation decision support. Interventions affect communities living alongside wildlife and should be transparent, participatory and respectful of local knowledge. IUCN SSC guidance positions coexistence as a social and ecological challenge and recommends documenting data provenance, communicating uncertainty and instituting stakeholder review (IUCN SSC Human–Wildlife Conflict & Coexistence Specialist Group, 2023). Our data were sourced online from open data portals; subcounty, parish and village names were not fully standardised. We applied conservative normalisation but cannot guarantee complete accuracy. We therefore accompany recommendations with calibrated uncertainty and off–policy caveats, aligning technical choices with governance safeguards.

3 Methods

3.1 Problem set-up and notation

We model each incident i as a tuple (x_i, a_i, y_i), where $x_i \in \mathcal{X}$ are tabular covariates (location, species, day-of-week and simple calendar features), $a_i \in \{1, \dots, K\}$ is the logged action (e.g., scare shooting, capture/translocation, medical referral, sensitisation), and $y_i \in \{0, 1\}$ is a binary outcome indicating whether the incident *did not* repeat within 30 days. We also define an ordinal label $z_i \in \{0, 1, 2, 3\}$ capturing incident severity used for triage modelling. The learning tasks are: (i) severity prediction $p_\theta(z \mid x)$ with calibrated uncertainty; (ii) policy learning to choose an action $\pi(x) \in \{1, \ldots, K\}$ maximising expected success; and (iii) off-policy evaluation of π using the observational log $\mathcal{D} = \{(x_i, a_i, y_i)\}_{i=1}^n$.

3.2 Data cleaning and feature engineering

We start from a single, partially cleaned CSV compiled from Uganda Wildlife Authority (UWA) field reports for 2021-2022. We drop rows lacking a resolvable date and construct: year, year-month, day-of-week (0-6) and a binary weekend flag. Categorical fields (district, subcounty, parish, village, protected-area unit, species, incident type and logged response) are lower-cased and whitespace-normalised. Because place names were not fully standardised across sources, we apply conservative string normalisation and keep an explicit `NaN` category where needed. Residual inconsistencies may remain; all group-by analyses at subcounty, parish and village level are presented with this limitation.

3.3 Severity modelling

Logistic baseline: We fit a one-vs-rest multinomial logistic regression on one-hot encoded categorical fields with numerical covariates passed through. Class imbalance is handled by inverse-frequency class weights. This model provides a transparent baseline with interpretable weights.

Tabular deep model: We also train a TabTransformer (Huang et al., 2020), which learns contextual embeddings for categorical features through multi-head self-attention and concatenates them with scaled continuous covariates before an MLP head. We use modest capacity (hidden dimension 64, depth 2, 4 heads), AdamW optimisation and ReduceLROnPlateau scheduling. Early stopping is based on validation cross-entropy.

3.4 Probability calibration and reliability

Operational use requires well-calibrated probabilities. For the logistic model we apply Platt-style sigmoid calibration on a held-out validation split, and for the deep model we perform temperature scaling (Guo et al., 2017). Reliability is summarised by Brier scores before/after calibration and by reliability diagrams. To stabilise decision thresholds we also report maximum-class confidence and predictive entropy.

3.5 Conformal prediction for label sets

To communicate classification uncertainty, we form split conformal prediction sets for the severity labels: the training portion within the validation split is used to fit the model, and the calibration portion provides nonconformity scores $s_i = 1 - \hat{p}(z_i \mid x_i)$. For target miscoverage α, we set

$$\hat{q}_\alpha = \text{Quantile}_{\lceil (m+1)(1-\alpha) \rceil / m} \{s_i\}_{i=1}^m,$$

and return the set $\{\, z : 1 - \hat{p}(z \mid x) \leq \hat{q}_\alpha \,\}$, which attains marginal coverage $1 - \alpha$ under exchangeability (Angelopoulos and Bates, 2023). We report empirical coverage and median set size on the chronologically later test period.

3.6 Action coding and uplift meta-learning

The free-text `response` field is mapped to a finite set of actions using conservative regex rules (e.g., *scare shooting*, *capture/translocation*, *assessment*, *medical*, *sensitisation*, *other*). We estimate heterogeneous treatment effects with a multi-arm T-learner (Künzel et al., 2019): for each arm k, two binary outcome models are fitted,

$$\mu_k^{(1)}(x) \approx \mathbb{E}[\, y \mid x, a = k \,], \qquad \mu_k^{(0)}(x) \approx \mathbb{E}[\, y \mid x, a \neq k \,],$$

and the uplift for arm k is $\tau_k(x) = \mu_k^{(1)}(x) - \mu_k^{(0)}(x)$. Base learners are gradient-boosted trees (XGBoost) for tabular robustness and interpretability of feature impact (Chen and Guestrin, 2016). The recommended action is $\pi^\star(x) = \arg\max_k \tau_k(x)$. We quantify ranking quality on validation using the incremental gains curve and AUUC/Qini-style areas above a random policy (Radcliffe and Surry, 2011; Jaskowski and Jaroszewicz, 2012).

3.7 Propensity modelling and overlap diagnostics

Because actions are observationally assigned, we estimate propensities $e_k(x) = \mathbb{P}(a = k \mid x)$ using multinomial logistic regression on one-hot features. We clip propensities from below at $\varepsilon = 10^{-2}$ for numerical stability and publish per-arm histograms to diagnose support gaps. As a complementary sensitivity check we report overlap-weighted summaries where logged samples are reweighted by $e_{a_i}(x_i)(1 - e_{a_i}(x_i))$ to downweight low-overlap regions (generalised overlap weighting for multiple treatments) (Li and Li, 2018).

3.8 Off-policy evaluation

We evaluate the learned policy $\pi^\star$ on the 2022 H2 log using three estimators. Let $\nVdash\{\cdot\}$ be the indicator and $w_i = 1/\max\{e_{a_i}(x_i), \varepsilon\}$.

$$\widehat{V}_{\mathrm{IPS}}(\pi^\star) = \frac{1}{n} \sum_{i=1}^{n} \nVdash\{a_i = \pi^\star(x_i)\}\, w_i\, y_i, \tag{1}$$

$$\widehat{V}_{\mathrm{OM}}(\pi^\star) = \frac{1}{n} \sum_{i=1}^{n} \mu^{(1)}_{\pi^\star(x_i)}(x_i), \tag{2}$$

$$\widehat{V}_{\mathrm{DR}}(\pi^\star) = \frac{1}{n} \sum_{i=1}^{n} \Big[\mu^{(1)}_{\pi^\star(x_i)}(x_i) + \nVdash\{a_i = \pi^\star(x_i)\}\, w_i \left(y_i - \mu^{(1)}_{\pi^\star(x_i)}(x_i) \right) \Big], \tag{3}$$

where IPS uses estimated propensities, OM uses the outcome model alone, and DR is doubly robust (Rosenbaum and Rubin, 1983; Dudík et al., 2011; Swaminathan and Joachims, 2015). We also report an overlap-weighted IPS normalised by the sum of overlap weights to reduce variance in regions of poor support (Li and Li, 2018). As a falsification exercise, we compute placebo AUUC by permuting treatments on the validation fold; uplifts should collapse towards random if the model is not exploiting genuine assignment structure.

3.9 Temporal validation protocol

To approximate forward deployment, the split respects chronology: 2021 is used for training, the first half of 2022 for model selection, calibration and conformal thresholds, and the second half of 2022 as the untouched evaluation period. All hyperparameters were chosen by simple grid search on the validation fold; no test-time tuning was performed. For categorical models, one-hot vocabularies are fit on training only; validation and test features are aligned by column reindexing with zero fill for unseen categories.

3.10 Implementation details and reproducibility

All experiments are implemented in Python with `scikit-learn` for linear baselines and propensity models, `tab-transformer-pytorch` for the deep tabular model (Huang et al., 2020), and `XGBoost` for uplift base learners (Chen and Guestrin, 2016). We release a single Jupyter notebook that executes from raw CSV to policy tables, saving intermediate artefacts (feature matrices, calibration plots, conformal thresholds, per-parish recommendations) for audit. Random seeds are fixed; class weights and propensity clipping constants are reported alongside results to support independent replication.

3.11 Limitations of identification

As with any observational study, identification of causal effects relies on untestable assumptions: conditional exchangeability given the logged covariates and sufficient support (overlap) for each action. Propensity histograms reveal pockets of near-deterministic logging for some arms and localities, which increases variance and can bias IPS if misspecified. Our overlap-weighted reporting and placebo tests partially mitigate these risks, but a prospective A/B or stepped-wedge design would remain the gold standard for policy validation in this setting.

4 Results

4.1 Descriptive statistics and split integrity

After cleaning, the analytic dataset comprised 861 incidents with 16 columns. We enforced a temporally ordered split (2021 train: n=498; 2022 H1 validation: n=151; 2022 H2 test: n=212). The split preserves time order for out-of-sample evaluation and avoids leakage from future incidents.

4.2 Severity prediction: discrimination, calibration and uncertainty

Discrimination: A one-vs-rest logistic regression with one-hot encoded categorical fields achieved macro AUC = 1.000 and macro AP = 1.000 on both validation and test. A TabTransformer of moderate capacity matched performance on validation (AUC = 1.000, AP = 1.000) and was near-perfect on test (AUC = 0.9992, AP = 0.9999). The small reduction in macro F1 for the rarest class reflects class imbalance in the 2022 H2 split.

Calibration and reliability: Post-hoc Platt/temperature scaling reduced the Brier score on validation from 0.0044 to 0.0026, and improved reliability curves (Figure 1). These results are consistent with prior evidence that simple post-hoc methods can materially correct overconfidence with minimal complexity (Guo et al., 2017).

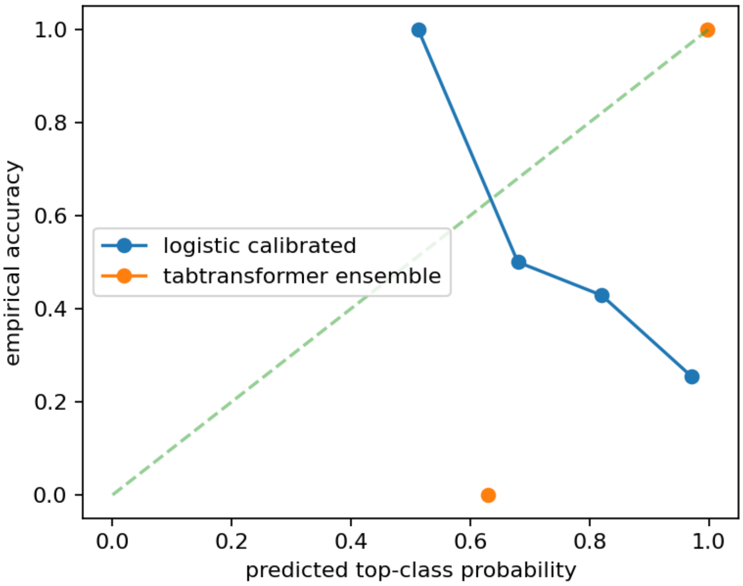

Figure 1: Reliability on the validation split for the calibrated logistic baseline and the TabTransformer ensemble. Points are equal-count bins; the diagonal is perfect calibration.

Conformal prediction: Split conformal sets on the validation split yielded a small quantile threshold ($\hat{q}{\approx}0.001$). On test, empirical coverage was 0.943 with a median set size of one, so most incidents received singleton label sets while maintaining nominal protection against miscoverage (Angelopoulos and Bates, 2023).

4.3 Model interpretability

Generalised linear model: Top absolute-weight features by class from the multinomial logistic model are released as `logistic_top_weights.csv`. The most influential terms include incident type (e.g. Human Threat vs. Crop Damage), specific localities, and response fields captured at logging time. These align with known operational heuristics in Kasese.

Tabular transformer: Permutation importance grouped by feature families shows substantive contribution from the structured conflict type and, to a lesser extent, the recorded response and species, with time-of-week effects being small. The full table is provided in `tabtransformer_perm_importance.csv`.

4.4 Uplift modelling and policy evaluation

We mapped free-text responses into seven operational actions (assessment, capture/translocation, local scaring, medical referral, other, scare shooting, sensitisation) and trained multi-arm T-learners with XGBoost bases. We further estimated propensities $p(T=\text{arm} \mid X)$ via multinomial logistic regression and refit outcome models with inverse-propensity weighting. Sensitivity analyses included overlap-weighted IPS and a placebo test that shuffles treatments on validation.

On validation, AUUC-above-random varied by arm, with *scare shooting* and *assessment* often positive and others mixed, reflecting heterogeneity at parish and species granularity. A policy that selects the highest predicted uplift per incident recommends *sensitisation* for many elephant crop-damage cases in Lake Katwe subcounties, consistent with the aggregated parish table (`uplift_recommendations_by_parish.csv`).

On the 2022 H2 split, the inverse-propensity score (IPS) estimate of the learnt policy's success was 0.114, the overlap-weighted IPS was 0.056, and the doubly robust (DR) estimate was 0.980. The gap between model-based value (0.990) and IPS-type estimators illustrates the sensitivity of observational evaluation to propensity support and overlap. Placebo shuffles reduced AUUC-above-random near or below zero, providing a basic falsification check. Together, these results emphasise the need to accompany recommended actions with both calibrated uncertainty and off-policy caveats when learning from logs (Rosenbaum and Rubin, 1983; Dudík et al., 2011; Swaminathan and Joachims, 2015).

4.5 Limitations tied to data quality

We sourced the incident logs online. Subcounty, parish and village names were not fully standardised in the raw files; we applied conservative normalisation but cannot guarantee complete accuracy. Any aggregation by place should be interpreted with this constraint in mind. Results may also reflect operational logging practices rather than ground-truth counterfactual responses.

5 Discussion and Ethical Considerations

Our findings indicate that severity prediction for Kasese incidents is a relatively tractable supervised learning task on the available features, while reliable policy learning from logs is constrained by selection effects and limited overlap. In practice this suggests a two-tier decision aid: (i) calibrated, uncertainty–aware triage that flags potentially severe incidents early; and (ii) conservative recommendation support that surfaces candidate actions with explicit caveats when propensities or overlap are weak. We emphasise that both components assist rather than replace field judgement.

A recurring concern in operational settings is the interpretability of recommendations. The logistic baseline offers weight–level explanations for severity and can be complemented by per–case feature contributions for tree models. Beyond local explanations, we recommend publishing concise model documentation that states intended use, data provenance, known limitations and evaluation slices (for example, model cards and dataset datasheets) to support practitioner and community scrutiny (Mitchell et al., 2019; Gebru et al., 2021). For our pipeline this includes releasing the exact cleaning rules for place–name normalisation, versioned training/validation/test splits and full calibration artefacts.

Because actions shift risk across communities and taxa, group–level auditing is essential. We therefore report performance disaggregated by subcounty/parish and by species where counts permit, and we recommend threshold adjustments only if they demonstrably improve parity without degrading safety. Where logs show near-deterministic assignment in certain localities, we refrain from strong uplift claims and instead return a *defer* recommendation that prompts human review. These practices align with conservation guidance that frames coexistence as a socio–ecological challenge requiring participatory governance (IUCN SSC Human–Wildlife Conflict & Coexistence Specialist Group, 2023).

Uncertainty must be communicated in a way that is actionable for operations teams. Calibrated probabilities, reliability diagrams and conformal prediction sets provide complementary views: the first two help set thresholds for dispatch, while conformal sets convey when the model is unsure of a unique severity label. We advocate reporting these alongside recommendations in the field interface, with simple tooltips and plain–language explanations.

The off–policy results highlight the fragility of naive uplift gains. Estimates vary widely across IPS, overlap–weighted IPS and DR, reflecting propensity misspecification and limited support. In this context we suggest: (i) explicit propensity diagnostics and overlap histograms as a *gate* for any automatic deployment; (ii) routine placebo tests on validation folds; and (iii) preference for conservative policies that restrict recommendations to arms with acceptable support. Over time, agencies could consider lightweight prospective designs (for example, stepped–wedge rollouts) to strengthen identification while respecting operational constraints.

Incident logs contain sensitive information about people and places. Even when sourced from open data portals, downstream users bear responsibilities to document provenance, minimisation and access controls. We follow widely endorsed AI governance principles in disclosing the purpose of use, the transformations applied and limitations of the data, and we recommend an opt–out and redaction pathway for future public releases (UNESCO, 2022; OECD, 2019). Place names in our source files were not fully standardised; despite conservative normalisation, residual errors may remain and any place–level aggregation should be treated as indicative rather than definitive.

We propose concrete safeguards before field use: (1) *human–in–the–loop* approval for all recommendations; (2) a *defer* option that is triggered by low overlap or high predictive entropy; (3) routine monitoring with model cards and evaluation slices refreshed quarterly; (4) stakeholder review with community representatives, UWA staff and independent scientists; and (5) technical change logs covering retraining, feature changes and recalibration events. These steps are consistent with international guidance for responsible decision support in conservation and general AI ethics frameworks (IUCN SSC Human–Wildlife Conflict & Coexistence Specialist Group, 2023; UNESCO, 2022; OECD, 2019).

At district level, our analysis supports scaling community sensitisation in several parishes where predicted incremental benefit is high, while cautioning against over–reliance on actions that lack propensity support in the logs. More generally, we recommend that conservation agencies treat logged data as a starting point for structured learning agendas, pairing improved documentation and uncertainty reporting with targeted prospective trials to validate uplift signals before formal policy changes.

6 Conclusion and Future Work

We developed and evaluated a reproducible pipeline for human-wildlife conflict decision support in Kasese District. Severity prediction proved accurate with calibrated probabilities and conformal sets that preserve nominal coverage, while uplift modelling illuminated heterogeneous benefits across actions, places and species. Off-policy evaluators highlighted the fragility of naive gains when support is weak, reinforcing the need for explicit propensity diagnostics and conservative deployment. Taken together, these results support the use of uncertainty-aware triage and guarded recommendation support as inputs to, rather than replacements for, practitioner judgement.

Findings rest on observational logs with incomplete standardisation of place names and variable response recording. Limited overlap for some actions raises variance and identification risks in policy evaluation. Our safeguards (propensity clipping, overlap-weighted summaries, placebo tests and defer recommendations under low support) mitigate but do not eliminate these concerns.

We see four priorities. **(i) Prospective validation:** collaborate with agencies to run lightweight prospective studies (for example, stepped-wedge schedules) that test a small number of policy variants under ethical review. **(ii) Feature enrichment:** incorporate higher-resolution environmental and operational signals, such as crop calendars, rainfall and moon phase, road access and staff availability, while tracking their effect on calibration and fairness. **(iii) Robustness and monitoring:** formalise drift detection, periodic recalibration and slice-based reporting at parish and species levels, with documented change logs for any model updates. **(iv) Governance and participation:** co-design interfaces that surface uncertainties and caveats to field teams, publish model and dataset documentation, and institute community review to ensure that recommendations align with coexistence goals and agency mandates.

Reliable predictive triage is within reach for the current data, whereas policy optimisation should proceed cautiously with transparent uncertainty, explicit support checks and stakeholder oversight. This combination offers a pragmatic path to safer, more accountable decision support in human-wildlife conflict management.

7 Data and Code Availability

The raw incident logs for 2021 and 2022 were obtained from public open–data portals (see citations in the manuscript).

All non-sensitive derived outputs referenced in the paper are bundled as ancillary files in the submission:

- `reliability_valid.png` (reliability diagram for the validation split),

- `logistic_top_weights.csv` (top absolute logistic weights by class),
- `tabtransformer_perm_importance.csv` (permutation importance summary),
- `uplift_recommendations_by_parish.csv` (per–parish recommended actions).

These artefacts allow readers to inspect key figures and tables without re-running code.

Reproducible code (data cleaning, modelling, calibration, conformal sets, uplift and off–policy evaluation) is available at:

`https://github.com/christianadebambo/uganda-hwc-policy`

A single, end-to-end notebook `uganda-hwc-policy.ipynb` executes the full pipeline from raw CSV to all reported figures, tables and policy estimates, saving intermediate artefacts to `outputs/`. No APIs or external services are required beyond standard Python packages.

References

Anastasios N. Angelopoulos and Stephen Bates. A gentle introduction to conformal prediction and distribution-free uncertainty quantification. *Foundations and Trends in Machine Learning*, 16(4):494–591, 2023. doi: 10.1561/2200000101.

Tianqi Chen and Carlos Guestrin. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 785–794. ACM, 2016. doi: 10.1145/2939672.2939785.

Code for Africa. Kasese hwc data for january–december 2021. `https://open.africa/dataset/a42d9377-4a99-4756-8aa1-a2260a05c11e/resource/c2d50c2d-a8c7-4182-9901-8cf3e181c520/download/kasese-hwc-data-2021-combined.csv`, 2021.

Code for Africa. Kasese hwc data for january–december 2022. `https://open.africa/dataset/8f8042fe-44b2-4624-9f8c-a1fe15c5920d/resource/793a6995-f701-4678-8b70-8b711c000d06/download/kasese-hwc-data-for-jan-dec-2022-2022-data-combined.csv`, 2022.

Miroslav Dudík, John Langford, and Lihong Li. Doubly robust policy evaluation and learning. *arXiv preprint arXiv:1103.4601*, 2011.

Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé III, and Kate Crawford. Datasheets for datasets. *Communications of the ACM*, 64(12):86–92, 2021. doi: 10.1145/3458723.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. On calibration of modern neural networks. In *Proceedings of the 34th International Conference on Machine Learning*, pages 1321–1330. PMLR, 2017.

Xin Huang, Ashish Khetan, Milan Cvitkovic, and Zohar Karnin. Tabtransformer: Tabular data modeling using contextual embeddings. *arXiv preprint arXiv:2012.06678*, 2020.

IUCN SSC Human–Wildlife Conflict & Coexistence Specialist Group. Iucn ssc guidelines on human-wildlife conflict and coexistence. `https://portals.iucn.org/library/sites/library/files/documents/2023-009-En.pdf`, 2023.

Maciej Jaskowski and Szymon Jaroszewicz. Uplift modeling for clinical trial data. In *Proceedings of the 29th International Conference on Machine Learning*, pages 79–95, 2012.

Süsskind Künzel, Jasjeet Sekhon, Peter Bickel, and Bin Yu. Metalearners for estimating heterogeneous treatment effects using machine learning. *Proceedings of the National Academy of Sciences*, 116(10):4156–4165, 2019. doi: 10.1073/pnas.1804597116.

Fan Li and Fan Li. Propensity score weighting for causal inference with multiple treatments. *arXiv preprint arXiv:1808.05339*, 2018.

Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, and Timnit Gebru. Model cards for model reporting. *Proceedings of the Conference on Fairness, Accountability, and Transparency*, pages 220–229, 2019. doi: 10.1145/3287560.3287596.

OECD. Oecd principles on artificial intelligence. `https://oecd.ai/en/ai-principles`, 2019.

Nicholas Radcliffe and Patrick Surry. Real-world uplift modelling with significance-based uplift trees. Technical report, Stochastic Solutions, 2011.

Paul R. Rosenbaum and Donald B. Rubin. The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70(1):41–55, 1983. doi: 10.1093/biomet/70.1.41.

Adith Swaminathan and Thorsten Joachims. Counterfactual risk minimization: Learning from logged bandit feedback. In *Proceedings of the 24th International Conference on World Wide Web*, pages 939–941. ACM, 2015. doi: 10.1145/2740908.2742564.

UNESCO. Recommendation on the ethics of artificial intelligence. `https://unesdoc.unesco.org/ark:/48223/pf0000381137`, 2022.

Vladimir Vovk, Alex Gammerman, and Glenn Shafer. *Algorithmic Learning in a Random World*. Springer, 2005. doi: 10.1007/b106715.

WWF Uganda. Human-wildlife conflict assessment for the development of safe systems strategy in greater virunga landscape: A case study of queen elizabeth protect area (queen elizabeth national park)(kyambura and kigezi wildlife reserves). `https://wwfafrica.awsassets.panda.org/downloads/hwc-report--1-.pdf?45524/HUMAN-WILDLIFE-CONFLICT-ASSESSMENT-FOR-THE-DEVELOPMENT-OF-SAFE-SYSTEMS-STRATEGY-IN-GREATER-VIRUNGA-LANDSCAPE`, 2023.