2025 ASEE Northeast Section Conference, March 22, 2025, University of Bridgeport, Bridgeport, CT, USA.

A Decision-Making Framework for Addressing the Imbalanced Learning Problem in Standoff Detection

Paul Cotae
Department of Electrical and
Computer Engineering
University of the District of Columbia
Washington, D.C., USA
pcotae@udc.edu

Nian Zhang
Department of Electrical and
Computer Engineering
University of the District of Columbia
Washington, D.C., USA
nzhang@udc.edu

Onyinye Obioha-Val
Department of Electrical and
Computer Engineering
University of the District of Columbia
Washington, D.C., USA
onyinye.obiohaval@udc.edu

Abstract—**Decision making, particularly in Machine Learning and Data Sciences, faces a major challenge of skewing to the majority dataset. The disproportion creates difficulties for traditional machine learning models, which most times gives false predictions and inaccurate results. We propose in this paper a decision-making framework for addressing imbalanced learning problems in standoff detection of threat chemicals. Our goal is to formulate a decision-making framework where detection thresholds and confidence scores are optimized to minimize false negatives, using Synthetic Minority Oversampling Technique (SMOTE) and Random Forest training classifier.**

Keywords—Decision making, Machine Learning, Imbalanced Data, false negative, Standoff Detection.

1. Introduction

In recent years, machine learning and data science have become indispensable tools in a wide range of applications, from healthcare to national security. However, one persistent challenge is the issue of imbalanced data—where the minority class is significantly underrepresented compared to the majority. This imbalance often skews decision-making processes, leading traditional algorithms to favor the majority class, which in turn results in false predictions and reduced overall accuracy. In critical applications, such as standoff detection of threat chemicals, even a small number of false negatives can have severe consequences for safety and security.

The complexity of accurately detecting threat chemicals is compounded by the inherent rarity of these events relative to normal background signals, making standard machine learning models ill-suited for such tasks. To overcome these challenges, advanced data augmentation techniques and robust classifiers are necessary. In this paper, we propose a novel decision-making framework that optimizes detection thresholds and confidence scores to minimize false negatives. By integrating the Synthetic Minority Oversampling Technique (SMOTE) with a Random Forest training classifier, our approach aims to provide a more reliable and accurate solution for imbalanced learning problems in the context of standoff chemical detection.

The detection of threat chemicals is crucial in ensuring security across various sectors, including national defense, environmental monitoring, and industrial safety. Standoff detection methods allow for the identification of hazardous substance from a distance, thus minimizing the risk of exposure to the personnel and enabling timely intervention. Standoff detection also plays a vital role in various applications including screening at airports, border crossing and critical infrastructure site. The ability to remotely detect explosives, toxic chemicals and narcotics helps mitigate risks associated with these threats and contributes to overall public safety.

The method used in [1] to conduct experiments is safe, both to humans and the environment, and it will be able to differentiate between the explosive and harmless background chemicals. To achieve this purpose, Eye-safe infrared laser interrogation was used coupled with infrared sensors or imaging arrays.

Standoff detection technologies, such as infrared backscatter hyperspectral imaging, have been increasingly adopted for their ability to detect and identify chemical threats without direct contact.

In this paper we proposed an approach that balances detection accuracy and false alarm rate, while improving the decision making process where detection threshold and confidence scores are optimized to minimize false negatives. The key contributions to achieving this task are listed as follows:

- SMOTE technique was applied to generate synthetic samples for the minority class using random numbers. This is done to balance the dataset, using the in-built resampling method. The dataset consists of background samples (majority class) and Threat chemical samples (minority class).
- The balanced dataset was trained using a Classifier (Random Forest) to ensure that the model can effectively learn from a data and generalizes well to new samples.
- The effectiveness, accuracy and precision of this model were evaluated by using a confusion matrix, precision recall F1 score and ROC curve to assess how well the model handles the imbalanced data.

The primary objective of this paper is to develop a robust decision-making framework to address the imbalanced learning problem in the standoff detection of threat chemicals context.

The sections of this paper are organized as follows. Section II presents a review of existing works within the field. In Section III the formulation of the problem is stated. Section IV deals with data preparation and the Methodology used for the experiment. The results of the algorithm implemented in Matlab are found in Section V. Section VI serves as the conclusion while in Section VII the future work is presented.

II. Literature Review

Imbalanced datasets are a pervasive challenge in hyperspectral imaging, largely due to the inherent rarity of certain chemicals of interest. This imbalance skews the data distribution, complicating the accurate detection of minority class instances and leading to suboptimal performance of conventional machine learning models. To mitigate these issues, a variety of methodologies have been proposed.

One widely adopted approach is **oversampling**, which seeks to balance class distribution by increasing the number of minority class samples. A popular oversampling technique is the **Synthetic Minority Over-sampling Technique (SMOTE)**, which generates synthetic samples by interpolating between existing minority class instances [2–4]. An alternative variant, **Adaptive Synthetic (ADASYN)**, focuses on generating additional synthetic data in regions where the minority class is particularly underrepresented and harder to classify [5].

Conversely, **undersampling** techniques aim to balance the dataset by reducing the number of majority class samples. Methods such as **Random Undersampling** randomly remove instances from the majority class [6], while **Tomek Links** identify and eliminate borderline majority class samples that are nearest neighbors to the minority class, thereby reducing class overlap and bias [9–11].

Another effective strategy is the **cost-sensitive learning approach**, which modifies the learning algorithm to assign higher misclassification costs to minority class instances. By penalizing errors more heavily for the minority class, the model is encouraged to focus on accurately classifying these critical cases — a particularly important adjustment in domains where misclassification of the minority class can have severe consequences, such as in medical diagnosis and fraud detection [7].

Additionally, **hyperspectral image analysis** itself offers powerful tools for chemical detection. By capturing detailed spectral information across numerous narrow bands, hyperspectral imaging facilitates the identification of materials and detection of chemical components. Techniques such as spectral unmixing, classification, and anomaly detection leverage these detailed signatures, especially in the infrared range—to accurately quantify chemical substances. These methods have been successfully applied in remote sensing, agriculture, and environmental monitoring, with recent advancements significantly enhancing sensitivity and accuracy [7, 8].

Complementing these data preprocessing strategies, robust decision-making frameworks are essential for addressing imbalanced learning challenges in threat detection. **Bayesian Decision Theory**, for instance, employs Bayes' theorem to incorporate prior knowledge and adjust for the minority class by modifying prior probabilities, thereby reducing bias and enhancing detection of rare events [12, 13]. Similarly, **Decision Tree models**, which partition data based on spectral features, can be augmented with methods such as Convolutional Neural Networks (CNNs) to improve classification accuracy in imbalanced datasets [14].

III. Problem Formulation

Detecting harmful chemicals from a standoff perspective poses significant challenges for sensor and detection systems. In these scenarios, the inherent data imbalance—where the minority class (threat chemicals) is vastly underrepresented compared to the majority (background signals)—often skews the outcomes. Traditional machine learning models struggle with this imbalance, exhibiting issues such as bias toward the majority class, insufficient representation of the minority class, overfitting, and sensitivity to noise and outliers. These limitations lead to a lack of trust in model predictions, ultimately resulting in poor performance and inaccurate decision making [2].

The objective of this work is to develop a robust decision-making framework that effectively addresses the challenges of imbalanced learning in standoff chemical detection. This framework will optimize detection thresholds and confidence scores to enhance the accuracy of identifying the minority class—specifically, threat chemicals—without compromising the overall performance of the detection system. Ultimately, our approach aims to provide a more reliable and trustworthy solution for critical applications where even a small number of false negatives can have severe consequences.

IV. Methodology

We employed a combination of the Synthetic Minority Over-sampling Technique (SMOTE) and a Random Forest Classifier to address the imbalance. Again, synthetic data is used, as actual real hyperspectral data is not available. The methodology and Hyperspectral data are preprocessed in six stages as stated below. For our simulation results we used MATLAB software.

A. STEP 1: Generate Synthesis Hyperspectral Data

Since actual hyperspectral data is not available, synthetic Hyperspectral Data is generated using random numbers. The dataset consists of background samples (majority class) and threat chemical samples (minority class). The threat sample is set to 50, indicating that there are 50 samples of the threat chemicals, while background sample is set at 950, thus bringing the total number of the samples to 1000. Each pixel or observations (Rows) in the image data consists of multiple wavelengths bands (spectral bands) of 100 wavelengths. Furthermore, the labels array is initialized with zero (background) and the first 50 elements of entries are updated to 1 (Threat chemical). Y is also initialized as a column vector of zero, with 1000 elements. From the synthetic hyperspectral data generated, the number of threat chemical samples is much smaller than background samples. This results in an imbalanced dataset.

B. STEP 2: Visualize the Class Distribution (imbalance in the Dataset)

This stage visualizes the imbalanced data class distribution by plotting the histogram as in Fig. 1.

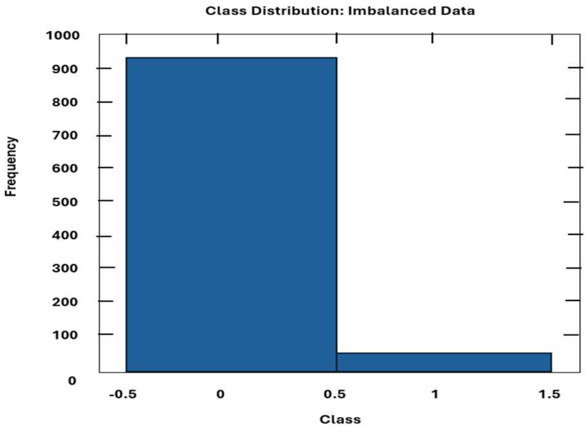

Fig. 1. Histogram representing Class Distribution

The histogram shows frequency of threat classes and background classes by creating bars representing each class label. Y represents the frequency, while X represents each class label.

C. STEP 3: Use SMOTE To Address the Imbalance

We applied SMOTE algorithm to generate synthetic samples for the minority class to balance the dataset, using the in-built resampling method. The threat sample, represented by X, is added to resample the set. Then we initiated a loop (j) in Matlab, that will run 10 times iteration. The variable j will take on value from 1-10 in each iteration. This ensures that each threat sample is oversampled 10 times. Then, noise of about 0.05 scale value is added for synthetic generation. The noise introduces small random variation to the minority threat class, which makes the threat class more realistic and diverse. Finally, most of the class (background samples) are added to the resampled dataset before the two datasets are trained.

D. STEP 4: Train a Classifier (Random Forest)

Leveraging on the balanced dataset created in Step 3, the training on a Random Forest classifier ensures that the model can effectively learn from the data and generalize well to new samples. In training the classifier, we first split the dataset into training sets and testing sets. Splitting the data is a fundamental step in machine learning, crucial for building robust and generalized models. This splitting ensures unbiased evaluation and prevents the overfitting of data. In this experiment, 20% of the data is held out for testing. After splitting the 20%, the test sets are assigned, for example: X_Train and Y_Train are the Training data features and Labels respectively, while X_Test and Y_Test are the Testing data features and Labels respectively. Then, the Fitensemble script implemented in Matlab trains the ensemble, which is built using 100 iterations for classification task purposes. We used Random Forest Matlab Classifier as illustrated in Fig.2.

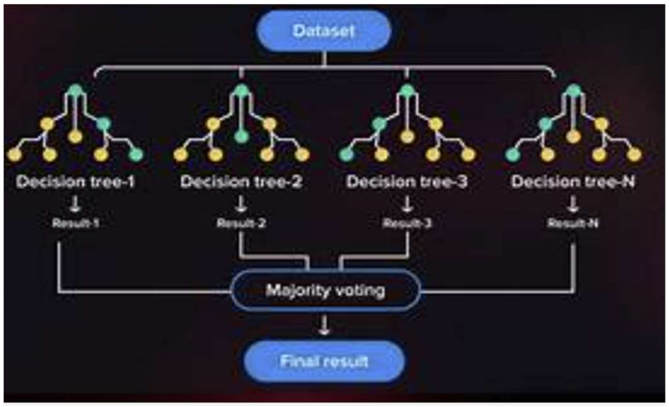

Fig. 2. Random Forest Algorithm

E. STEP 5: Test the Model

After the model has been trained in Step 4 using the training dataset, their performance is evaluated to ensure that the model generalizes well with the new data. Then, the predict function takes the dataset X_ and uses the trained model to generate predictions. These predictions (y_pred) represent the model's best guess for the output based on the test data's features. Then, the model accuracy is calculated by comparing the predictions (y_pred) with the true test labels (y_test). This step test provides an insight into the effectiveness of using Random Forest Classification model. It helps with evaluation and generalization. The Fitcensemble routine is used to train the ensemble.

F. STEP 6: Evaluate Performance

After testing the model, its performance was evaluated using several metrics—namely, a confusion matrix, precision, recall, F1 score, and ROC curve—to assess its ability to handle imbalanced data. The confusion matrix, which details true positives, true negatives, false positives, and false negatives, provides a clear comparison between actual labels and predicted outcomes. Precision is reported with two decimal places to quantify the accuracy of the model's positive predictions, while recall measures the proportion of actual positive cases correctly identified. The F1 score, representing the harmonic mean of precision and recall, offers a balanced overall assessment. Notably, the experiment achieved an F1 score of 1, indicating optimal performance in identifying the minority class.

V. Experimental Results

The results obtained are visualized in the Receiver Operating Characteristics Curve (ROC). The ROC curve plots the True Positive Rate (sensitivity) against the False Positive

Rate (FPR) (1-sensitivity) at various threshold settings. ROC assesses how well the model handled the imbalanced data.

AUC (Area Under Curve) measures the entire two-dimensional area underneath the entire ROC curve from (0 0) to (0 1). It is a single metric summarizing the overall performance of the classification across all thresholds. From the experiment, the AUC=1. This means the model has high accuracy and excellent performance.

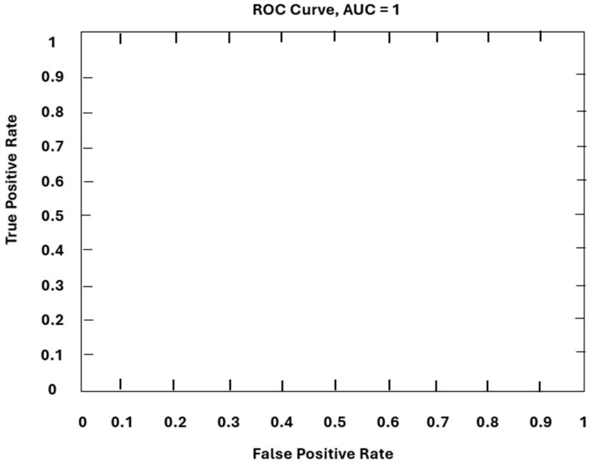

Fig. 3. ROC Curve

$$\text{Confusion Matrix: } \begin{bmatrix} 190 & 0 \\ 0 & 110 \end{bmatrix}$$

TABLE 1: INTERPRETATION OF THE CONFUSION MATRIX

	Predicted Negative	**Predicted Positive**
Actual Negative	190	0
Actual Positive	0	110

True Negative (TN) = 190. The model correctly predicted 190 instances as negative (class 0) which are negative.
False Positive (FP) = 0; The model did not incorrectly predict any negative instances as positive. There are no false positives.
False Negative (FN) = 0; There are no false negatives.
True Positive (TP) = 110; The model correctly predicted 110 instances as positives (class 1) which are positive.
Then, the performance matrix is calculated thus:

Precision: $\frac{TP}{TP + FB} = \frac{110}{110 + 0} = \mathbf{1.00}$

Recall: $\frac{TP}{TP + FN} = \frac{110}{110 + 0} = \mathbf{1.00}$

F1 Score: 2* [(Precision * Recall)
(Precision + Recall)] = **1.00**

Trade Off in Decision Making:

In the context of chemical threat detection, both False Positive Rate (FPR) and False Negative Rate (FNR) are critical metrics that impact the effectiveness and reliability of detection systems. High FPR leads to alert fatigue, where operators become desensitized to alarms, due to frequent false alarms (missing real threats), while High FNR poses a significant safety risk, as undetected threat can lead to exposure to harmful chemicals, resulting in health hazards or environmental damage.

Real-world Applications:

In terms of response times, Standoff detection systems, such as laser-based spectroscopy and radar imagination used for security and surveillance can detect threats from a distance, allowing for a rapid response. Again, the accuracy of standoff detection system used in medical diagnosis for diagnosing either benign or malignant cancerous disease and disease screening are crucial to minimizing false positives and false results. Other real-world applications are in autonomous systems which include self-driving cars (for object detection) and Drones, which use the model for surveillance and threat detection. Banks and Insurance companies also used the model for fraud detection and insurance claims. All these are deployed with ease considering cost, size, and operational environment.

VI. CONCLUSIONS

From the experiment performed, the proposed methodology of using SMOTE and Random Forest classifier has effectively mitigated the challenges posed by class imbalance in the dataset. The result demonstrates outstanding performance, with a Precision of 1.00, Recall of 1.00, and F1 Score of 1.00. These metrics indicate that the framework achieves a perfect identification of threat instances while maintaining no false positive or false negatives. Such exemplary performance underscores the robustness and efficacy of the combined approach in enhancing the detection capabilities of standoff detection systems.

In conclusion, the proposed decision-making framework for addressing the imbalanced learning problem in standoff detection has proven to be a robust and efficient solution, capable of significantly improving the accuracy and reliability of threat detection system.

VII. FUTURE WORK

As the standoff detection challenges keep evolving, which include different types of standoff detection scenarios, various threat types and environment, it is expected that this model be developed further and tested for different data sets. This paper open up to new avenues for research in Advanced Deep learning models and effective imbalanced leaning system for highly overlapped imbalanced classeses involving rare diseases, abnormal behaviour or trace explosive, which can save billions of dollars and human life.

ACKNOWLEDGMENT

This work is supported by the National Science Foundation (NSF) grants # 2401880.

REFERENCES

[1] Jarvis, Jan, Haertelt, Marko, Hugger, Stefan, Butschek, Lorenz, Juergen. "Hyperspectral data acquisition and analysis in imaging and real-time active MIR backscattering spectroscopy" Avanced Optical Technologies, vol. 6, no. 2, 2017, pp. 85-93. https://doi.org/10.1515/aot-2016-0068

[2] C. J. Breshike, C.A. Kendziora, R.Furstenberg, T.J. Huffman, V.K. Nguyen, N.Budack, Y. Yoon, and R.A. McGrill, "Hyperspectral imaging using active infrared backscatter spectroscopy for standoff detection of trace explosives, "Optical Engineering, vol.59,no. 9,p.092009,2020.doi: 10.1117/1.OE.59.9.092009.

[3] Chawla, Nitshe, Bowyer K, Hall L., Kegelmeyer W. "SMOTE: Synthetic Minority Over-Sampling Technique." Journal of Artificial Research, vol.16, 2002, pp.321-357.

[4] C.J. Breshike, C. A. Kendziora, R. Furstenberg, and R. A. McGrill, "Infrared backscatter imaging spectroscopy for standoff detection of trace explosive, "Journal of Applied Physics, vol.125, no.10, p.104901, 2019. doi:10.1063/1.5079622.

[5] C.Liu, J. Li, M.E. Paoletti, J.M. Haut, A. Plaza and Q. Shi, "Accessibility-Free Active Learning for Hyperspectral Image Classification," IGAPSS 2019-2019 IEEE International Geoscience and Remote Sensing Symposium, Yokohama, Japan, 2019, pp. 409-412, doi:10.1109/IGARSS.2019.8897920.

[6] Li, J; Du, Q; Li, Y; Li, W. Hyperspectral Image Classification with Imbalanced Data Based on Orthogonal Complement Subspace Projection. IEEE Trans.Geosci. Remote Sens. 2018, 56,3838-3851.

[7] Sun,T.; Jiao, L.; Feng, J.; Liu, F.; Zhang, x. Imbalanced Hyperspectral Image Classification Based on Maximum Margin. IEEE Geosci. Remote Sens.Lett.2015,12,522-526.

[8] X. Zheng, J. Jia, J. Chen, S.Guo, L. Sun, and Y. Wang, Hyperspectral Image classification with imbalanced data based on semi-supervised learning, "Applied Sciences, Vol. 12, no. 8, p.3943, 2022. doi: 10.3390/app12083943.

[9] K. Pang, Y. Liu, S. Zhou, Y. Liao, Z. Yin, L. Zhao, and H. Chen, "Proto-DS: A self-supervised learning-based nondestructive testing approach for food adultration with imbalanced hyperspectral data, "Foods, vol. 13, no.22, p.3598, 2024. Doi:10.3390/foods13223598.

[10] S. Galli, 'xploring Oversampling Techniques for Imbalanced Datasets, "Training Data's Blog, 2023. Available:https//www.blog.tranindata.com/sampling-techniques-for-imbalanced-data/

[11] M. Mujahid et al., "Data oversampling and imbalanced datasets: an investigation of performance for machine learning and feature engineering, "Journal of Big Data, vol.11, no.1, p.87, 2024, doi: 10.1186/s40537-024-00943-4.

[12] M. G. Villar, E.S. Alvarado, I. D. L. L. T. Diez, "Data oversampling and imbalanced datasets: an investigation of performance for machine learning and feature engineering, "Journal of Big Data, vol. 11, no. 1, p.87, 2024, doi: 10.1186/s40537-024-00943-4.

[13] S. Prasad and J. Chanussot, "Hyperspectral Image Analysis: Advances in Machine Learning and Signal Processing, "Springer, 2020.

[14] Rokach, Lior, and Oded Maimon. "Decision Trees". Data Mining and Knowledge Discovery Handbook, edited by H. Liu and M. Yu, Springer, 2005, pg. 165-192.