IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 1, January 2026

DOI: 10.17148/IJARCCE.2026.151155

Class-imbalance-aware Multiclass Attack Classification Using Random Forest on the Unsw-nb15 Dataset

Dadavali S. P.

Department of Computer Science, Government First Grade College, Kengeri, Bengaluru, India

Abstract: Binary intrusion detection separates normal traffic from malicious traffic, but an operational security system also benefits from identifying the specific attack category associated with an alert. This paper investigates multiclass attack-category classification on the UNSW-NB15 dataset using Random Forest under the dataset's uneven class distribution. A reproducible stratified sample of 20,000 training records and 10,000 testing records was extracted from the supplied training and testing partitions. The target was changed from the binary label used in the earlier studies to the ten-class attack_cat variable, consisting of Normal and nine attack categories. Identifier and binary-label fields were excluded from model inputs to avoid leakage. Numerical attributes were median-imputed and standardized, while categorical attributes were imputed and one-hot encoded. Two Random Forest configurations were evaluated: a standard model and a class-balanced model using balanced_subsample weighting. The standard model achieved 76.13% accuracy and 44.41% macro-F1, whereas the balanced model achieved 75.97% accuracy, 45.04% macro-F1, 77.93% weighted-F1, 46.85% balanced accuracy, and an MCC of 0.686. Class weighting slightly improved macro-level performance and weighted F1 while maintaining comparable overall accuracy. The results also reveal severe difficulty in detecting the rare Analysis, Backdoor, and Worms classes, demonstrating that overall accuracy alone is insufficient for multiclass intrusion-detection evaluation.

Keywords: UNSW-NB15, multiclass classification, intrusion detection, Random Forest, class imbalance, attack categories, network security

I. INTRODUCTION

Machine-learning-based intrusion detection is commonly evaluated as a binary problem in which traffic is assigned to Normal or Attack. The published study associated with this research compared Logistic Regression, Decision Tree, K-Nearest Neighbors, Random Forest, and Support Vector Machine under that binary formulation. Random Forest provided the best overall balance among the evaluated metrics. A subsequent paper examined Random Forest robustness under feature perturbation, feature reduction, class-distribution changes, and repeated sampling. A third paper added an explainability layer using SHAP.

The binary formulation, however, does not identify what type of attack is present. UNSW-NB15 provides an attack_cat field containing Normal and nine attack categories: Fuzzers, Analysis, Backdoors, DoS, Exploits, Generic, Reconnaissance, Shellcode, and Worms [1]. Multiclass classification is therefore a natural extension of the earlier binary work while addressing a different research question.

A second challenge is class imbalance. The rare categories contain far fewer records than Normal and Generic. Consequently, a classifier can obtain relatively strong overall accuracy while performing poorly on minority attack types. This paper therefore evaluates both a standard Random Forest and a class-balanced Random Forest, with macro-F1, balanced accuracy, and MCC reported alongside accuracy and weighted F1.

II. RESEARCH GAP AND OBJECTIVES

The earlier papers established a binary classifier comparison, robustness analysis, and explainability analysis. They did not report a complete multiclass experiment using attack_cat. The robustness paper explicitly identified multiclass attack classification as future work, while the explainability paper also listed extension to the UNSW-NB15 attack categories as future work. This study addresses that specific gap by changing the target definition from Normal/Attack to ten-class attack-category prediction.

- To construct a reproducible ten-class classification problem from the UNSW-NB15 training and testing partitions.

International Journal of Advanced Research in Computer and Communication Engineering
Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 1, January 2026
DOI: 10.17148/IJARCCE.2026.151155

- To evaluate Random Forest without class weighting as a multiclass baseline.
- To evaluate a class-balanced Random Forest using balanced_subsample weighting.
- To determine whether class weighting improves minority-sensitive metrics without relying only on overall accuracy.
- To identify attack categories that remain difficult under the selected experimental configuration.

III. DATASET AND TARGET DEFINITION

The UNSW-NB15 dataset was created at UNSW Canberra using the IXIA PerfectStorm platform and contains normal activities together with contemporary synthetic attack behaviors. The official dataset description lists nine attack types and 49 features including the class label [1]. The present experiments use the supplied CSV training and testing files.

Item	Available	Experimental	Description
Training partition	175,341	20,000	Stratified sample
Testing partition	82,332	10,000	Stratified sample
Original fields	49	45 predictors	id, label, attack_cat excluded from X
Target classes	10	10	Normal + 9 attack categories
Sampling seed	—	42	Fixed for reproducibility

Table I. Dataset and experimental sampling configuration.

Unlike the binary experiments, attack_cat is retained as the target variable. The fields id and label are removed from the predictor matrix. Removing label is necessary because it directly indicates whether a record is normal or an attack and would leak information about the attack_cat target. The id field is an identifier rather than a traffic characteristic.

Class	Training sample	Testing sample
Analysis	228	82
Backdoor	199	71
Dos	1399	497
Exploits	3809	1352
Fuzzers	2074	736
Generic	4562	2292
Normal	6388	4494
Reconnaissance	1197	425
Shellcode	129	46
Worms	15	5

Table II. Stratified sample distribution across the ten target classes.

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 1, January 2026

DOI: 10.17148/IJARCCE.2026.151155

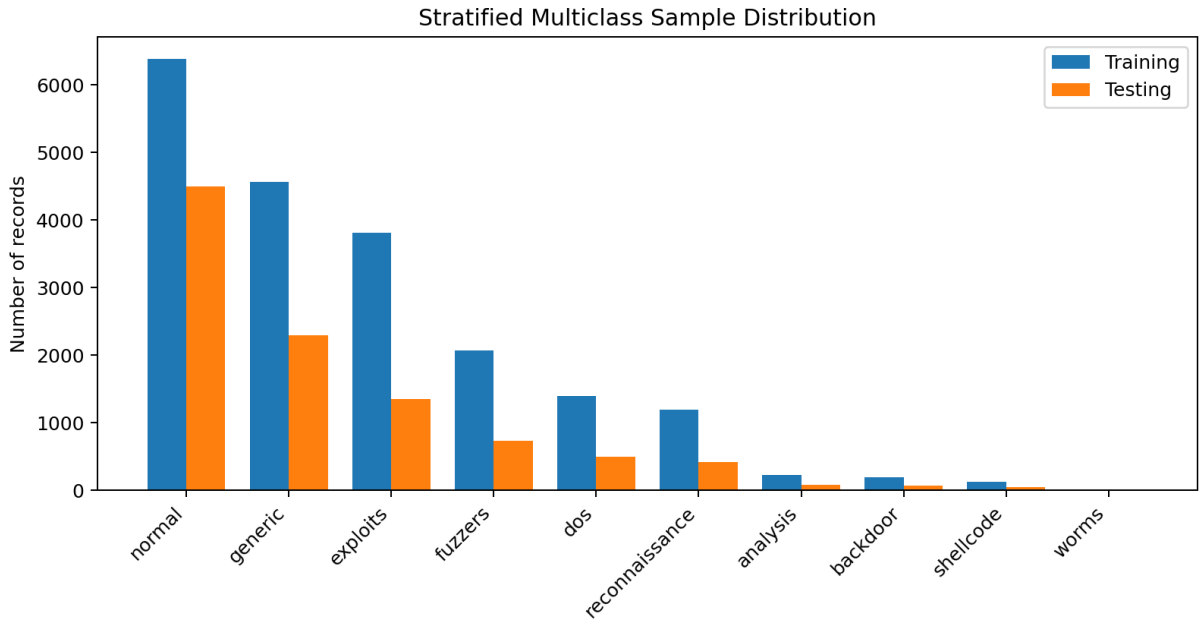

Fig. 1. Distribution of the stratified multiclass training and testing samples.

IV. DATA PREPROCESSING

Numerical features are processed using median imputation followed by standardization. Categorical features are processed using most-frequent imputation followed by one-hot encoding. The preprocessing transformer is fitted only on the training sample and then applied to the testing sample. With the selected data, the transformed representation contains 189 features after excluding id, label, and attack_cat.

Stratified sampling is applied independently to the training and testing partitions with random_state = 42. This preserves the approximate class proportions while making the experiment computationally manageable and comparable with the earlier 20,000/10,000 experimental design.

V. PROPOSED MULTICLASS CLASSIFICATION FRAMEWORK

The framework consists of five stages: dataset loading, stratified multiclass sampling, leakage-aware preprocessing, Random Forest classification, and class-imbalance evaluation. The main methodological change from the earlier binary studies is the use of attack_cat as the target rather than label.

Random Forest constructs an ensemble of decision trees and aggregates their predictions. Two configurations are examined. The first uses the standard unweighted class objective. The second uses class_weight='balanced_subsample', which increases the influence of under-represented classes independently across bootstrap samples. The objective is not to maximize accuracy alone, but to determine whether weighting improves the treatment of minority attack types.

Algorithm 1: Multiclass Class-Imbalance Evaluation

Input: UNSW-NB15 training and testing CSV files.

1. Read the training and testing partitions.
2. Set attack_cat as the ten-class target.
3. Remove id and label from the predictor matrix.
4. Draw 20,000 training and 10,000 testing records using stratified sampling with seed 42.
5. Fit median imputation and standardization on numerical training attributes.
6. Fit most-frequent imputation and one-hot encoding on categorical training attributes.
7. Transform training and testing data using the fitted preprocessing pipeline.
8. Train a standard Random Forest with 100 trees and random_state = 42.
9. Train a second Random Forest with 100 trees and balanced_subsample class weighting.
10. Predict all ten classes on the fixed testing sample.
11. Compute accuracy, precision, recall, F1, macro-F1, weighted-F1, balanced accuracy, and MCC.
12. Compare minority-class performance and inspect the confusion matrix.

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 1, January 2026

DOI: 10.17148/IJARCCE.2026.151155

Output: multiclass performance and class-wise error analysis.

VI. EXPERIMENTAL SETUP AND EVALUATION METRICS

Both Random Forest models use 100 trees, random_state = 42, and parallel training. The only intentional difference is the class-weighting configuration. The same testing sample is used for both models.

Metric	Purpose
Accuracy	Correct predictions divided by total predictions.
Macro-F1	Mean F1 across all ten classes; each class has equal weight.
Weighted-F1	F1 averaged according to the number of samples in each class.
Balanced Accuracy	Mean recall across classes; useful under class imbalance.
MCC	Correlation-based multiclass summary of prediction quality.

Table III. Evaluation metrics used in the multiclass experiment.

VII. RESULTS AND ANALYSIS

Model	Accuracy	Macro-P	Macro-R	Macro-F1	Weighted-F1	Bal. Acc.	MCC
Standard RF	76.13%	44.26%	46.72%	44.41%	77.32%	46.72%	0.689
Balanced RF	75.97%	44.47%	46.85%	45.04%	77.93%	46.85%	0.686

Table IV. Overall multiclass performance.

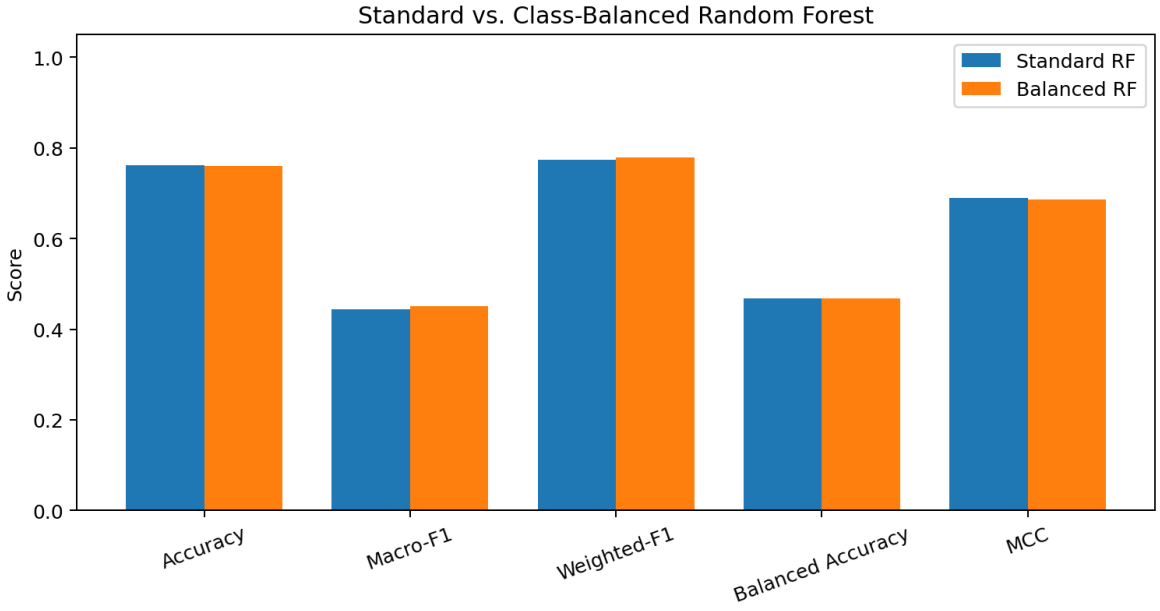

Fig. 2. Comparison of standard and class-balanced Random Forest.

The standard Random Forest obtains 76.13% accuracy and 44.41% macro-F1. The balanced model obtains 75.97% accuracy and 45.04% macro-F1. Class balancing reduces accuracy by 0.16 percentage points but increases macro-F1 by 0.63 points and weighted-F1 by 0.60 points. Balanced accuracy increases from 46.72% to 46.85%. This indicates a modest shift toward minority-sensitive performance rather than a universal improvement in every metric.

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 1, January 2026

DOI: 10.17148/IJARCCE.2026.151155

VIII. CLASS-WISE PERFORMANCE

Class	Support	Precision	Recall	F1
Analysis	82	0.00%	0.00%	0.00%
Backdoor	71	3.55%	7.04%	4.72%
Dos	497	26.11%	28.37%	27.19%
Exploits	1352	63.36%	75.96%	69.09%
Fuzzers	736	30.51%	55.84%	39.46%
Generic	2292	99.91%	96.38%	98.11%
Normal	4494	95.22%	76.75%	84.99%
Reconnaissance	425	81.77%	78.12%	79.90%
Shellcode	46	44.23%	50.00%	46.94%
Worms	5	0.00%	0.00%	0.00%

Table V. Class-wise results for the balanced Random Forest.

The strongest performance is obtained for Generic, with an F1-score of 98.11%, followed by Normal (84.99%) and Reconnaissance (79.90%). Exploits reaches 69.09%, while Fuzzers reaches 39.46%. The minority categories remain difficult: Analysis has zero detected instances, Backdoor has an F1-score of 4.72%, and Worms has zero detected instances. Class weighting alone is therefore not sufficient to solve extreme minority-class scarcity.

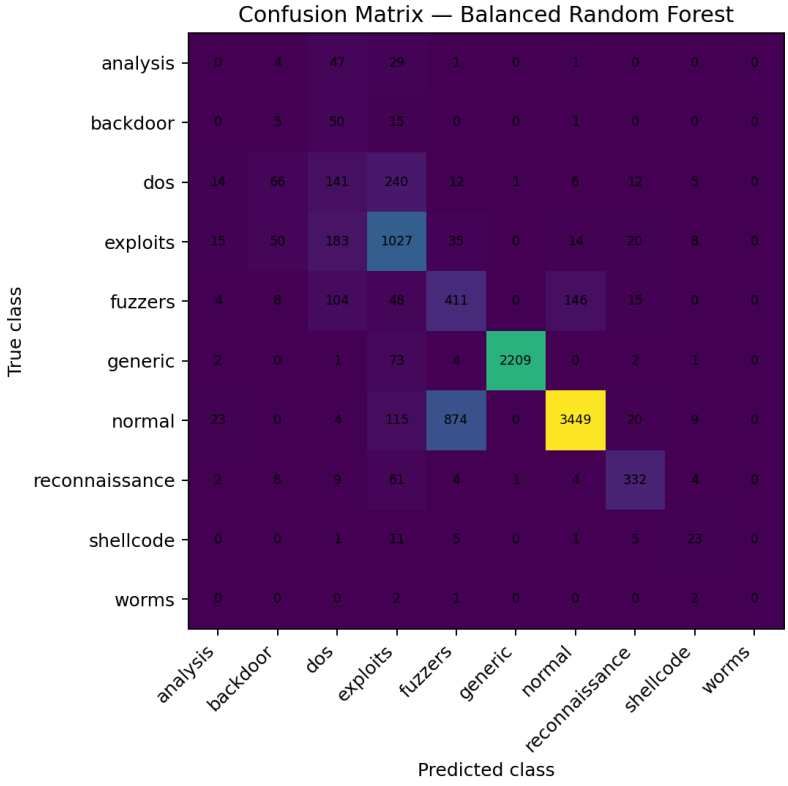

Fig. 3. Confusion matrix for the balanced Random Forest.

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 1, January 2026

DOI: 10.17148/IJARCCE.2026.151155

IX. DISCUSSION

The experiment demonstrates a clear difference between aggregate and class-balanced evaluation. Weighted-F1 remains relatively high because the largest classes, particularly Normal and Generic, are classified more successfully. Macro-F1 is much lower because every class contributes equally and the rare attack categories are poorly recognized. A model can therefore appear strong under accuracy or weighted averages while failing to identify specific low-frequency attacks.

The balanced Random Forest increases recall for DoS from 13.68% to 28.37% and slightly improves F1 for Exploits, Fuzzers, Generic, and Normal. At the same time, the increased emphasis on minority classes produces more confusion among several attack categories. This explains why overall accuracy does not increase. The result should be interpreted as a trade-off rather than a universal improvement.

The experiment also provides a clear boundary for the previous binary studies. The binary label answers "Is this traffic an attack?" whereas attack_cat asks "Which category best describes this traffic?" These are different predictive tasks and should not be compared using the same single accuracy number. The multiclass setting is substantially harder because statistically similar attack behaviors must be separated into specific categories.

X. COMPARISON WITH EARLIER PAPERS

Study	Target	Main focus	Research question
Published Paper 1	Binary Normal/Attack	Five lightweight classifiers	Which classifier performs best?
Paper 2	Binary Normal/Attack	Random Forest robustness	How stable is performance under controlled variations?
Paper 3	Binary Normal/Attack	Random Forest + SHAP	Why does the model make a prediction?
Paper 4	Ten-class attack_cat	Standard vs. class-balanced RF	Can attack categories be distinguished under class imbalance?

Table VI. Distinction among the four studies.

The main distinction of this paper is the target variable and evaluation objective. It does not reuse the binary Normal/Attack target from the earlier papers. It also does not repeat the feature-perturbation robustness experiments or SHAP explainability analysis. Instead, it investigates ten-class attack identification and the effect of class weighting under the same UNSW-NB15 source data.

XI. LIMITATIONS

- The experiments use a 20,000/10,000 stratified sample rather than the complete UNSW-NB15 partitions.
- The Worms testing class contains only five sampled records, so its zero recall should not be interpreted as a stable estimate of real-world Worms detection.
- Only Random Forest is evaluated; other multiclass algorithms may produce different trade-offs.
- Class weighting is tested without synthetic oversampling or specialized cost-sensitive learning.
- The transformed representation contains one-hot encoded categorical variables, and no feature-selection method is applied in this paper.

XII. CONCLUSION

This paper extended the earlier UNSW-NB15 binary intrusion-detection work to ten-class attack-category prediction. Using a reproducible stratified sample and leakage-aware preprocessing, standard and class-balanced Random Forest models were evaluated. The standard model achieved 76.13% accuracy and 44.41% macro-F1, while the balanced model achieved 75.97% accuracy and 45.04% macro-F1. The balanced model therefore provided a small improvement in minority-sensitive aggregate measures, but severe errors remained for Analysis, Backdoor, and Worms. The findings support the use of macro-F1 and balanced accuracy alongside accuracy when evaluating multiclass intrusion-detection systems.

International Journal of Advanced Research in Computer and Communication Engineering
Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 1, January 2026
DOI: 10.17148/IJARCCE.2026.151155

XIII. FUTURE WORK

- Evaluate cost-sensitive learning and controlled oversampling for minority attack categories.
- Investigate specialised strategies for Analysis, Backdoor, Shellcode, and Worms.
- Compare multiclass Random Forest with gradient-boosting models and neural architectures.
- Aggregate results over multiple independent stratified samples to quantify statistical uncertainty.
- Extend the study to additional intrusion-detection datasets and evaluate cross-dataset generalization.
- Combine multiclass classification with the SHAP-based explainability framework

REFERENCES

[1]. N. Moustafa and J. Slay, "UNSW-NB15: A comprehensive data set for network intrusion detection systems," in Proc. 2015 Military Communications and Information Systems Conference (MilCIS), Canberra, Australia, 2015, pp. 1–6, doi: 10.1109/MilCIS.2015.7348942.
[2]. L. Breiman, "Random forests," Machine Learning, vol. 45, pp. 5–32, 2001, doi: 10.1023/A:1010933404324.
[3]. F. Pedregosa et al., "Scikit-learn: Machine learning in Python," Journal of Machine Learning Research, vol. 12, pp. 2825–2830, 2011.
[4]. H. He and E. A. Garcia, "Learning from imbalanced data," IEEE Transactions on Knowledge and Data Engineering, vol. 21, no. 9, pp. 1263–1284, 2009, doi: 10.1109/TKDE.2008.239.
[5]. G. M. Weiss, "Mining with rarity: A unifying framework," SIGKDD Explorations, vol. 6, no. 1, pp. 7–19, 2004.
[6]. V. García, J. S. Sánchez, and R. A. Mollineda, "On the effectiveness of preprocessing methods when dealing with different levels of class imbalance," Knowledge-Based Systems, vol. 25, no. 1, pp. 13–21, 2012.
[7]. T. Fawcett, "An introduction to ROC analysis," Pattern Recognition Letters, vol. 27, no. 8, pp. 861–874, 2006.
[8]. T. Saito and M. Rehmsmeier, "The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets," PLoS ONE, vol. 10, no. 3, 2015.
[9]. "An Approach for the Application of a Dynamic Multi-Class Classifier for Network Intrusion Detection Systems," Electronics, vol. 9, no. 11, 1759, 2020.
[10]. V. Z. Mohale and I. C. Obagbuwa, "Evaluating machine learning-based intrusion detection systems with explainable AI: enhancing transparency and interpretability," Frontiers in Computer Science, 2025, doi: 10.3389/fcomp.2025.1520741.

