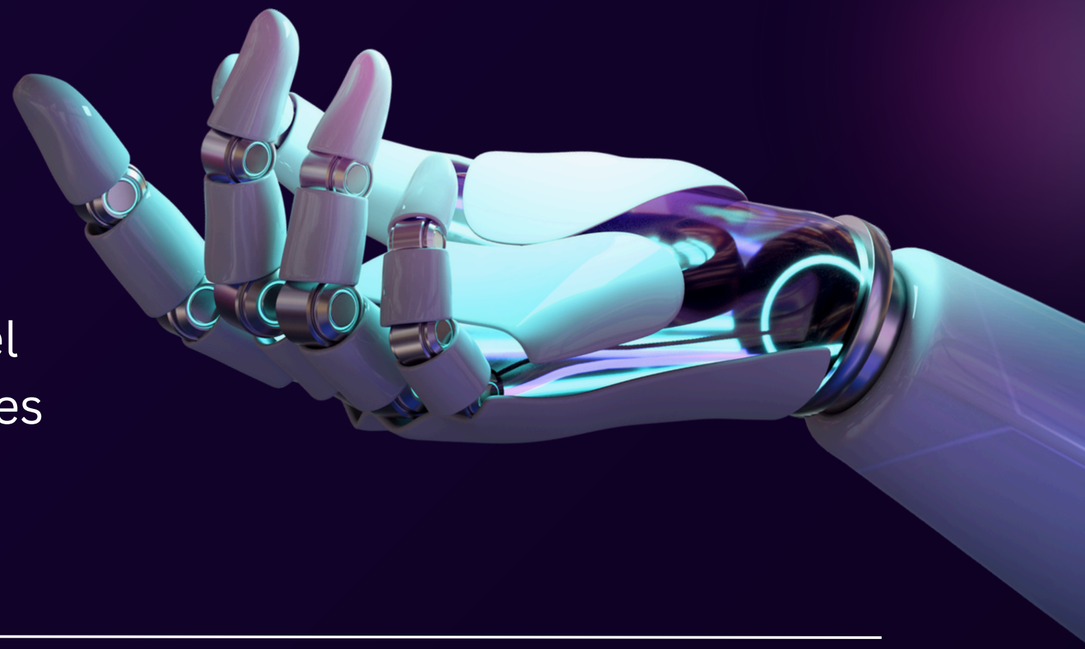
ENHANCING DNS ANOMALY DETECTION USING RANDOM FORESTS



A comprehensive analysis of synthetic data generation, label noise simulation, and hyperparameter optimization techniques in cybersecurity.

01 **Malak Mohamed Osman Mohamed 2205059**

02 **Fadya Hesham Mahmoud ElOraby 2205177**

03 **Abdelrahman Ayman Saad 2205033**

INTRODUCTION

Understanding DNS

01 The Domain Name System (DNS) serves as the backbone of web navigation by converting domain names into IP addresses, facilitating internet connectivity.

Critical Infrastructure

02 DNS is essential for internet functionality, making it a target for various cyber attacks, including cache poisoning.

Role of Anomaly Detection

03 Implementing anomaly detection in DNS is crucial as it helps to identify unusual traffic patterns that can signify potential malicious activities.

Challenges with Real Data

04 Access to real-world DNS traffic data is often limited or biased, which poses challenges in accurately training detection algorithms and understanding typical patterns.

Benefits of Synthetic Data

05 Control anomalies and normal traffic, decide how much traffic will be normal and how much will be anomalous.

Addressing Data Imbalances

06 Synthetic data generation helps overcoming imbalances in datasets.

HINT ABOUT THE GENERATED DATASET.

Simulation of DNS Queries

01 Conducting a simulation of 30,000 DNS queries to gather relevant data for analysis.

TTL (Time to Live)

02 Indicates how long a response is valid, crucial for understanding DNS caching behavior.

Transaction ID

03 A unique identifier for each query, ensuring that responses can be matched to their requests.

Response Codes

04 Categorized into types such as 'No Error,' 'SERVFAIL,' and 'NXDOMAIN,' these codes provide insight into query outcomes.

Anomaly Label

05 Indicates whether the query is normal or anomalous, assisting in the detection of unusual patterns.

Query Volume

06 Counts how often a domain is queried, useful for identifying trends and potential issues.

Anomaly Simulation

07 Randomly introducing anomalies with a 40% probability to create a realistic dataset for testing.

Types of Anomalies

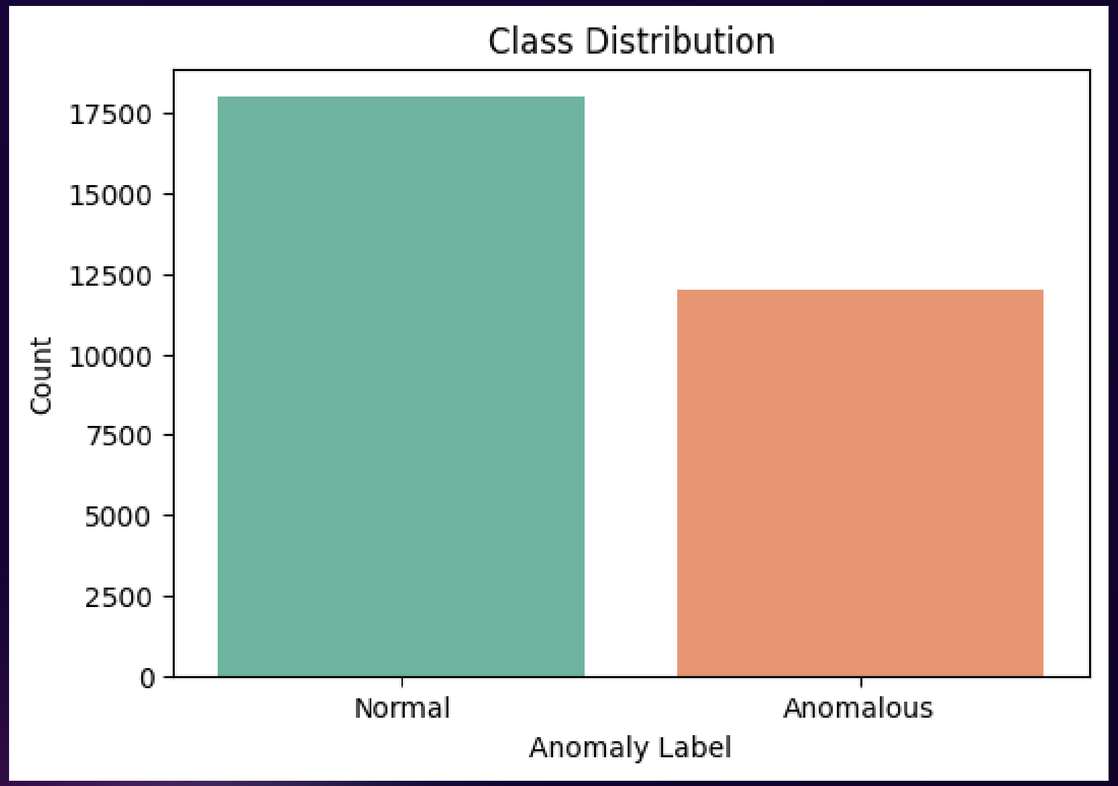
08 Includes low/high TTL values, unusual response codes, and delayed responses to mimic real-world scenarios.

Real-World Traffic Variations

09 Small random variations are added to simulate actual DNS traffic patterns more accurately.

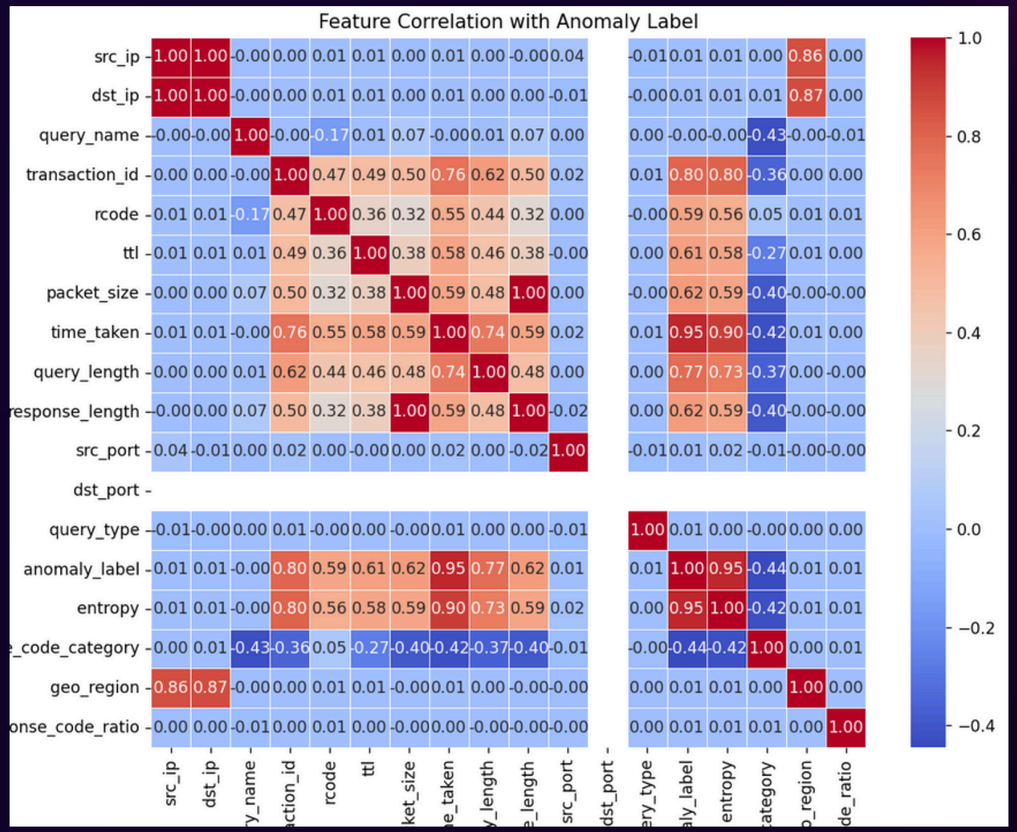
01

Normal 17981
Anomalous 12019
There is no significant class imbalance



02

According to this correlation heatmap, We will exclude features such as “src_ip, dst_ip, query_name, src_port, dst_port, geo_region, and response_code_ratio” as they have minimal impact on the model's decision-making process.



03 Data before scaling

A sample of data before:

	src_ip	dst_ip	query_name	transaction_id	rcode	ttl	...	query_type	anomaly_label	entropy	response_code_category	geo_region	response_code_ratio
0	8.8.8.8	8.8.8.8	test.example.org	0	NXDOMAIN	109	...	MX	Normal	1	Error	US	0.000000
1	1.1.1.1	1.1.1.1	rare.tld	44911	SERVFAIL	50000	...	MX	Anomalous	3	Error	Global	1.000000
2	8.8.4.4	8.8.4.4	my.site.info	0	NXDOMAIN	116	...	MX	Normal	1	Error	US	2.000000
3	8.8.4.4	8.8.4.4	my.site.info	33164	SERVFAIL	1	...	MX	Anomalous	4	Error	US	3.000000
4	8.8.4.4	8.8.4.4	nonexistent.example	30766	REFUSED	1	...	AAAA	Anomalous	4	Error	US	4.000000
5	1.1.1.1	1.1.1.1	example.com	30843	NXDOMAIN	1	...	TXT	Anomalous	4	Error	Global	4.000000
6	8.8.8.8	8.8.8.8	openai.com	24590	SERVFAIL	1	...	TXT	Anomalous	5	Error	US	5.000000
7	1.1.1.1	1.1.1.1	google.com	0	No Error	59	...	A	Normal	1	No Error	Global	6.000000
8	1.1.1.1	1.1.1.1	example.com	49857	REFUSED	50000	...	TXT	Anomalous	5	Error	Global	6.000000
9	1.1.1.1	1.1.1.1	my.site.info	0	NXDOMAIN	56	...	AAAA	Normal	1	Error	Global	6.000000
10	8.8.4.4	8.8.4.4	google.com	57471	REFUSED	1	...	TXT	Anomalous	4	Error	US	7.000000

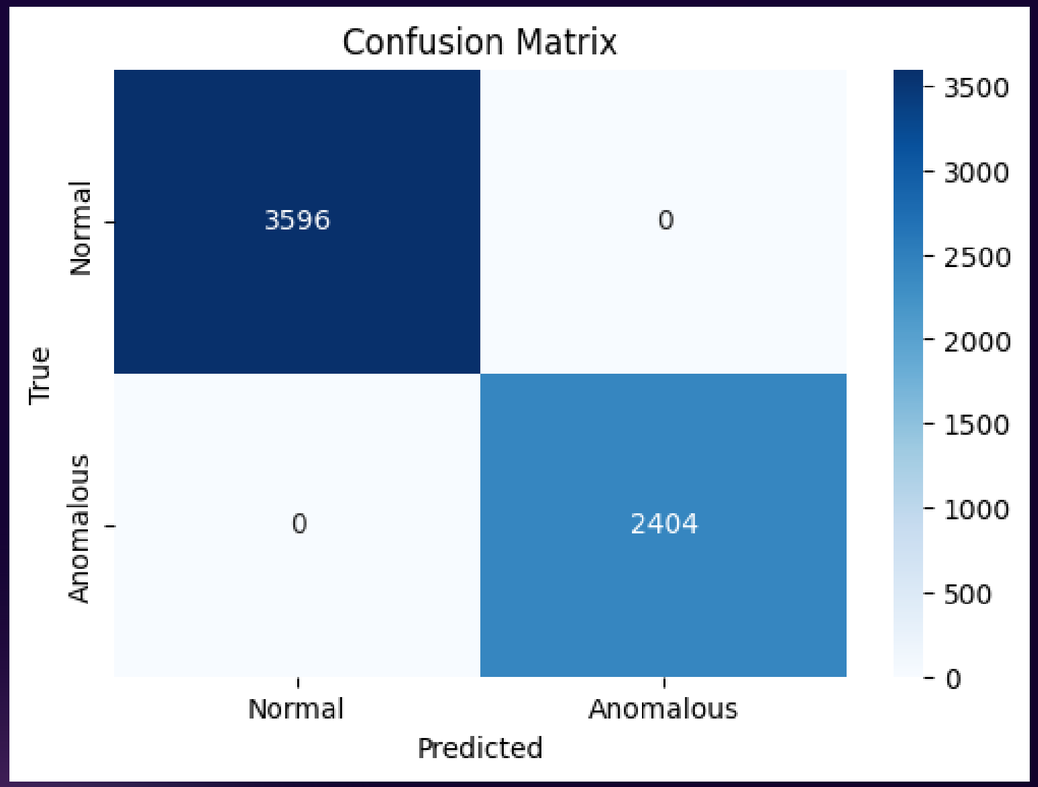
04 After scaling/encoding /dropping low-correlated features with our target

A sample of data after scaling:

	transaction_id	rcode	ttl	packet_size	time_taken	query_length	response_length	query_type	entropy	response_code_category
0	0.000000	0.000000	0.00216	0.213307	0.011582	0.059055	0.214844	0.666667	0.00	0.0
1	0.685298	1.000000	1.00000	0.358121	0.955190	0.799213	0.359375	0.666667	0.50	0.0
2	0.000000	0.000000	0.00230	0.201566	0.134365	0.043307	0.203125	0.666667	0.00	0.0
3	0.506050	1.000000	0.00000	0.078278	0.627755	0.350394	0.080078	0.666667	0.75	0.0
4	0.469459	0.666667	0.00000	0.277886	0.470135	0.244094	0.279297	0.333333	0.75	0.0
5	0.470634	0.000000	0.00000	0.673190	0.641636	0.897638	0.673828	1.000000	0.75	0.0
6	0.375219	1.000000	0.00000	0.125245	0.777480	0.854331	0.126953	1.000000	1.00	0.0
7	0.000000	0.333333	0.00116	0.138943	0.002286	0.035433	0.140625	0.000000	0.00	1.0
8	0.760769	0.666667	1.00000	0.256360	0.713597	0.019685	0.257812	1.000000	1.00	0.0
9	0.000000	0.000000	0.00110	0.219178	0.177154	0.043307	0.220703	0.333333	0.00	0.0
10	0.876951	0.666667	0.00000	0.369863	0.432195	0.787402	0.371094	1.000000	0.75	0.0

05

In our Output:
We observe a problem : overfitting.
Solution: introduce noise.

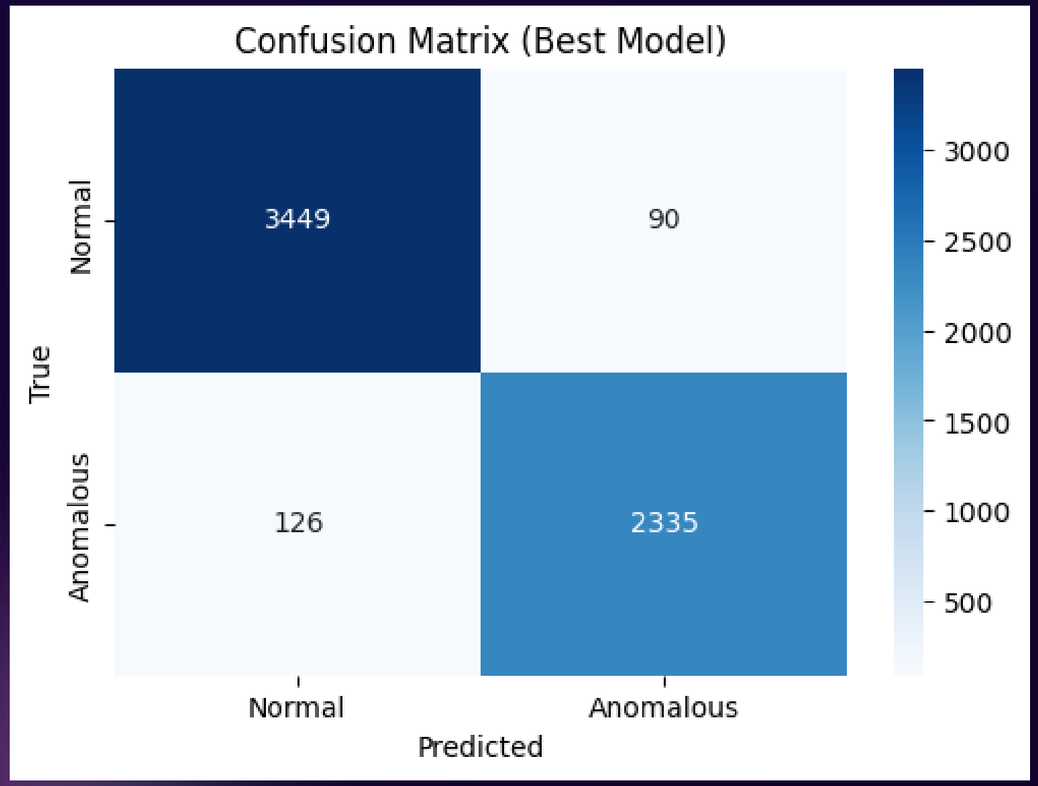



```
Accuracy: 1.0000
Classification Report:
              precision    recall  f1-score   support

           0       1.00      1.00      1.00      3596
           1       1.00      1.00      1.00      2404

    accuracy                           1.00      6000
   macro avg       1.00      1.00      1.00      6000
weighted avg       1.00      1.00      1.00      6000
```

06

Problem Solution using Grid Search and Adding Noise

- Introduce Artificial Noise: To address overfitting, I introduced label noise by flipping a small percentage of labels (0.04) in the dataset.
- Hyperparameter Tuning with Grid Search:
  - Grid Search CV was used to find the best hyperparameters for the model by evaluating all possible combinations in a defined parameter grid.
  - we tuned parameters such as n_estimators, max_depth, and min_samples_split to optimize performance.




```
Best Model Accuracy: 0.9640
Best Hyperparameters:  {'max_depth': 10, 'min_samples_split': 2, 'n_estimators': 100}
Classification Report (Best Model):
              precision    recall  f1-score   support

           0       0.96      0.97      0.97      3539
           1       0.96      0.95      0.96      2461

    accuracy                           0.96      6000
   macro avg       0.96      0.96      0.96      6000
weighted avg       0.96      0.96      0.96      6000
```