Model Research

MODEL RESEARCH BY THREAT TYPE

Threat Type	Dataset	Problem Type	Recommended Models	Key Features / Inputs	Notes
Malware Classification	**EMBER**	Tabular (binary / multiclass)	Random Forest, XGBoost, LightGBM, Autoencoder (for anomaly)	2,381 static PE features (entropy, imports, sections, strings)	Tree models handle sparse + non-linear numeric features well.
Phishing URL Detection	**UCI Phishing Websites**	Tabular (binary)	Logistic Regression, Random Forest, XGBoost, LightGBM	Lexical + host features (URL length, HTTPS, @ symbol, IP in URL)	Feature importance interpretable; can also use ANN for non-linearities.
Spam / Malicious Email	**Enron Email**	NLP (binary / multi-class)	Naive Bayes (baseline), TF-IDF + SVM, LSTM / BERT	Subject, body text, headers	Classical NB for baseline → transformer (BERT) for context semantics.

Threat Type	Dataset	Problem Type	Recommended Models	Key Features / Inputs	Notes
Static Code Vulnerability	**Juliet Test Suite**	Source Code analysis	CNN / RNN over tokens, CodeBERT / Graph Neural Net	Tokenized AST or code embeddings	GNN captures data flows; CodeBERT for semantic understanding.
Malware Image Classification	**Malimg**	Image classification	CNN (ResNet, EfficientNet)	Grayscale malware images from binaries	Use transfer learning for fast training on GPU (Colab).
Intrusion / DDoS / Port Scan	**CICIDS2017, CSE-CIC-IDS2018**	Network flow classification	Random Forest, XGBoost, LSTM, 1-D CNN	80 statistical flow features (bytes, pkts, duration)	Tabular ML for baseline → deep sequence for time features.
IoT / Botnet Threats	**BoT-IoT**	Network traffic classification	LSTM, GRU, CNN-LSTM, Autoencoder	Time-series IoT features (packet rates, flows)	Capture temporal patterns; handle class imbalance.
Network Anomaly	**UNSW-NB15**	Tabular binary / multi-class	Isolation Forest, Autoencoder,	49 network features	Try unsupervised first →

Threat Type	Dataset	Problem Type	Recommended Models	Key Features / Inputs	Notes
			Random Forest	(protocol, src bytes, dst bytes)	supervised with labels.
Real-Time Threat Intel	**AlienVault OTX**	Stream matching / ranking	Rule-based lookup + ML scoring (XGBoost)	IP/domain reputation, frequency, TTL, ASN	Combine threat feed scores + ML confidence.

⚙ Model Families Overview

□ Classical ML (fast baselines)

- **Random Forest / XGBoost / LightGBM:** Best for tabular network or static PE data.
- **Logistic Regression / SVM:** Good interpretable baseline for phishing URLs.

🔡 NLP Models

- **TF-IDF + SVM / Naive Bayes** for spam.
- **Transformer Models (BERT, RoBERTa)** fine-tuned on email or phishing text.

□ Deep Learning Models

- **CNN / ResNet / EfficientNet** — malware image classification.
- **LSTM / GRU / CNN-LSTM** — sequential network traffic.
- **Autoencoder / Variational AE** — unsupervised anomaly detection.
- **Graph Neural Networks (GNN)** — IP-domain-file relationship graphs.

🔧 Preprocessing & Feature Design

Data Type	Steps	Typical Features
Network flows	Standardize numeric features (0–1 scaler), encode categoricals (proto, service)	bytes, packets, duration, ports, ratios, entropy
URLs	Extract lexical tokens, domain age, length, special chars	binary flags, lengths, host entropy
Emails	Clean text, tokenize, TF-IDF / word embeddings	word freqs, topic vectors
Code	Tokenize AST / CFG graphs	function calls, flows, data dependencies
Images	Normalize pixels (0–1) / resize	image arrays