Malicious Activity Detection and Analysis using Machine Learning

:State of the Art and future opportunities

Presented By: Amit Singh

amitsingh@cert-in.org.in/singh.amit90@gov.in

LinkedIn: https://www.linkedin.com/in/amit-singh-11845437

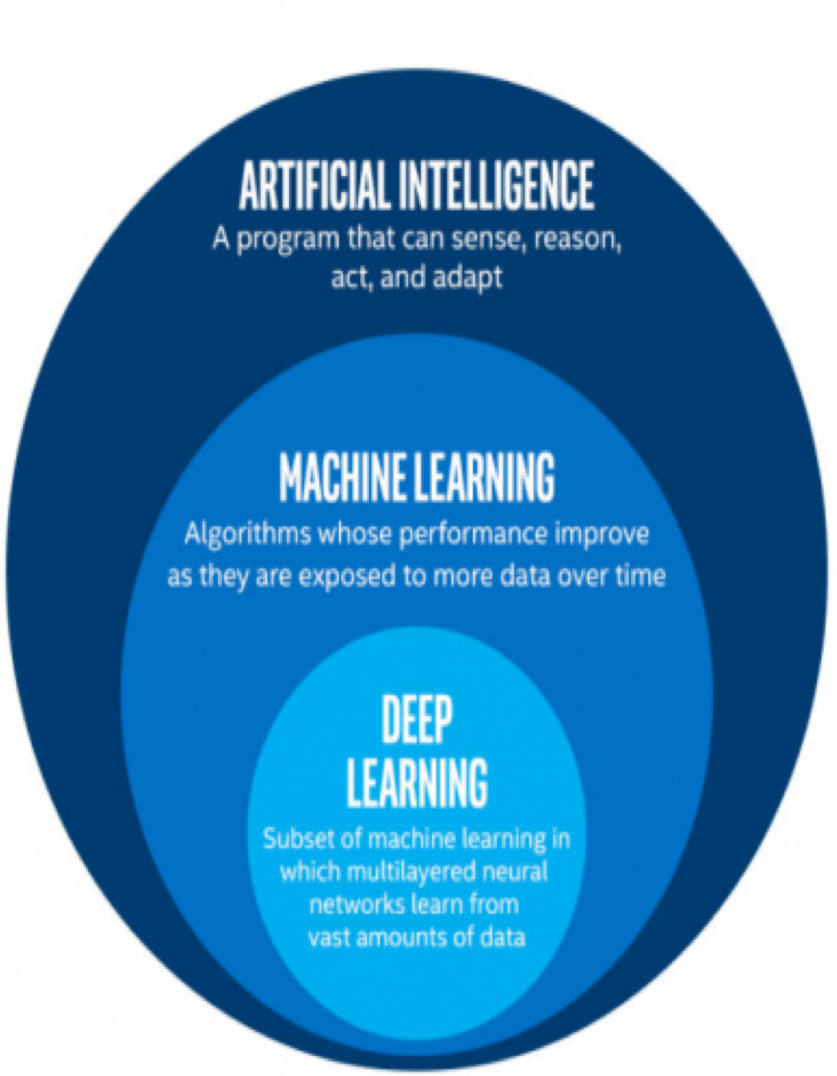
Outline

- Machine Learning
- Classification of Machine Learning algorithms
  - Classical ML vs. Deep Learning
- Performance Evaluation of ML based approaches
- ML meets Cyber Security
- Applications of Machine Learning algorithms in Malicious Activity Detection
  - Spam/Phishing detection
  - Anomaly Intrusion Detection
  - Malware detection and classification
- Future Direction and Challenges

Machine Learning

- A branch of **artificial intelligence**, concerned with the design and development of algorithms that allow computers to evolve behaviors based on empirical data.

- As intelligence requires knowledge, it is necessary for the computers to acquire knowledge.



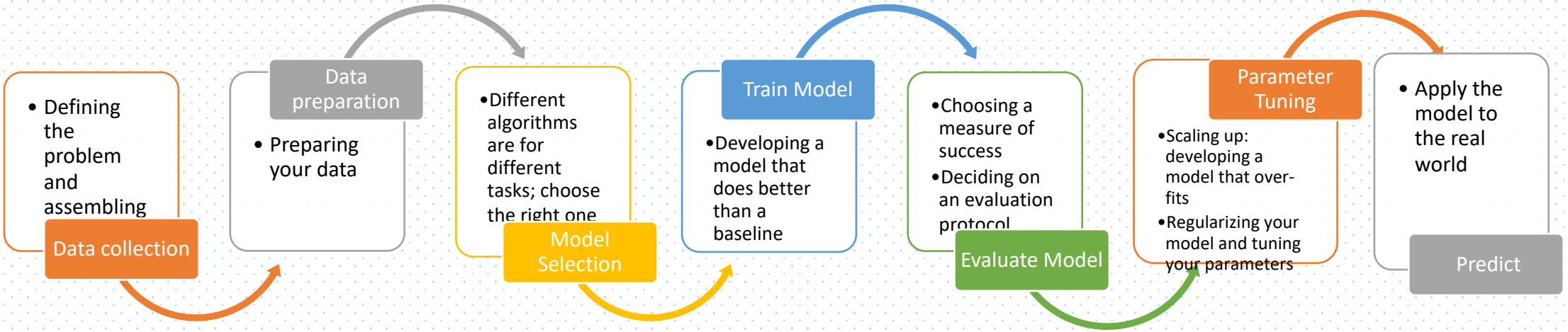
Source: *Difference between AI, Machine Learning & Deep Learning*

Machine Learning

- “A computer program is said to learn from experience E with respect to some class of tasks T and performance measure P if its performance at tasks in T, as measured by P, improves with experience E ”- **Tom Mitchell**

- “Machine learning is a type of artificial intelligence (AI) that provides computers with the ability to learn without being explicitly programmed. Machine learning focuses on the development of computer programs that can teach themselves to grow and change when exposed to new data.” - **Google’s definition**

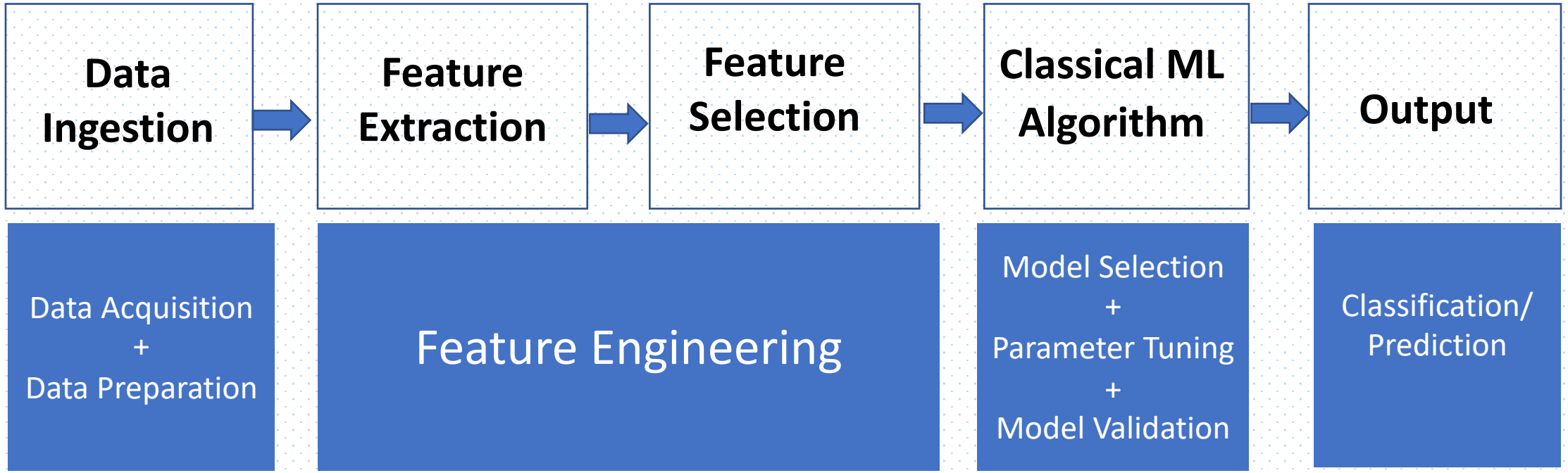
The Machine Learning Process



Taxonomy of ML Approaches

- **Shallow Learning/Classical Machine Learning**
  - **Supervised Learning**
    - Naive Bayes
    - Logistic
    - regression
    - Support Vector Machines
    - Random Forests
    - Shallow Neural Networks
  - **Unsupervised Learning**
    - Clustering
    - Association Rule Mining
- **Deep Learning**
  - **Supervised Learning**
    - Feed forward NN
    - Recurrent NN
    - Convolutional NN
  - **Unsupervised Learning**
    - Stacked auto encoders
    - Deep belief Networks

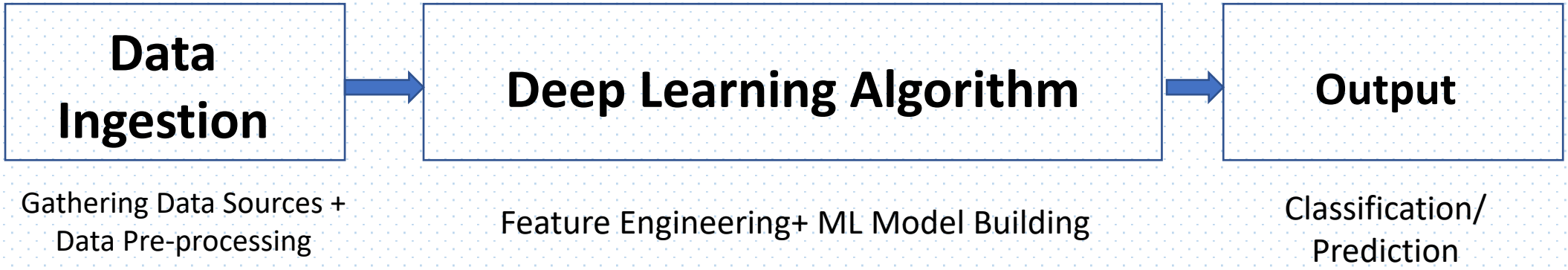
Classical ML



Traditional ML-

- Works better on small data
- Financially and computationally cheap
- Easier to interpret

Deep learning

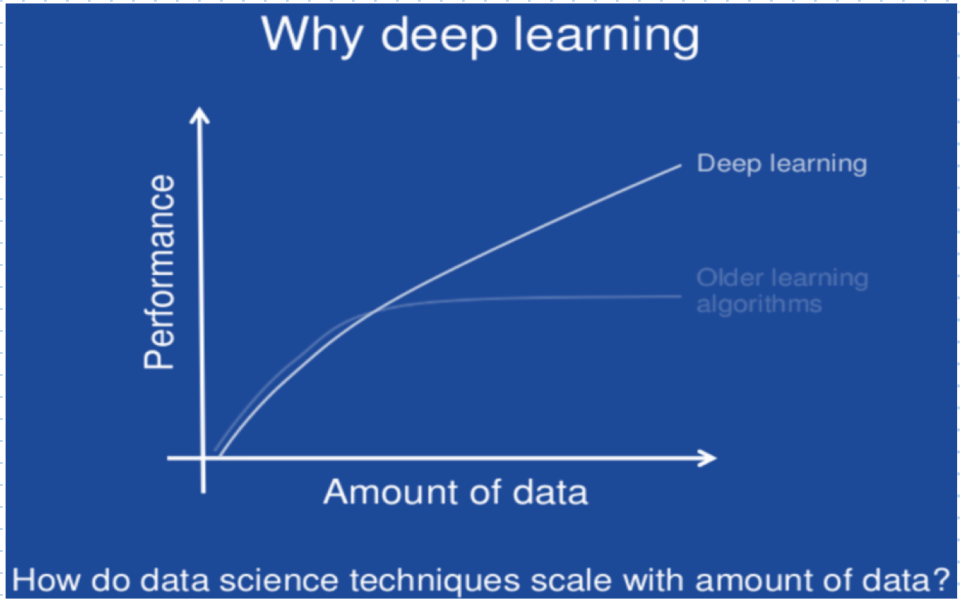


"Deep learning is a particular kind of machine learning that achieves great power and flexibility by learning to represent the world as nested hierarchy of concepts, with each concept defined in relation to simpler concepts, and more abstract representations computed in terms of less abstract ones."

Deep learning promises

- Best-in-class performance
- Scales effectively with data
- ‘Auto’ feature extraction
- High accuracy of detections
- Adaptable and transferable

Deep learning vs Classical ML



- Deep Learning is more hardware intensive
- Deep Learning provides highly accurate results but takes much longer to train
- Deep Learning is perfect for dealing with high volumes of unstructured data
- Deep Learning does not require a domain expert to specify each and every function

Performance Evaluation of ML based approaches

		Actual value	
Predicted value		Positive (1)	Negative(0)
	Positive (1)	TP (True positive)	FP (False positive)
	Negative (0)	FN (False Negative)	TN (True Negative)

- True positive when model predicted true and it is true
- True Negative when model predicted false and it is false.
- False positive when model predict true but it is false (**Type 1 error**)
- False Negative when model predicted false but it is true. (**Type 2 error**)

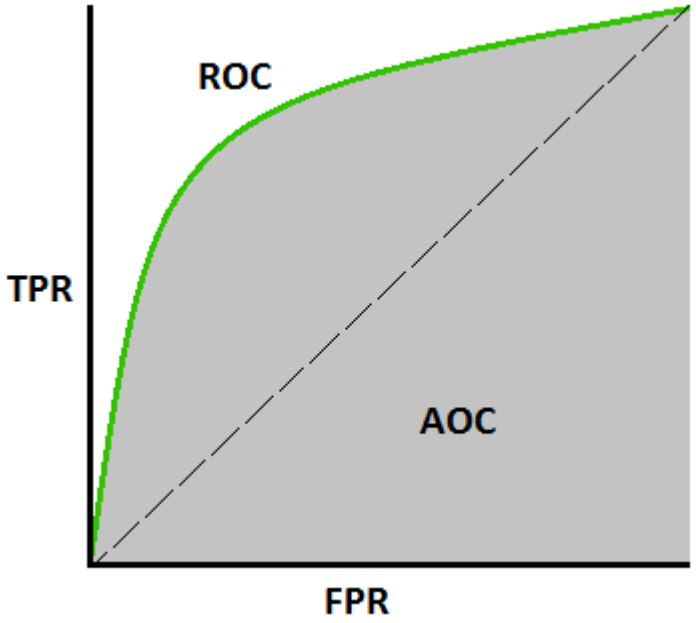
Performance Indicators

Metrics	Formula
$Accuracy$	$\frac{TP + TN}{TP + TN + FP + FN}$
Error rate	$\frac{FP + FN}{TP + TN + FP + FN}$
Sensitivity	$\frac{TP}{TP + FN}$
Specificity	$\frac{TN}{TN + FP}$
$Recall$	$\frac{TP}{TP + FN}$
$Precision$	$\frac{TP}{TP + FP}$
$F1 - Score$	$\frac{2 * Recall * Precision}{Recall + precision}$
$Recall$	$\frac{TP}{TP + FN}$

AUC - ROC Curve

AUC - ROC curve is a performance measurement for classification problem at various thresholds settings. ROC is a probability curve and AUC represents degree or measure of separability. It tells how much model is capable of distinguishing between classes. Higher the AUC, better the model is at predicting 0s as 0s and 1s as 1s.

The ROC curve is plotted with TP against the FP where TP is on y-axis and FP is on the x-axis.



Cyber Security

Cybersecurity is the body of technologies, processes and practices designed to protect networks, computers, programs and data from attack, damage or unauthorized access. In a computing context, **security** includes both **cybersecurity** and physical **security**.

ML meets Cyber Security

Why use ML in cybersecurity?

- ML is now considered crucial to the role of cyber security
- Protect sites from attacks and identify quickly new threats
- It can speed up the process of noticing attacks
- Beneficial in preventing a full-scale breach being unleashed

Machine learning specifics in cybersecurity

- Large representative datasets
- Interpretable model
- Low False positive rates
- Adaptable Algorithms to tackle evolving malware counter measures

Applications of Machine Learning algorithms in Malicious Activity Detection

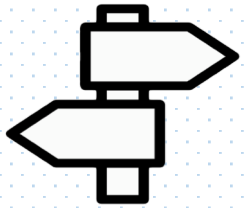
- Botnet Detection and classification
- Anomaly Intrusion Detection

  Examples
  – Malicious JavaScript and other scripts
  – Malicious Non-Executable Files
  – Malicious Executable Files
- Malware detection and classification
- Inappropriate Web and Email Content
- Spam/Phishing detection
  – Derive probabilistic models of phishing attacks
  – Derive probabilistic reputation models for URLs

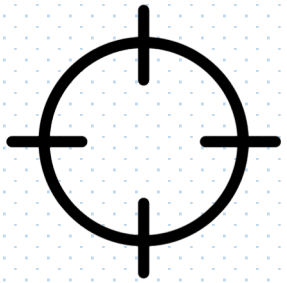
Spam/Phishing detection

PHISHING IS A TYPE OF SOCIAL ENGINEERING ATTACK USED TO STEAL SENSITIVE INFORMATION SUCH AS PASSWORDS OR FINANCIAL DETAILS

ATTACKERS PRETEND TO BE A TRUSTED ENTITY TO PUSH VICTIMS INTO OPENING FRAUDULENT LINKS OR ATTACHMENTS



THIS IS A GENERIC ATTACK USING COMMON MESSAGES THAT MAY BE RELEVANT TO THE VICTIMS CONTRIBUTING TO THEIR FALSE SENSE OF TRUST

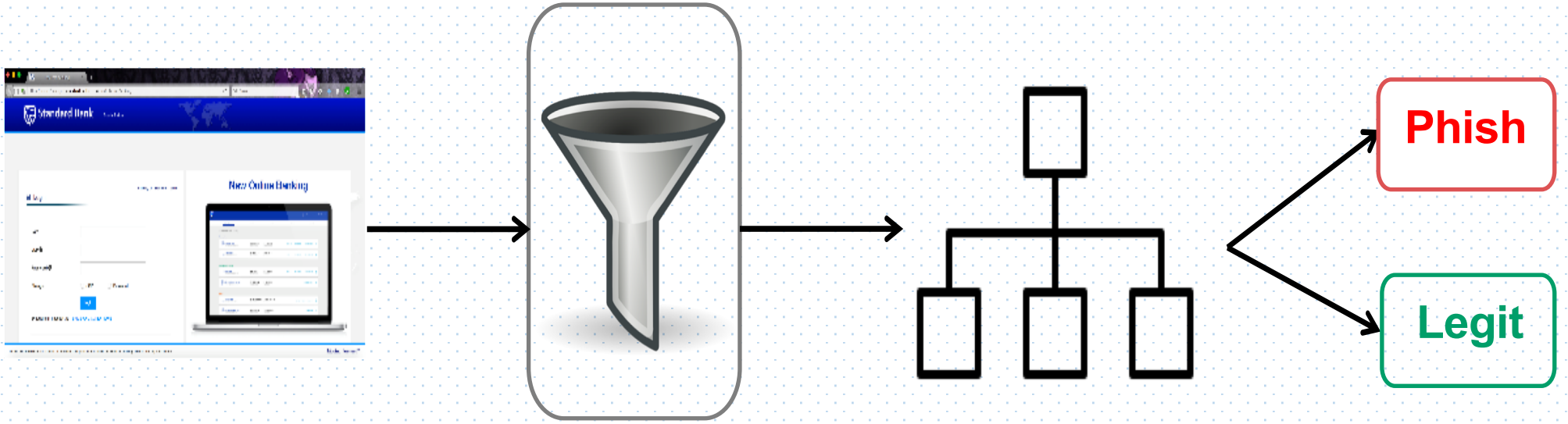


SPEAR PHISHING IS AN ADVANCED TYPE OF SOCIAL ENGINEERING ATTACK USED TO STEAL SENSITIVE INFORMATION SUCH AS PASSWORDS OR FINANCIAL DETAILS.

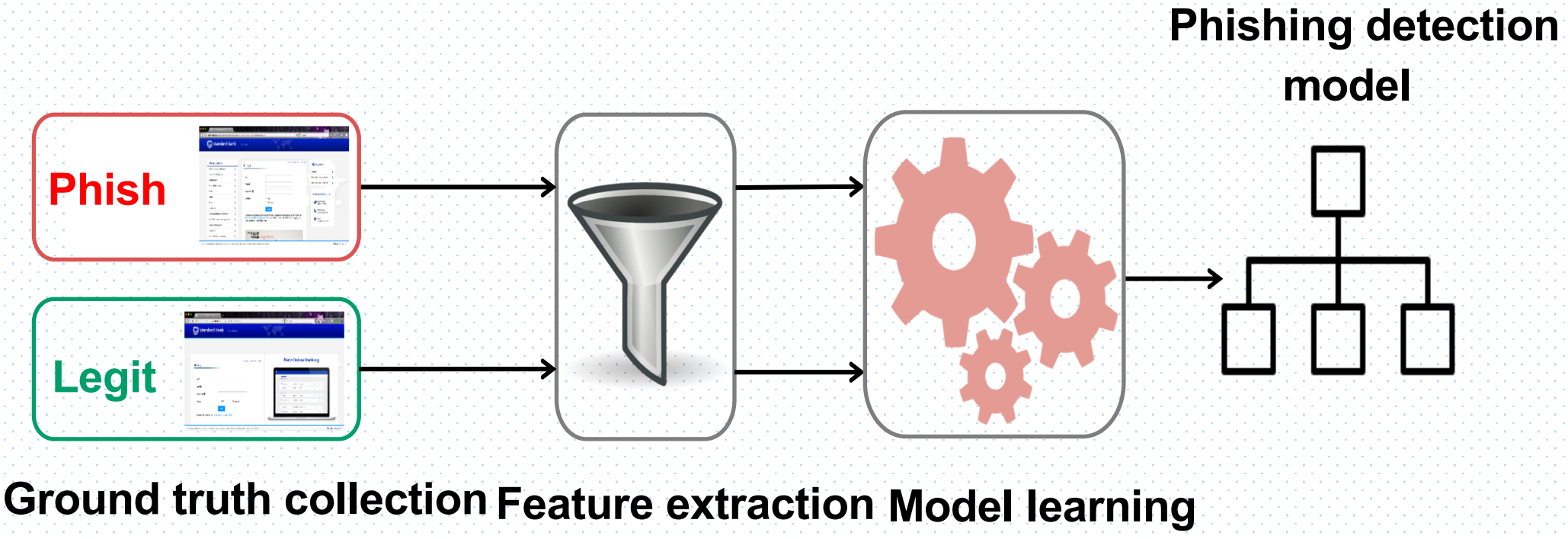
ATTACKERS PRETEND TO BE A TRUSTED ENTITY TO PUSH VICTIMS INTO OPENING FRAUDULENT LINKS OR ATTACHMENTS

THIS IS A VERY FOCUSED ATTACK USING SPECIFIC MESSAGES WITH PERSONAL AND RELEVANT INFORMATION TO THE VICTIMS INCREASING THEIR FALSE SENSE OF TRUST

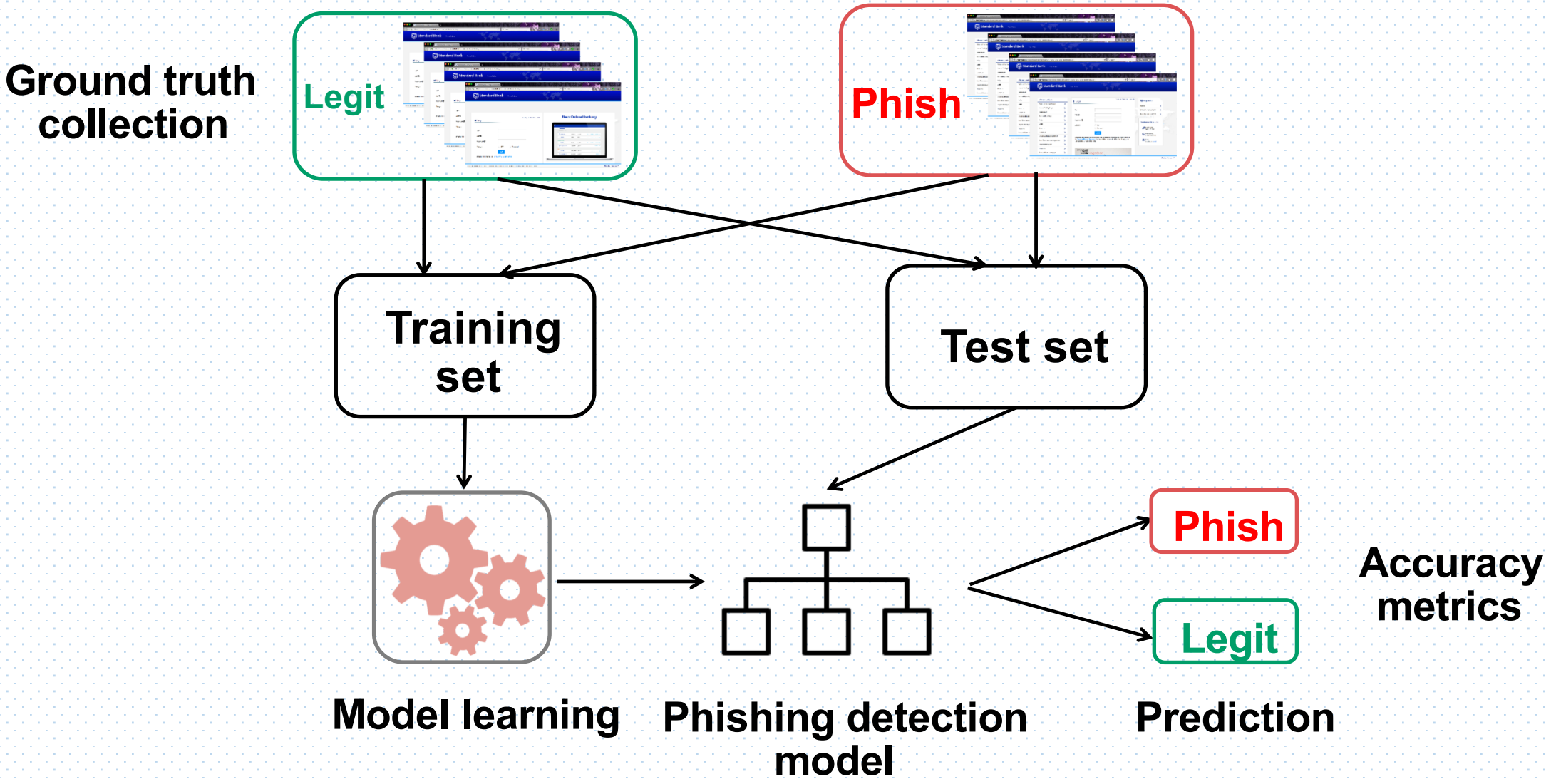
Machine learning based phishing detection



Phishing detector training



Evaluation setup



Example Features

Number of Dots	SubdomainLevel	PathLevel
Url Length	NumDash	NumDashInHostname
At Symbol	TildeSymbol	NumUnderscore
NumPercent	NumQueryComponents	NumAmpersand
NumHash	NumNumericChars	NoHttps
RandomString	IpAddress	DomainInSubdomains
DomainInPaths	HttpsInHostname	HostnameLength
PathLength	QueryLength	DoubleSlashInPath
NumSensitiveWords	EmbeddedBrandName	PctExtHyperlinks
PctExtResourceUrls	ExtFavicon	InsecureForms
RelativeFormAction	ExtFormAction	AbnormalFormAction
PctNullSelfRedirectHyperlinks	FrequentDomainNameMismatch	FakeLinkInStatusBar
RightClickDisabled	PopUpWindow	SubmitInfoToEmail
IframeOrFrame	MissingTitle	ImagesOnlyInForm
SubdomainLevelRT	UrlLengthRT	PctExtResourceUrlsRT
AbnormalExtFormActionR	ExtMetaScriptLinkRT	PctExtNullSelfRedirectHyperlinksRT

For our experimental work the dataset was obtained (Choon Lin Tan, 2018). In this Phishing, datasets, there are 48 features and 10,000 instances.

Classifier	Precision value for Feature selection & Dimensionality Reduction						
	Information gain feature selection	Select from model	Recursive feature eliminatio n	Univariate Feature selection	Variance Reduction Feature selection	Principal componen t analysis	L1 (Lasso) Regulariza tion
MultinomialNB	0.693	0.858	0.854	0.738	0.687	0.576	0.707
GaussianNB	0.745	0.936	0.920	0.934	0.824	0.718	0.954
Logistic Regression	0.838	0.927	0.919	0.875	0.876	0.739	0.940
Random Forest	**0.884**	**0.982**	**0.983**	**0.969**	**0.940**	**0.847**	**0.968**
AdaBoost	0.834	0.973	0.966	0.957	0.916	0.794	0.954
XGBoost	0.856	0.974	0.969	0.963	0.922	0.807	0.968
KNeighbors	0.766	0 .963	0.944	0.920	0.821	0.812	0.818
SVM	0.828	0.923	0.906	0.864	0.870	0.694	0.934
DecisionTree Classifier	0.842	0.966	0.971	0.957	0.891	0.801	0.941

Classifier	Recall value for Feature selection & Dimensionality Reduction						
	Information gain feature selection	Select from model	Recursive feature elimination	Univariate Feature selection	Variance Reduction Feature selection	Principal component analysis	L1 (Lasso) Regularizati on
MultinomialNB	0.871	0.748	0.738	0.928	0.893	0.922	0.910
GaussianNB	0.870	0.769	0.735	0.739	0.876	0.861	0.696
Logistic Regression	0.886	0.928	0.925	0.888	0.862	0.832	0.927
Random Forest	0.885	**0.980**	**0.975**	**0.975**	**0.938**	0.888	**0.958**
AdaBoost	0.902	0.979	0.956	0.965	0.933	0.867	**0.958**
XGBoost	**0.915**	**0.980**	0.972	0.966	0.933	0.889	0.949
KNeighbors	0.890	0.952	0.937	0.914	0.879	0.896	0.856
SVM	0.908	0.934	0.927	0.930	0.880	**0.931**	0.936
DecisionTreeClassifier	0.821	0.954	0.971	0.960	0.883	0.826	0.949

Classifier	Accuracy for Feature selection & Dimensionality Reduction						
	Information gain feature selection	Select from model	Recursive feature elimination	Univariate Feature selection	Variance Reduction Feature selection	Principal component analysis	L1 (Lasso) Regularizati on
Multinomial NB	73.600	80.733	80.100	79.433	73.700	62.733	76.033
GaussianNB	78.067	85.467	83.167	83.967	84.067	76.433	82.700
Logistic Regression	85.433	92.600	92.033	87.800	86.733	77.300	93.267
Random Forest	**88.200**	**98.067**	**97.833**	**97.133**	**93.767**	**86.567**	**96.200**
AdaBoost	86.433	97.567	96.000	95.967	92.167	82.333	95.467
XGBoost	87.800	97.667	96.967	96.367	92.533	84.000	95.767
KNeighbors	80.500	95.700	94.500	91.533	84.066	84.700	82.933
SVM	85.700	92.700	91.400	88.960	87.166	76.366	93.366
DecisionTreeClassifier	82.933	95.967	97.000	95.800	88.333	81.300	94.300

Classifier	F1-Sore value for Feature selection & Dimensionality Reduction						
	Information gain feature selection	Select from model	Recursive feature elimination	Univariate Feature selection	Variance Reduction Feature selection	Principal component analysis	L1 (Lasso) Regularizatio n
Multinomial NB	0.772	0.799	0.792	0.822	0.777	0.709	0.796
GaussianNB	0.803	0.844	0.817	0.825	0.849	0.783	0.805
Logistic Regression	0.861	0.927	0.922	0.881	0.869	0.783	0.933
Random Forest	0.884	**0.981**	**0.979**	**0.972**	**0.939**	**0.867**	**0.963**
AdaBoost	0.867	0.976	0.961	0.961	0.924	0.829	0.956
XGBoost	**0.885**	0.977	0.970	0.964	0.927	0.846	0.958
KNeighbors	0.823	0.952	0.940	0.917	0.849	0.852	0.837
SVM	0.866	0.928	0.916	0.896	0.875	0.795	0.935
DecisionTree Classifier	0.831	0.960	0.971	0.958	0.887	0.813	0.945

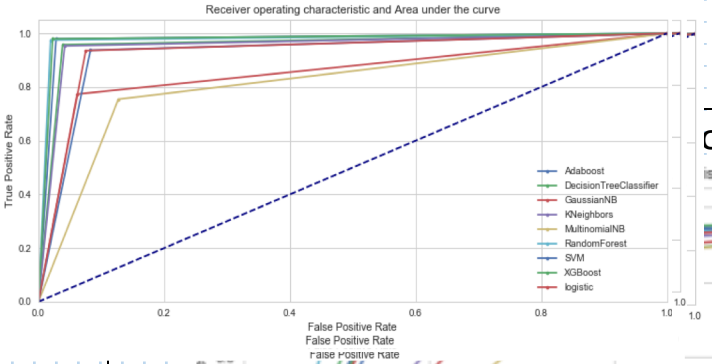


Figure 1 AUC-ROC curve on Information gain

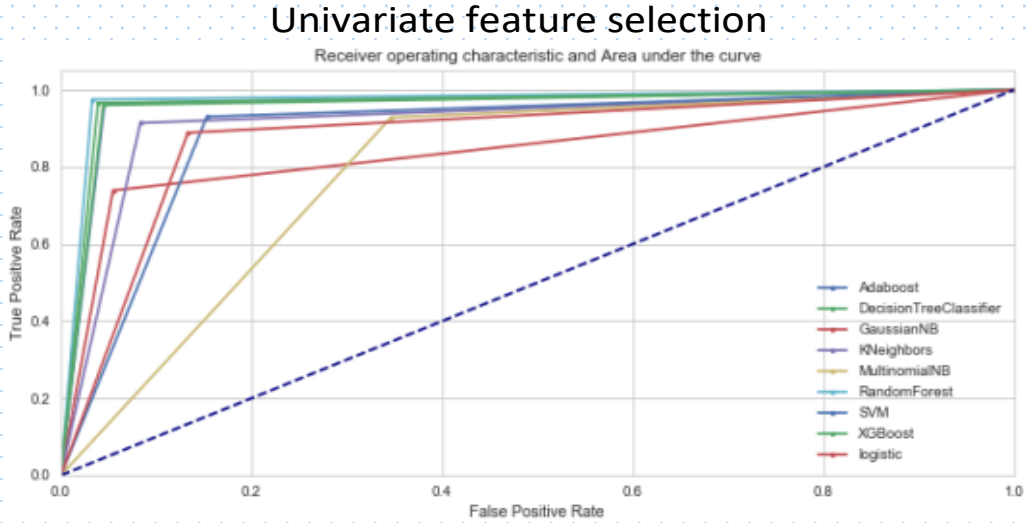


Figure 2 AUC-ROC curve on Univariate feature selection

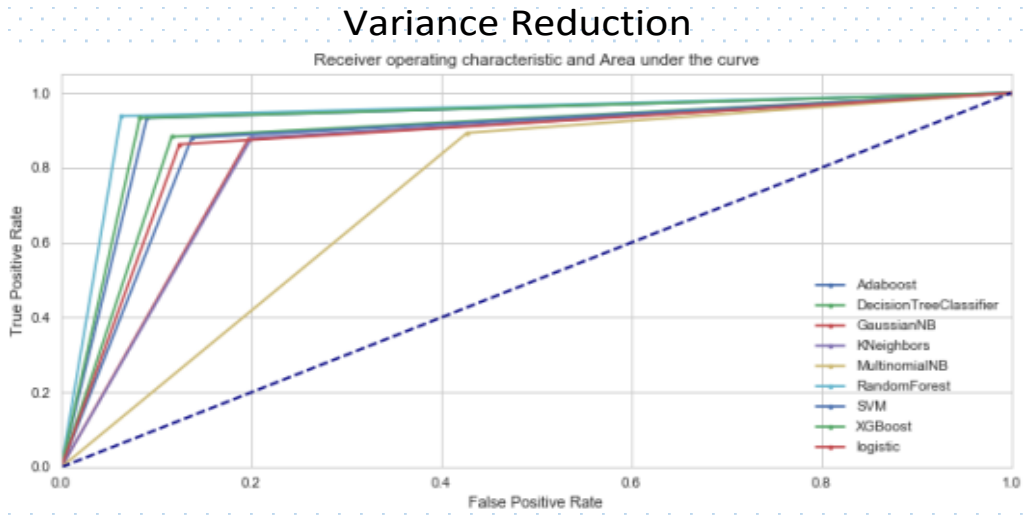


Figure 3 AUC-ROC curve on Variance reduction

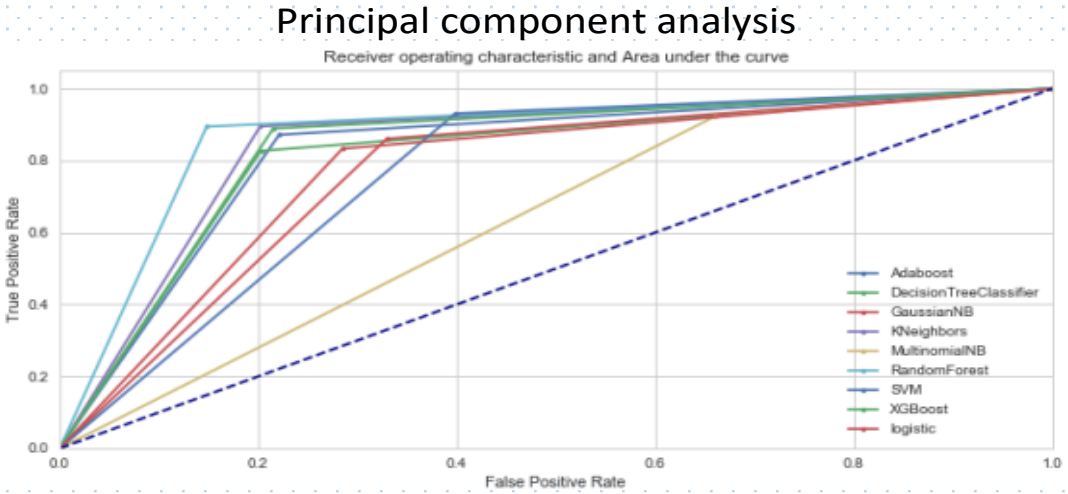


Figure 4 AUC-ROC curve on PCA

Recursive feature elimination

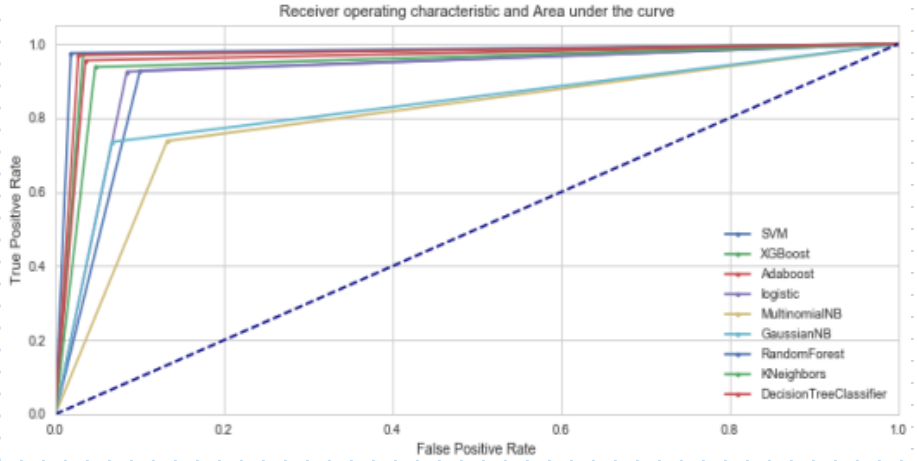


Figure 1 AUC-ROC curve on Recursive feature elimination

Lasso feature selection

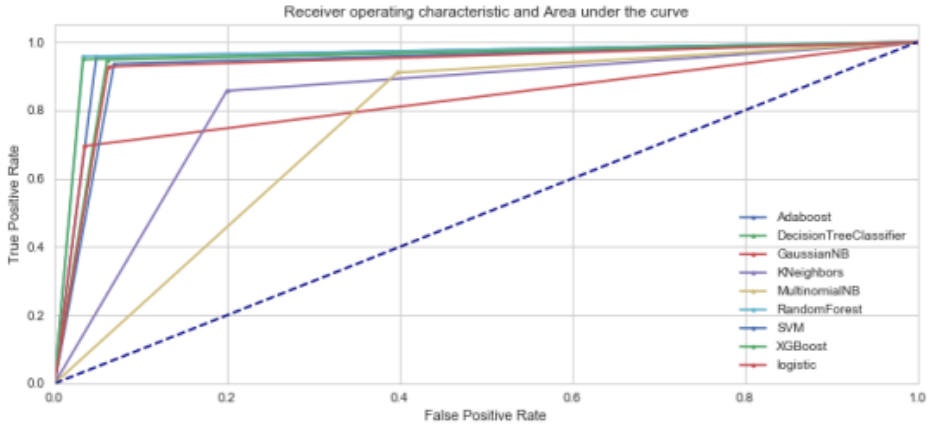


Figure 2 AUC-ROC curve on Lasso feature selection

Select from model

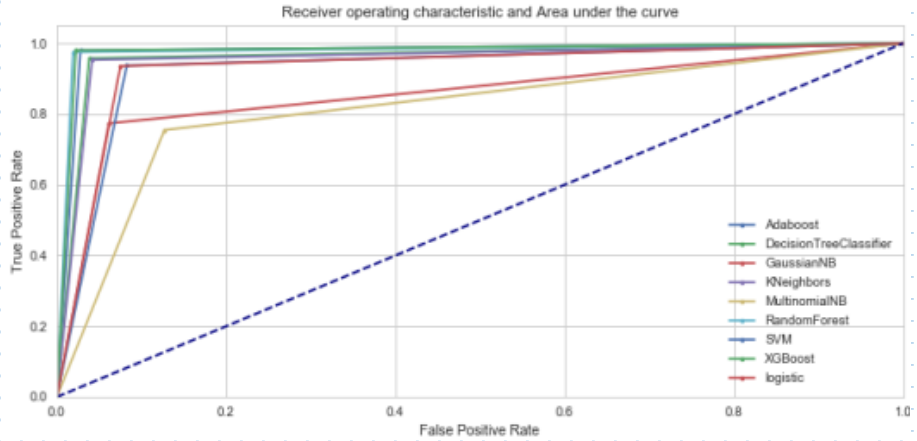
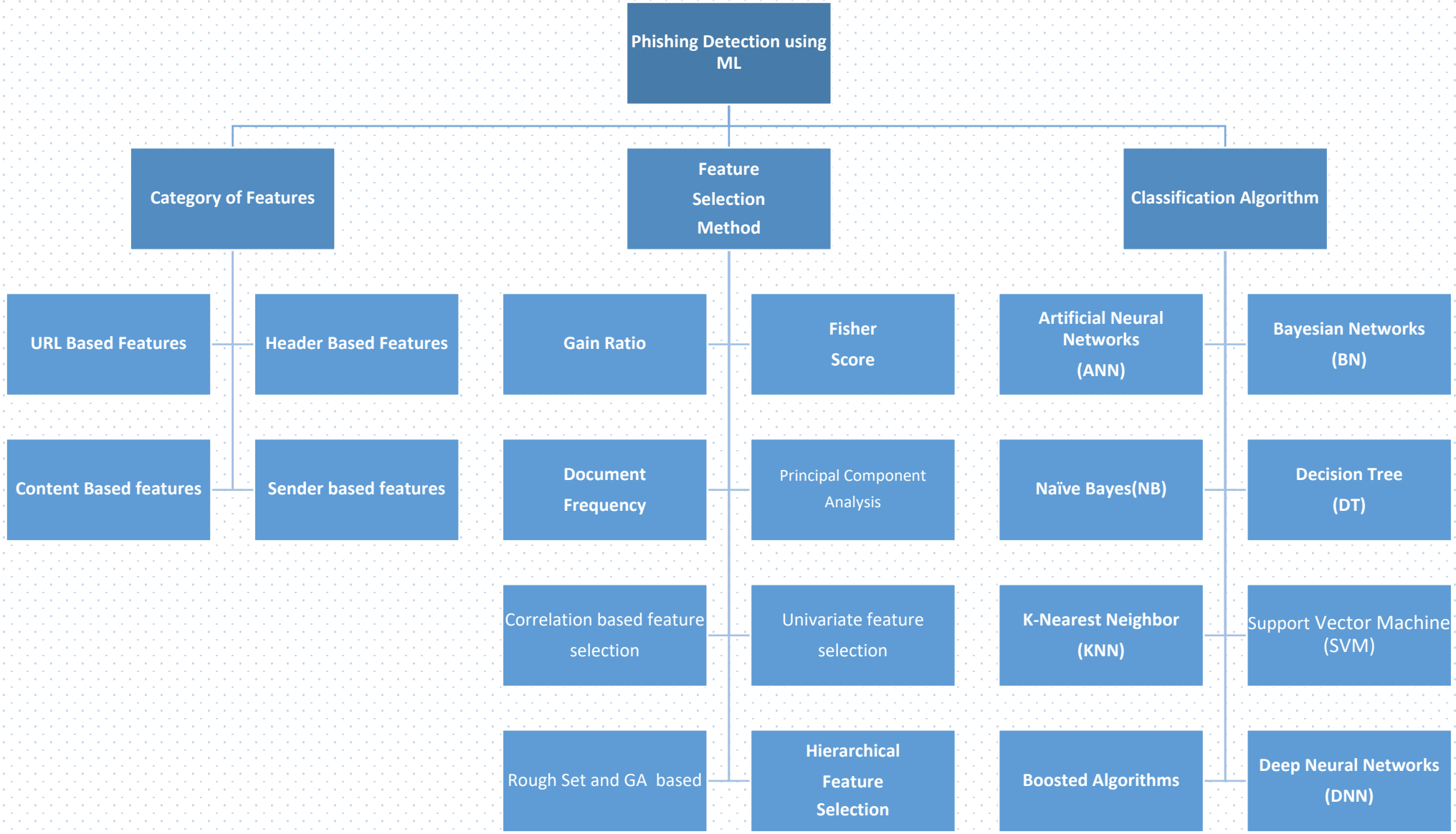


Figure 3 AUC-ROC curve on Select from model

Catalog of Approaches used for Phishing Detection



Malware detection/Classification

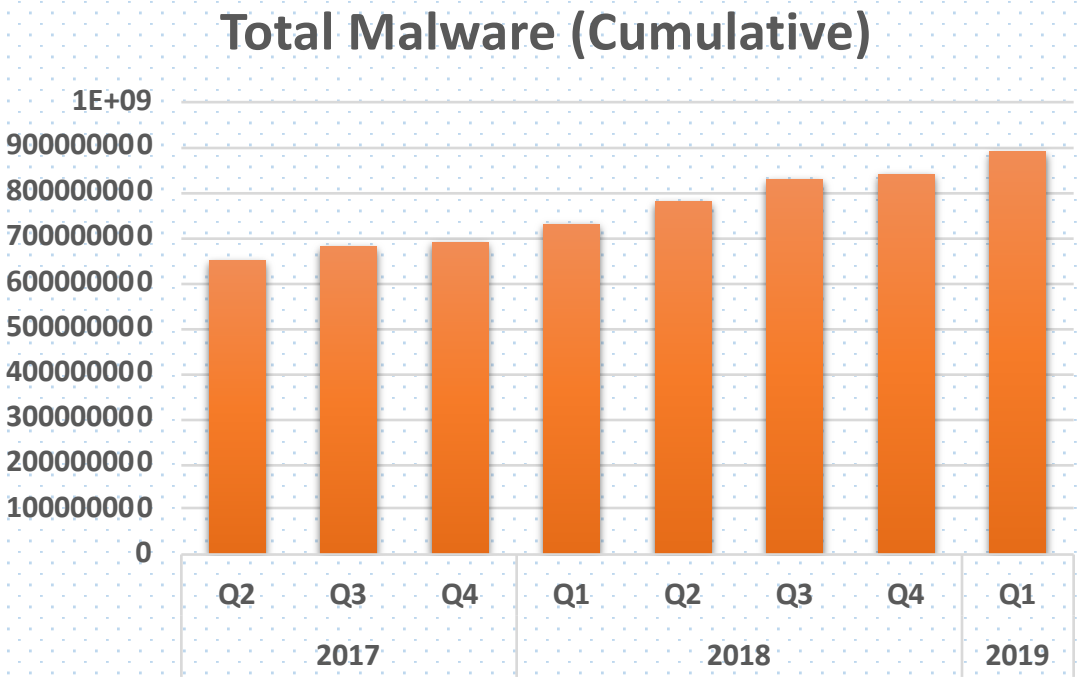
Malware, short for malicious software, is software designed to gain access to confidential information, disrupt computer operations, and/or gain access to private computer systems. Malware can be classified by how it infects systems:

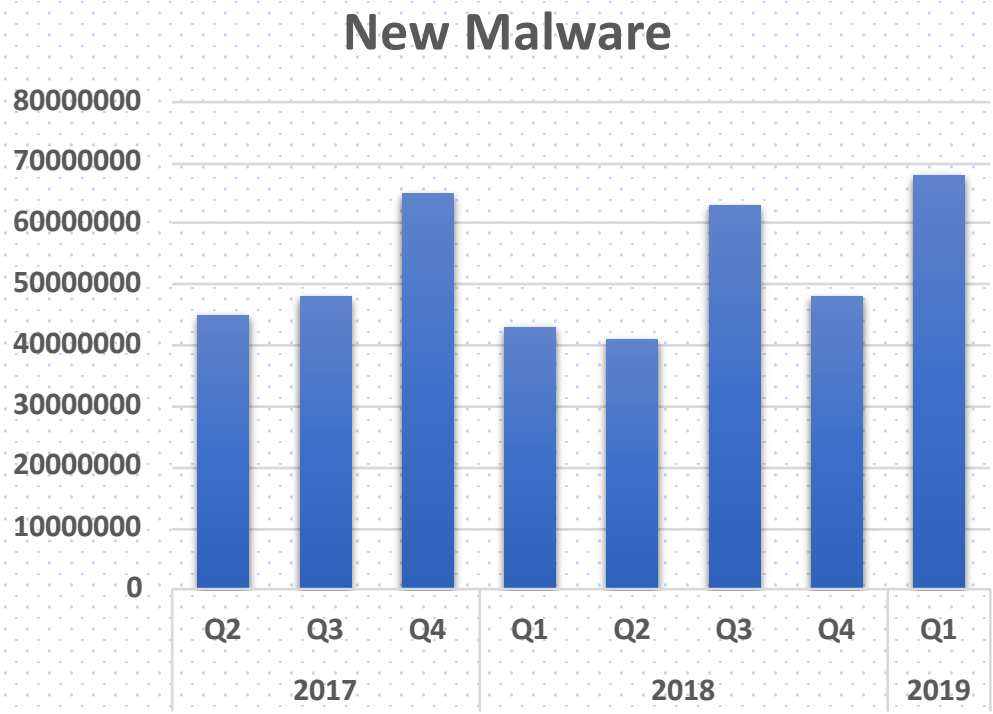
- Trojan Horses
- Viruses
- Worms

Or by what assets it targets:

- Ransomware
- Information stealers
- Spyware and adware
- Backdoors
- Rootkits
- Botnets

- Increasing threats……
  - ➢ Continuous and increased attack on infrastructure
  - ➢ Threats to business national security and personal security of PC’s





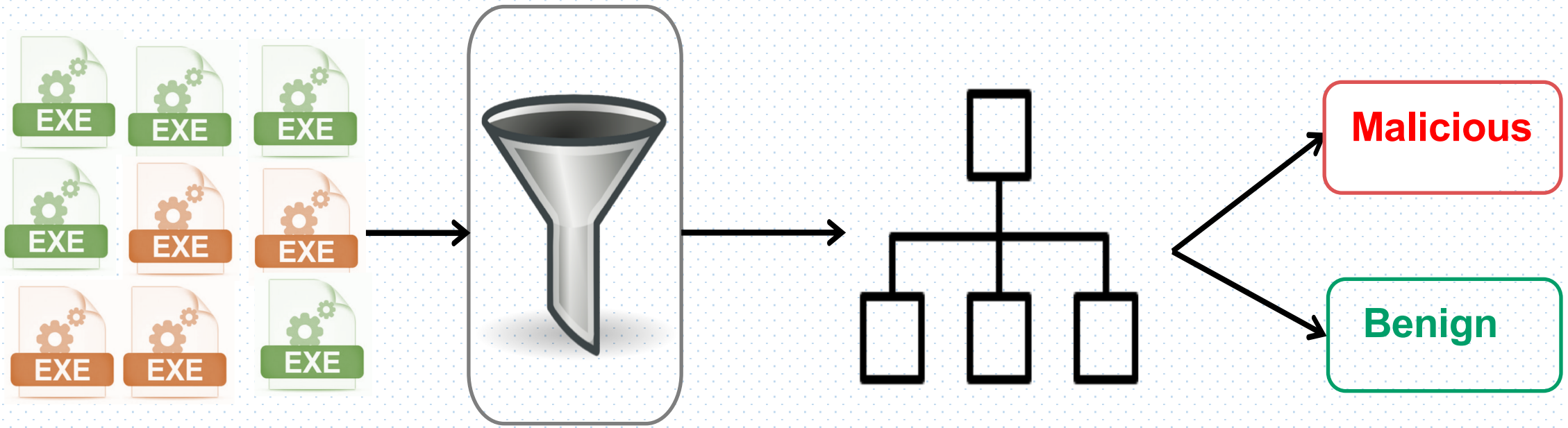
Source :[McAfee Threat Report]

The problem: Antivirus

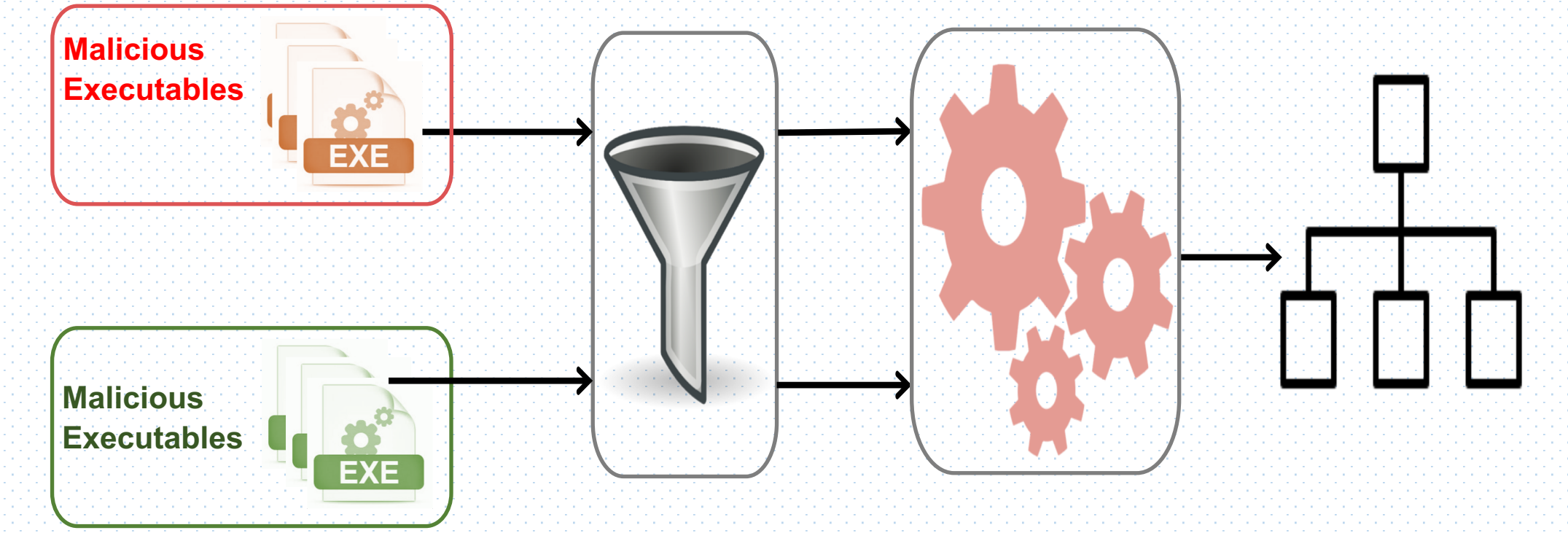
- How Antivirus detect malware?
  - Antivirus uses signature, heuristics and hand crafted rules that do not scale well.
- How Malwares evade antivirus?
  - Using polymorphism and obfuscation.
- Is Antivirus Dead?
  - The Security industry has declared Antivirus Solutions as dead but there is no universally accepted replacement.

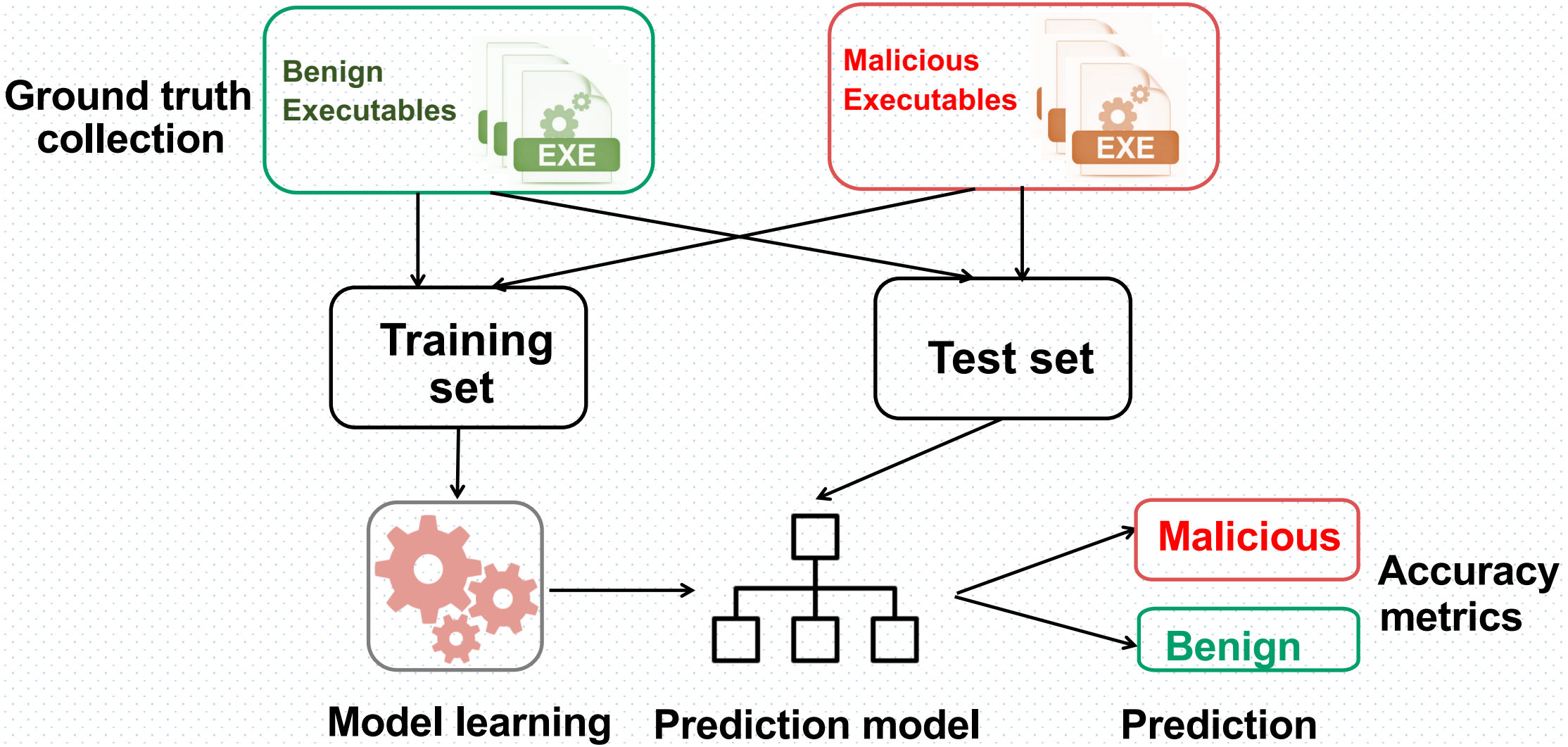
- Why Machine Learning?
  - Too many types of malware bots, virus
  - Too much data
  - Malwares are becoming more advanced and sophisticated.
  - Machine Learning can be the replacement.

Malware detection and Classification

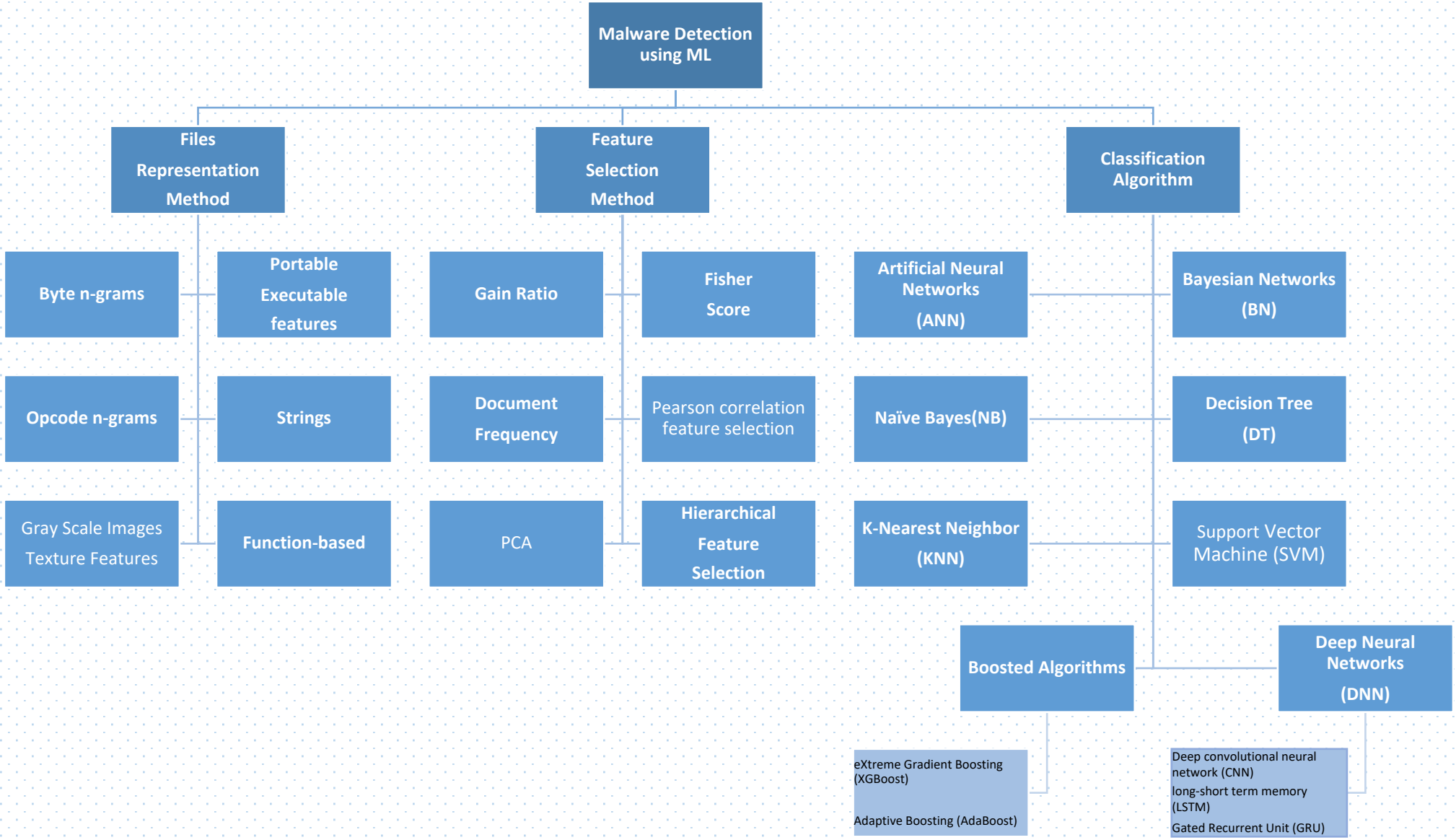


Malware detection and classification training




Ground truth collection
Benign Executables
EXE
Malicious Executables
EXE
Training set
Test set
Malicious
Benign
Accuracy metrics
Model learning
Prediction model
Prediction

Catalog of Approaches used for Malware Classification



Example Features

32/64 bit Executable	GUI Subsystem	Command Line Subsystem	File Size	Timestamp
Debug Information Present	Packer Type	File Entropy	Number of Sections	Number Writable
Header Information	Byte Histogram	Distribution of Section Entropy	Imported DLL Names	Imported Function Names
Compiler Artifacts	Linker Artifacts	Resource Data	Embedded protocol Strings	Embedded IPS/Domains
Embedded Paths	Embedded Product Meta Data	Digital Signature	Icon Content	Export Information

Experiment

- Obtaining data is the first step, our dataset consists of 10,123 benign and 40212 malicious PE files. The training set contained 70 % data and the test set contained 30 % data.
- LIEF library was used to parse these PE files and extract PE features.
- Feature Selection Using Extra Tree Classifier.
- Extra Tree Classifier selected 23 important features out of 57 features.

Model Evaluation Matrix			
Algorithm Name	Recall	Precision	F1 Score
XG BOOST	0.964	0.914	0.938
RANDOM FOREST	0.961	0.919	0.939
XG BOOST WITH SMOTE	0.990	0.989	0.990
RANDOM FOREST WITH SMOTE	0.997	0.994	0.996

Anomaly Intrusion Detection

- Intrusion Detection:
  - Process of monitoring the events occurring in a computer system or network and analyzing them for intrusions
  - Intrusions are defined as attempts to bypass the security mechanisms of a computer or network
- Challenges
  - Traditional signature-based intrusion detection systems are based on signatures of known attacks and cannot detect emerging cyber threats
  - Substantial latency in deployment of newly created signatures across the computer system
- Anomaly detection can alleviate these limitations

Type of Anomaly

- Point Anomalies
- Contextual Anomalies
- Collective Anomalies

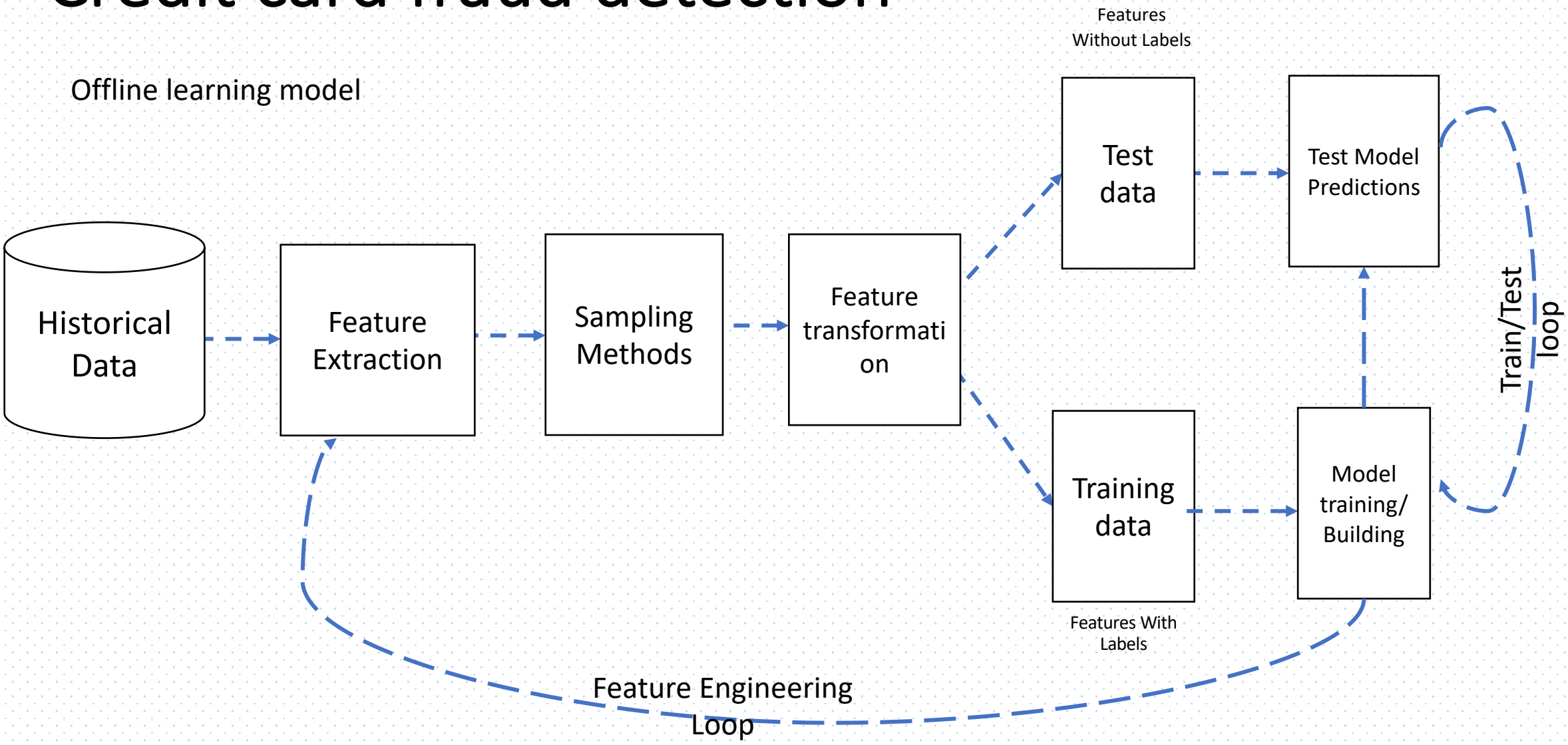
Applications of Anomaly Detection

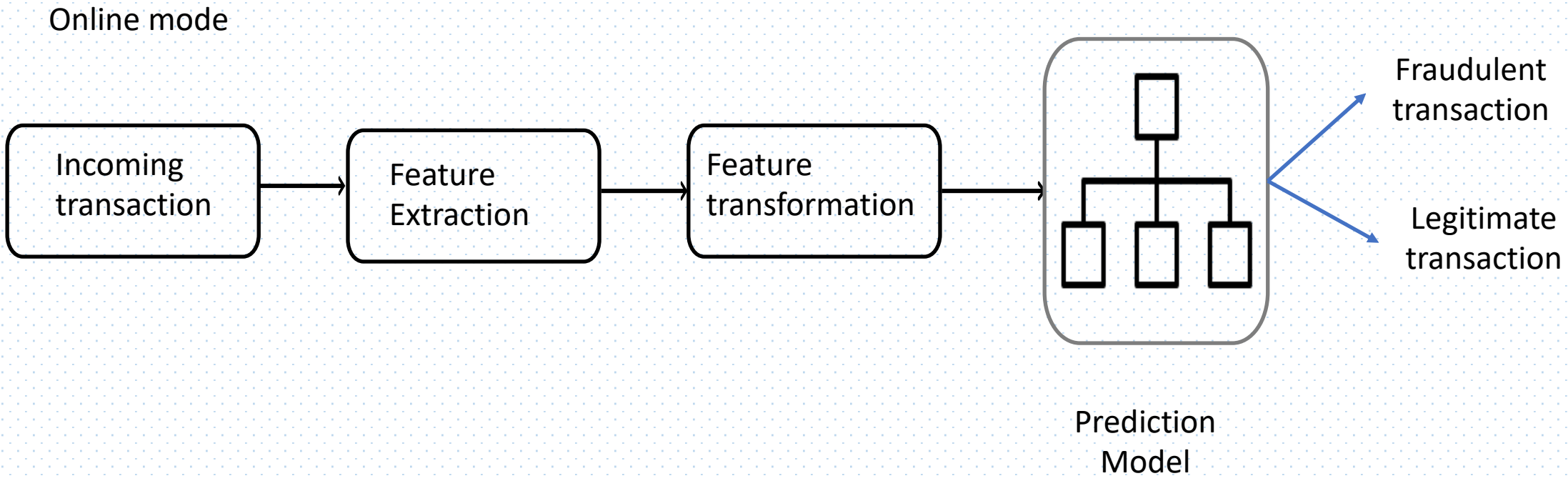
- Network intrusion detection
  - A web server involved in *ftp* traffic
- Insurance / Credit card fraud detection
  - An abnormally high purchase made on a credit card
- Healthcare Informatics / Medical diagnostics
- Industrial Damage Detection
- Image Processing / Video surveillance
- Novel Topic Detection in Text Mining

Fraud Detection

- Fraud detection refers to detection of criminal activities occurring in commercial organizations
  - Malicious users might be the actual customers of the organization or might be posing as a customer (also known as identity theft).
- Types of fraud
  - Credit card fraud
  - Insurance claim fraud
  - Mobile / cell phone fraud
  - Insider trading
- Challenges
  - Fast and accurate real-time detection
  - Misclassification cost is very high

Credit card fraud detection




Online mode
Incoming transaction
Feature Extraction
Feature transformation
Prediction Model
Fraudulent transaction
Legitimate transaction

Example Features

Features associated with the Card Holder

- Card Type
- Expiration date
- Home Address

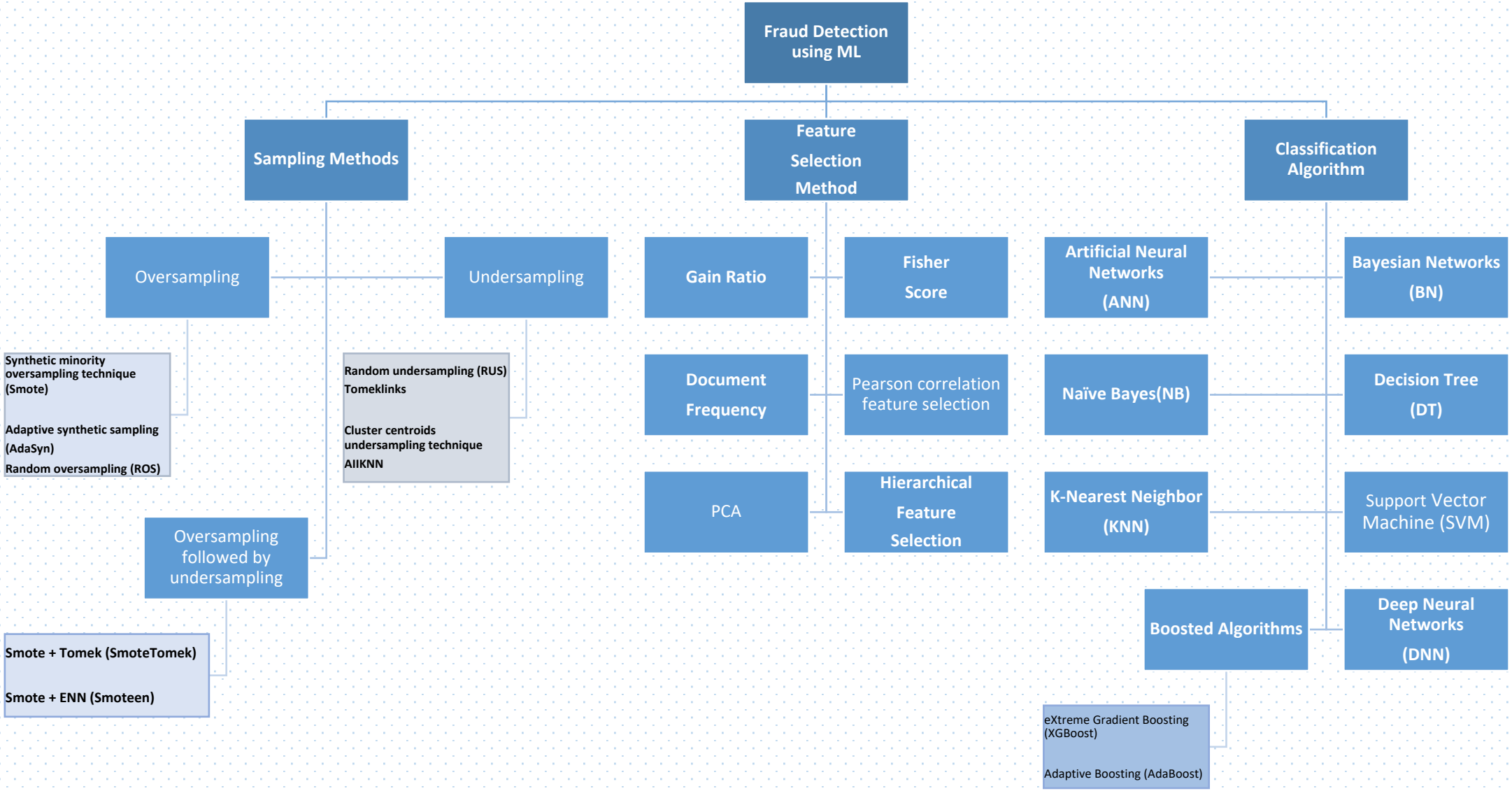
Features associated with the transaction

- POS Number
- Account Number
- Date and Time
- Transaction Amount
- Merchant category

Features derived from transaction history

- Number of Transactions in last 24 hours
- Total $ amount las 24 Hours
- Average amount last 24 hours
- Average amount last 24 hours compared to historical use
- Location and time difference since last transaction
- Average transaction
- Fraud risk of merchant type
- Merchant types for day compared to historical use

Catalog of Approaches used for Credit card Fraud Detection



- In the Credit card Fraud Detection system, datasets obtained from https://www.kaggle.com/mlg-ulb/creditcardfraud.
- This dataset contained transactions made by credit card in September 2013 by European cardholders. It is a highly Imbalance, dataset, containing 31 features and 284,807 instances. Upcoming Table is providing all the necessary information regarding dataset.

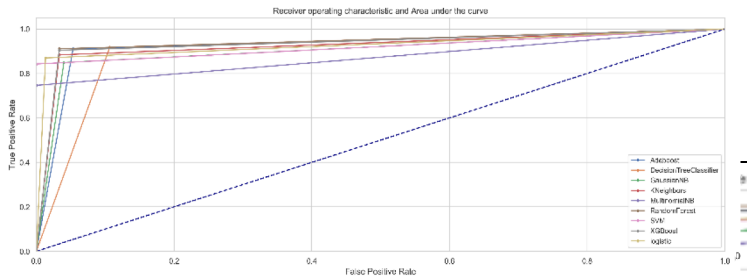
Description	Values
Number of features	31
Number of labels	2,84,807
Number of Positive labels (one)	492
Number of Negative labels (zero)	2,84,315
Number of Null values in instances	Zero

Classifier	Precision value for Balanced Dataset								
	SMOTE	ADASYN	ROS	RUS	TMLK	CC	AIIKNN	SMTMK	SMTN
AdaBoost	0.984	0.962	0.988	0.943	0.964	0.992	**1.000**	0.984	0.984
Decision Tree	0.997	0.997	0.999	0.893	0.964	0.992	**1.000**	0.997	0.996
GaussianNB	0.972	0.927	0.970	0.953	0.701	0.977	0.775	0.972	0.973
KNeighbors	0.998	0.998	0.999	0.962	0.969	0.991	1.000	0.998	**0.999**
MultinomialNB	0.999	0.909	0.999	**1.000**	----	0.990	----	**0.999**	0.996
Random Forest	**1.000**	**1.000**	**1.000**	0.963	0.964	0.984	1.000	**0.999**	0.999
SVM	0.977	0.914	0.980	1.000	0.969	0.991	1.000	0.977	0.976
XGBoost	0.989	0.964	0.993	0.963	0.965	0.992	1.000	0.989	0.988
Logistic Regression	0.975	0.914	0.976	0.984	0.959	0.983	0.991	0.975	0.974

Classifier	Recall value for Balanced Dataset								
	SMOTE	ADASYN	ROS	RUS	TMLK	CC	AIIKNN	SMTMK	SMTN
AdaBoost	0.969	0.967	0.982	0.911	**1.000**	**1.000**	0.984	0.968	0.970
Decision Tree	0.999	0.999	**1.000**	**0.917**	0.990	0.992	**0.992**	0.999	0.998
GaussianNB	0.850	0.526	0.859	0.849	**1.000**	**1.000**	0.991	0.852	0.852
KNeighbors	**1.000**	**1.000**	1.000	0.883	0.863	0.993	0.893	**1.000**	**1.000**
MultinomialNB	0.767	0.677	0.760	0.746	0.000	0.785	0.000	0.767	0.769
Random Forest	**1.000**	**1.000**	**1.000**	0.910	0.982	0.984	0.984	**1.000**	**1.000**
SVM	0.917	0.866	0.913	0.842	0.845	0.905	0.840	0.916	0.917
XGBoost	0.972	0.985	0.996	0.904	1.000	1.000	**0.992**	0.967	0.969
Logistic Regression	0.922	0.855	0.919	0.869	0.855	0.944	0.879	0.922	0.922

Abbreviations used in Tables:

ROS: Random Over Sampling.
TMLK: TomekLinks
CC: Cluster Centroids
SMTMK: SmoteTomek
SMTN: Smoteen
RUS: Random Under-Sampling



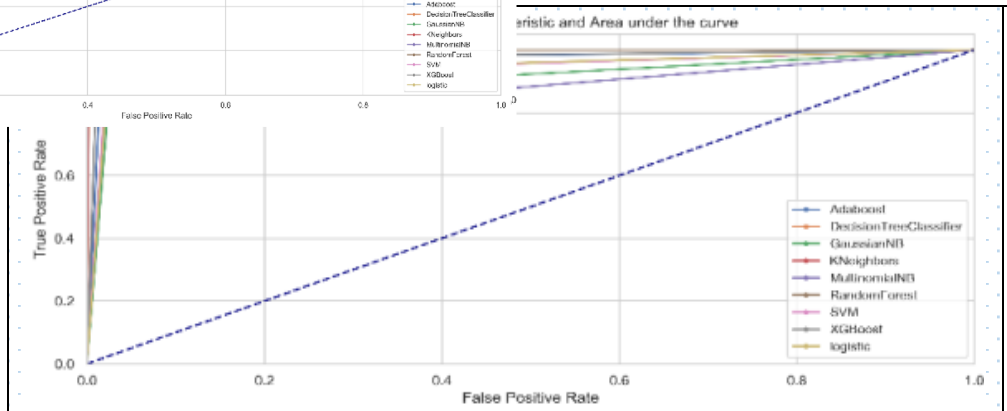


Figure 1 AUC-ROC curve on Smote

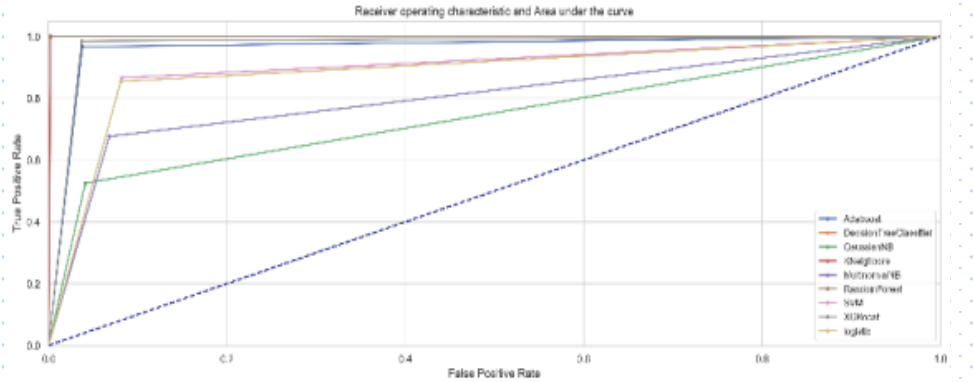


Figure 2 AUC-ROC curve on AdaSyn

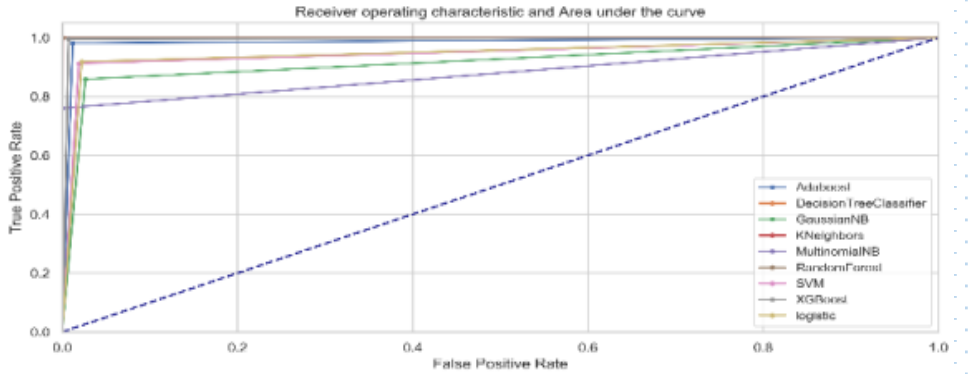


Figure 3 AUC-ROC curve on Random OverSampling

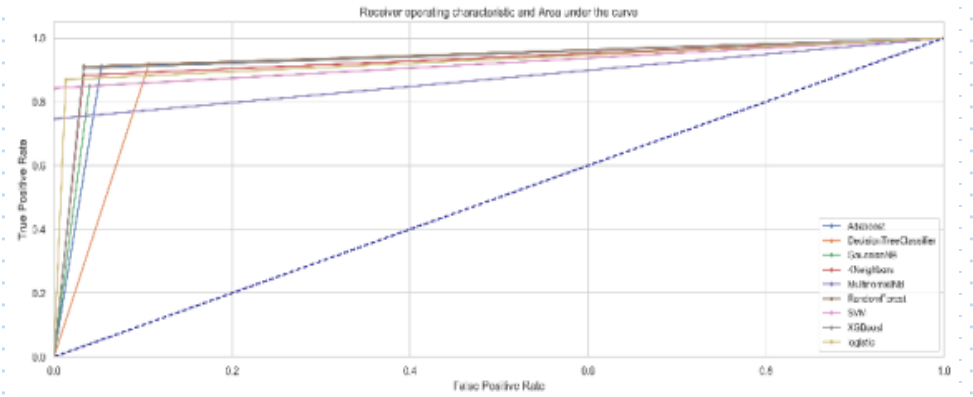


Figure 4 AUC-ROC curve on Random UnderSampling

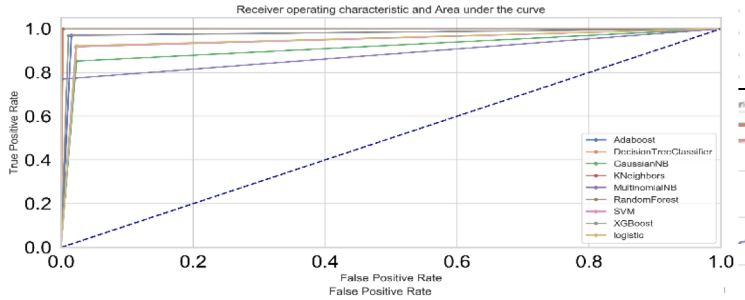


Figure 1 AUC-ROC curve on TomekLinks

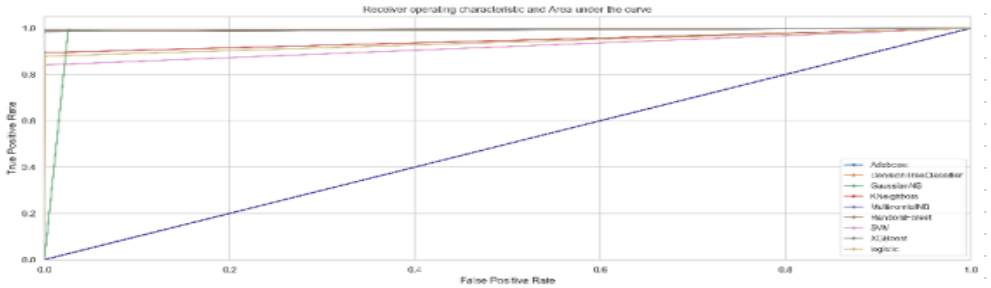


Figure 2 AUC-ROC curve on AllKNN

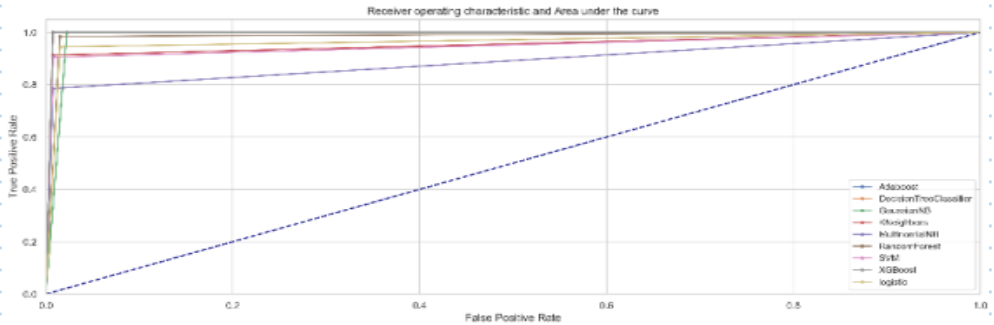


Figure 3 AUC-ROC curve on Center Centroids

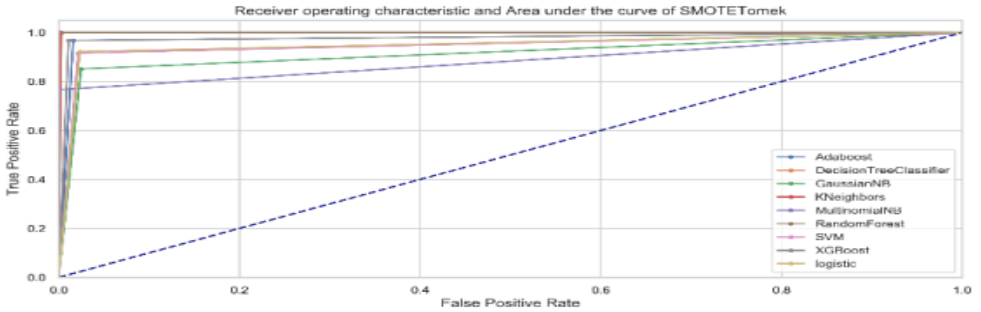


Figure 4 AUC-ROC curve on SmoteTomek

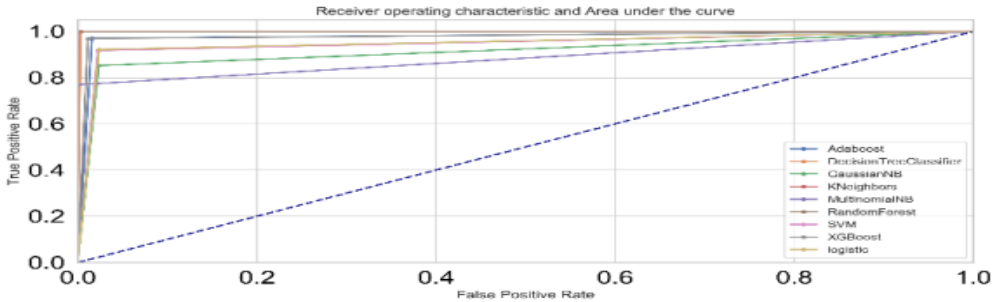


Figure 5 AUC-ROC curve on Smoteen

Future Direction and Challenges

Challenges

- No well-trained domain experts and data scientists to oversee the implementation.
- Bad understanding of the data to engineer meaningful **features** (e.g. Byte stream for binaries)
- Lack of Labeled Samples and Certainty in Ground Truth.
- Lack of publicly available and labelled data for training.
- Reliable Evaluation Data Sets or Data Generation Tools.
- Verifiability of output.
- Interpretation of output.
- Data **cleanliness** issues (timestamps, normalization across fields, etc.)
- Imbalanced Data Sets.
- Imbalanced Learning Problem and Advanced Evaluation Metrics.
- Concept Drift.
- Operational effectiveness, efficiency, or cost reduction.
- Weaponization of artificial intelligence in Cyber Security.

- **Future Direction**
  - Transfer Learning and “One-Shot Learning”
  - Continuous Learning
  - Reinforcement Learning
  - Adaptive Machine Learning
    - Adversarial Machine Learning
    - Emerging technologies such as IOT, Blockchain and cloud computing.
    - Swarm Viruses and Antiviruses.

Thank you for listening!