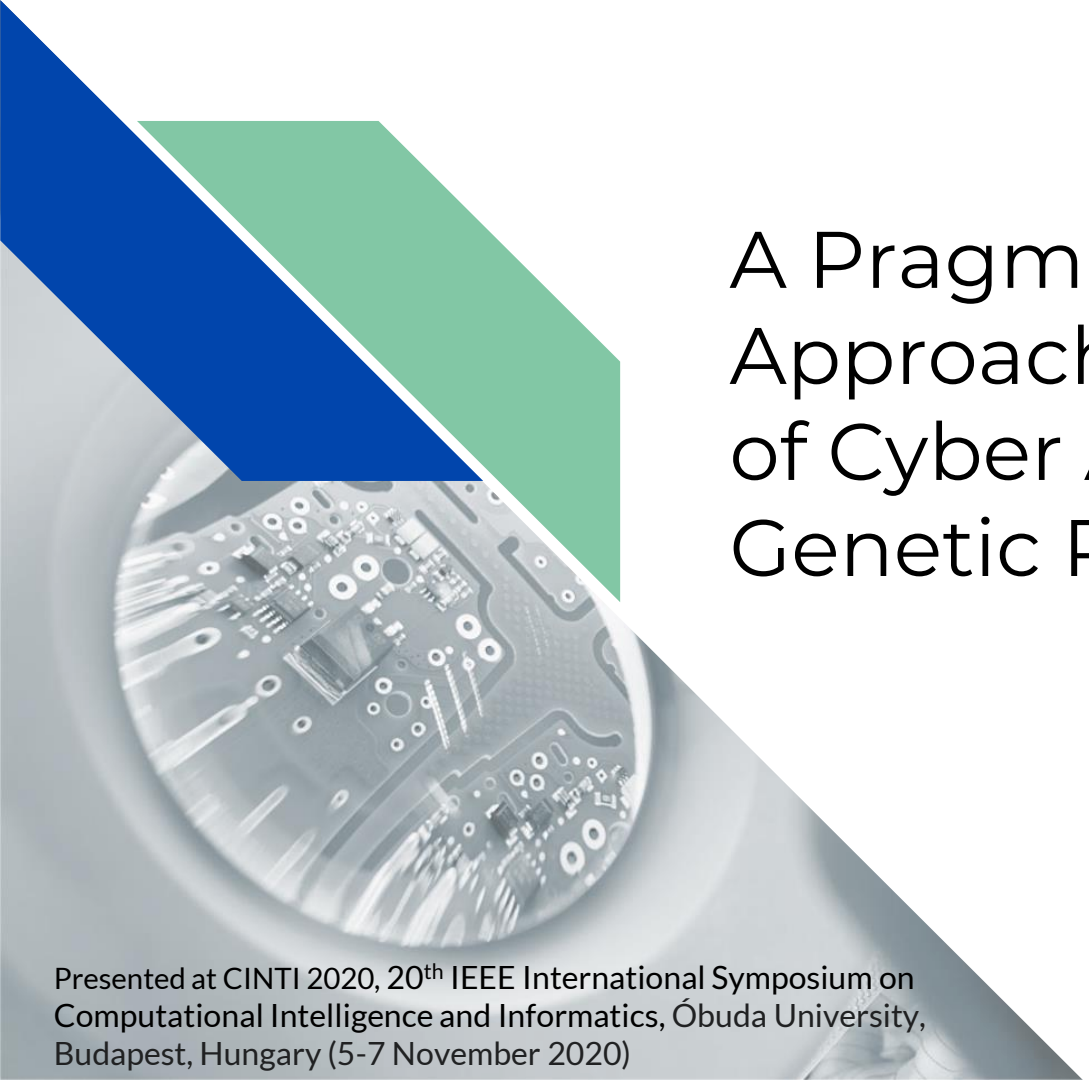
A Pragmatic Optimal Approach for Detection of Cyber Attacks using Genetic Programming



Authors:

- Nikhil Mane, PESIT, Bangalore
- Anjali Verma, PESIT, Bangalore
- Arti Arya, PES University, Bangalore

Presented at CINTI 2020, 20th IEEE International Symposium on Computational Intelligence and Informatics, Óbuda University, Budapest, Hungary (5-7 November 2020)

Overview:

- Problem Statement
- Introduction
- Genetic Programming
- Proposed Method:
  - -Data Acquisition
  - -Preprocessing
  - -Implementation Using Genetic Programming
- Experiment and Results
- Conclusion

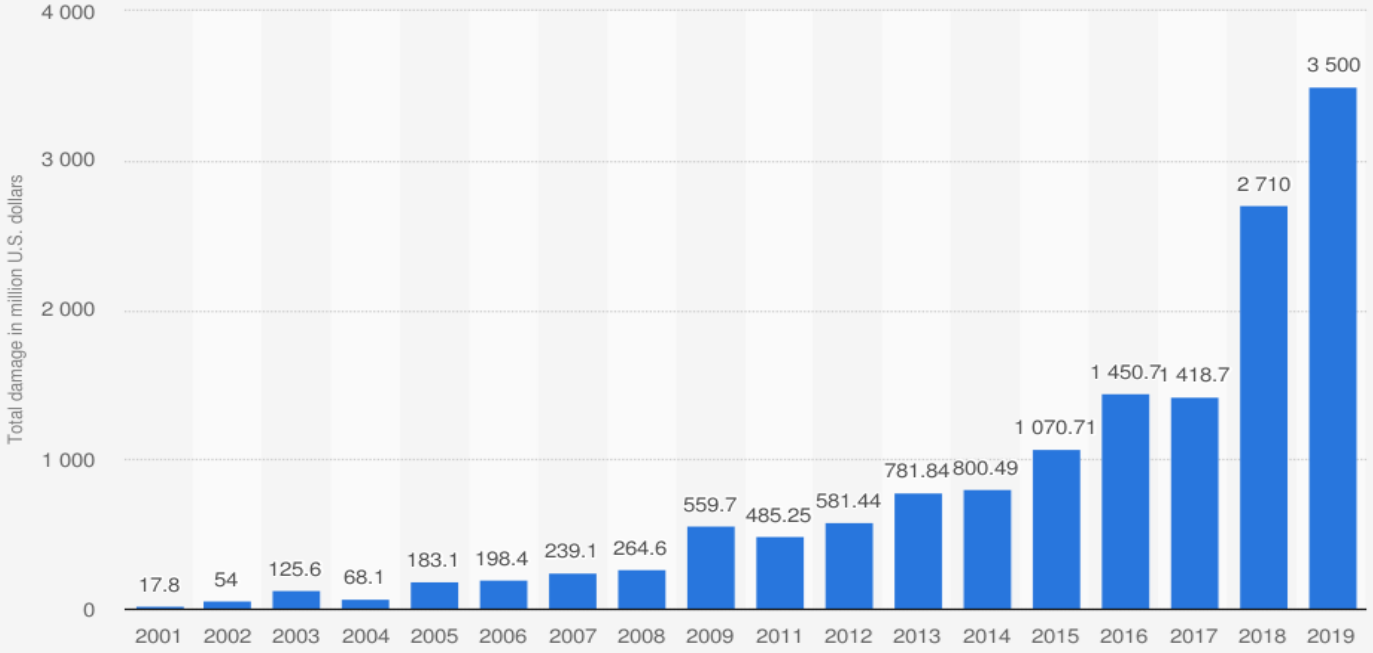
Problem Statement:

- Attackers have been evolving their strategies and methods with time. Using ML/DL methods will certainly improve their exploitations. Therefore, developing a model which fights against such threats is very essential.
- Solving a complex attack has become a very difficult task for security engineers. Hence, building a self learning model will enhance the ability to identify and rectify such attacks.
- Attacks happen when the systems are most vulnerable, such scenarios will be guarded by the developed model.

Statistics:

- **94%** of malware is delivered via mail.
- Phishing attacks account for more than **80%** of reported security incidents.
- **$17,700** is lost every minute due to phishing attacks
- **60%** of breaches involved vulnerabilities for which a patch was available but not applied
- **63%** of companies said their data was potentially compromised within the last 12 months due to hardware-or silicon-level security breach.
- Data breaches cost enterprises an average of **$3.92 million**.
- **40%** of IT leaders say cybersecurity jobs are the most difficult to fill.

Amount of monetary damage caused by reported cyber crime to the IC3 from 2001 to 2019 (in million U.S. dollars)

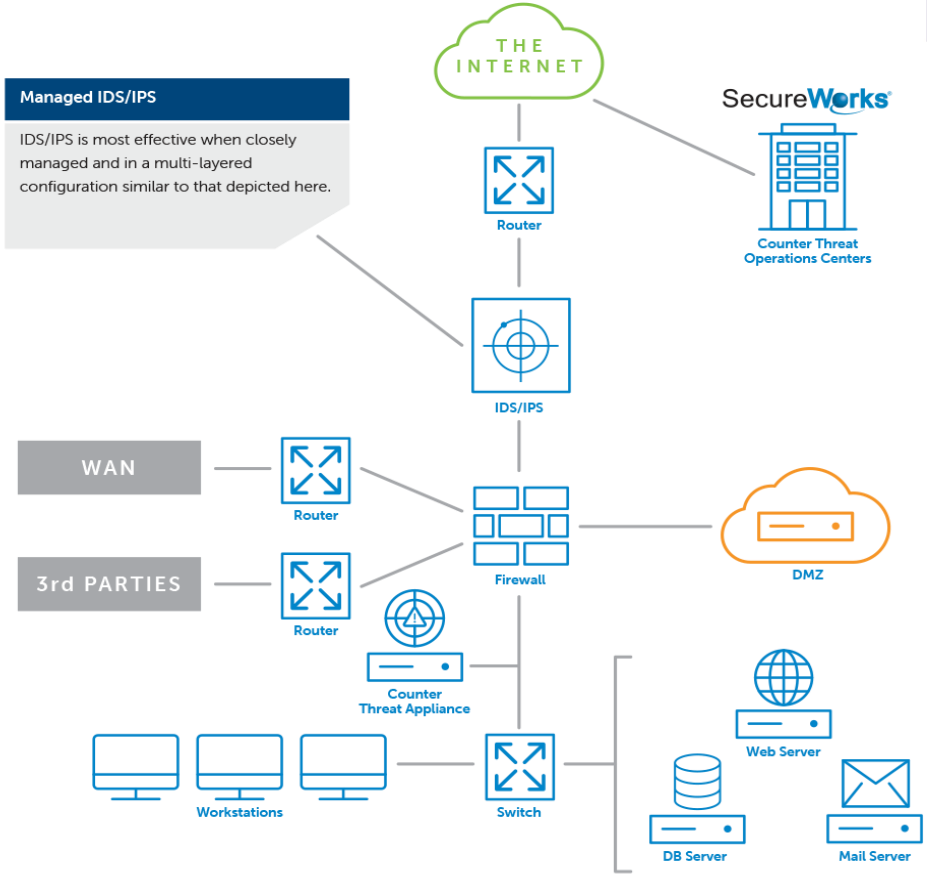


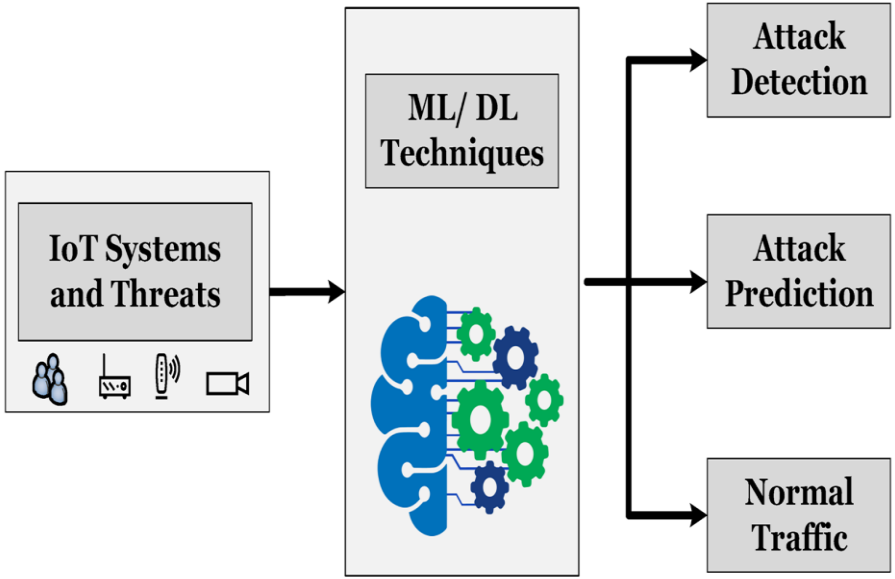
Sources
FBI; IC3; US Department of Justice

Additional Information:
Worldwide; IC3; 2001 to 2019, excluding 2010; Cybercrime reported to IC3

- “**Intrusion detection** is the process of monitoring the events occurring in a computer system or network and analysing them for signs of possible incidents, which are violations or imminent threats of violation of computer security policies,acceptable use policies, or standard security practices"

- An **Intrusion detection system** (IDS) is a software application or hardware appliance that monitors traffic moving on networks and through systems to search for suspicious activity and known threats, sending up alerts when it finds such items.





Machine learning/Deep learning is adopted in a wide range of domains where it shows its superiority over traditional rule-based algorithms. These methods are being integrated in cyber detection systems with the goal of supporting or even replacing the first level of security analysts.

The use of a branch of ML namely **Evolutionary Computation (EC)** plays an important role in tracking, analyzing, identifying digital security threats to combat viruses and hackers

Genetic Programming

- Executional steps for general GP:

1. Generate an initial population of random compositions (Computer Programs).

2. Run a tournament, which picks four programs randomly out of the population of programs.

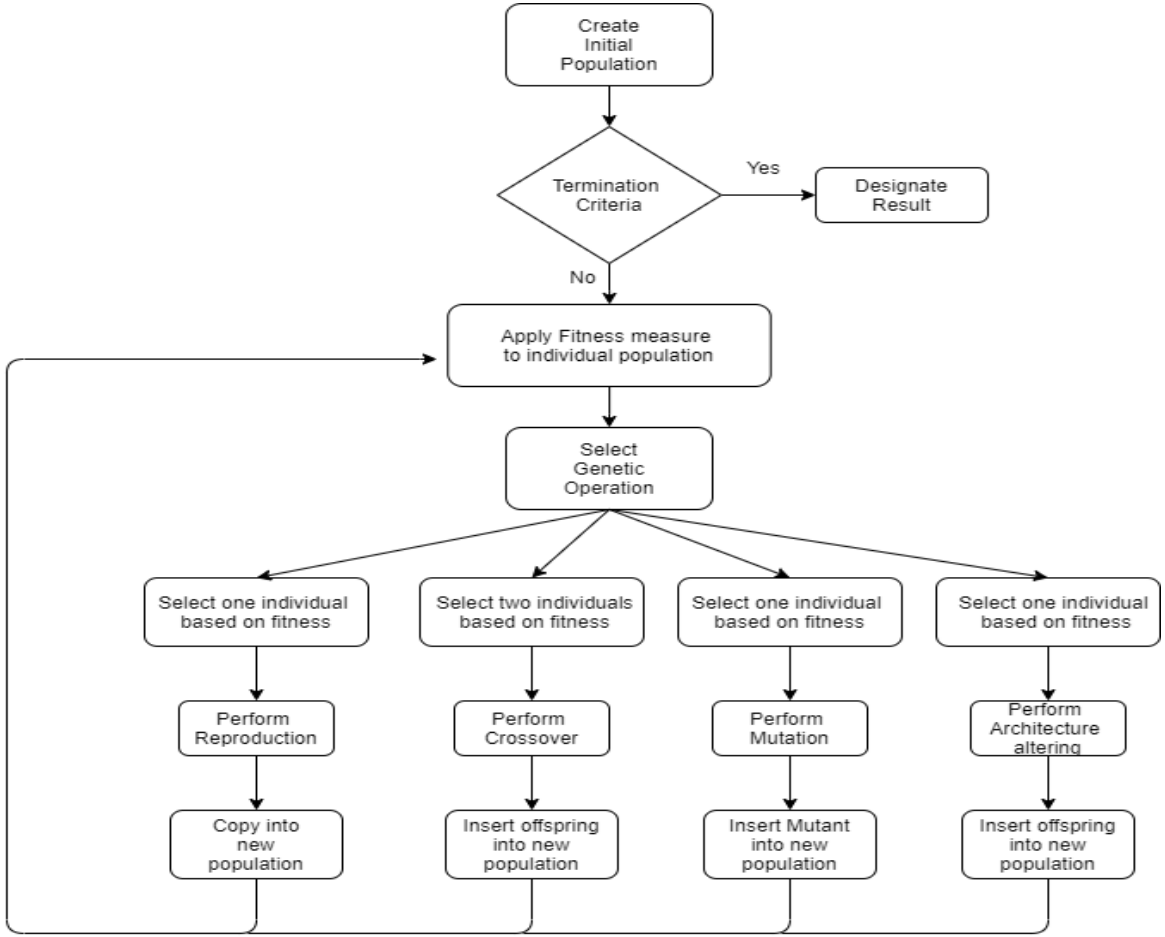
3. Apply the search operators crossover and mutation (and possibly others) to the winners

a. Copy the two winners
b. With **Crossover** Frequency, apply crossover to copies of the winners
c. With **Mutation** Frequency, mutate the programs from (a)

4. Replace the tournament losers with the new offspring.

5. Repeat until a predefined termination criterion has been satisfied, or a fixed number of generations have been explored.

6. The solution is the genetic program with the best fitness within all the generations.


Create Initial Population
Termination Criteria
Yes
Designate Result
No
Apply Fitness measure to individual population
Select Genetic Operation
Select one individual based on fitness
Select two individuals based on fitness
Select one individual based on fitness
Select one individual based on fitness
Perform Reproduction
Perform Crossover
Perform Mutation
Perform Architecture altering
Copy into new population
Insert offspring into new population
Insert Mutant into new population
Insert offspring into new population

Proposed Method:

1) Data Acquisition

- Modern DDoS Dataset is used for implementation.
- A novel Dataset which that contains modern kinds of DDoS attacks.
- Generated using NS2 Network simulator.
- The dataset had 2,160,668 number of instances.

Variable Number	Attribute Name	Description
1	SRC_ADD	Source Address
2	DES_ADD	Destination Address
3	PKT_ID	Packet Identifier
4	FROM_NODE	Source Node
5	TO_NODE	Destination Node
6	PKT_TYPE	Packet Type
7	PKT_SIZE	Total packet size in bytes
8	FLAGS	Flags
9	FID	Flow identifier
10	SEQ NUMBER	Sequence number
11	NUMBER_OF_PACKET	Total number of packets
12	NUMBER_OF_BYTE	Total number of bytes
13	NODE_NAME_FROM	Node Name From
14	NODE_NAME_TO	Node Name To
15	PKT_IN	Total time of packet inside queue
16	PKT_OUT	Total time of packet outside queue
17	PKT_R	Time of packet received
18	PKT_DELAY_NODE	Time packet delay within node
19	PKT_RATE	Average packet rate
20	BYTE_RATE	Average byte rate
21	PKT_AVG_SIZE	Average packet size
22	UTILIZATION	Bandwidth utilization
23	PKT_DELAY	Total time packet delay
24	PKT SEND TIME	Time of sending packet
25	PKT RESERVED TIME	Time of receiving packet
26	FIRST PKT SENT	Time of first packet sent
27	LAST PKT RESERVED	Time of last packet received

Dataset classes

- **Smurf** : The target server receives huge number of ICMP echo requests packet.
- **UDP Flood**: A massive amount of UDP traffic is sent to inundate the server.
- **SQL Injection DDOS**: Sql sentences are used to flood the server.
- **HTTP Flood**: attacker overwhelm the server using HTTP GET/POST methods.
- **Normal** transaction data.

Category	No. of Records
Smurf	12,590
UDP Flood	201,344
SIDDOS	6,665
HTTP Flood	4,110
Normal	1,935,959

2) Preprocessing

- **Principal Component Analysis (PCA)** was applied on the dataset.
- The features were reduced to 8, 16 and 20 principal components.
- There was no significant difference in the 16 and 20 principal components in terms of percentage loss.
- Hence for further processing, 16 principal components were considered for simplicity.

Number of Principal Components	% Information Gain	% Information Loss
8	94.92%	5.08%
16	98.48%	1.52%
20	99.6%	0.4%

3) Implementation of Genetic Programming:

- For implementation, Distributed Evolutionary Algorithms in Python (**DEAP**) and Tree-Based Pipeline Optimization Tool (**TPOT**) frameworks were used.

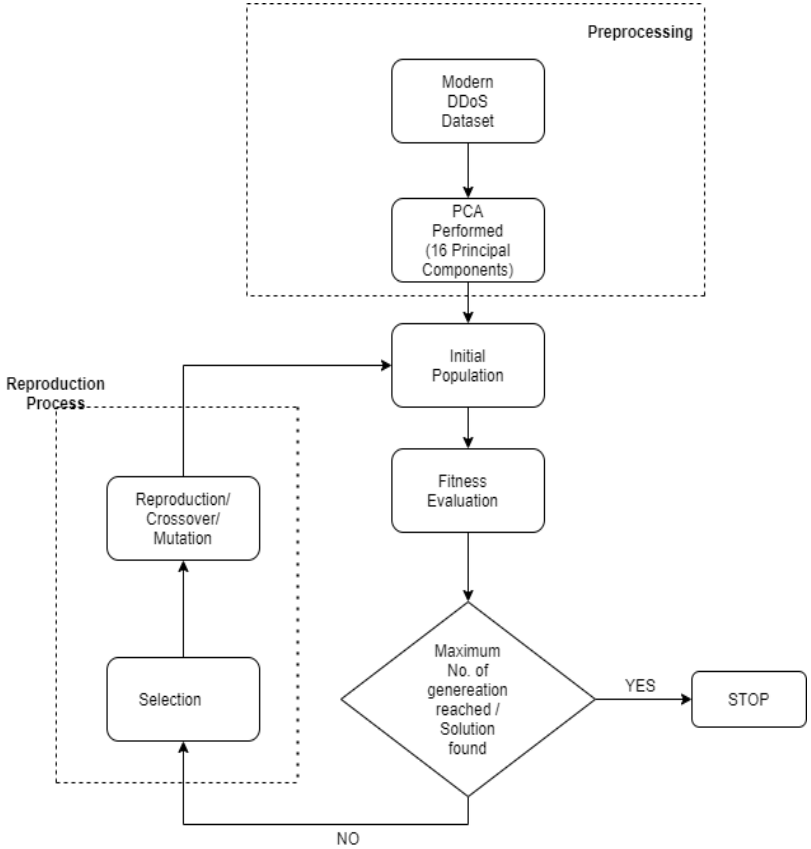
- **Performed in 4 steps**:

  1. Build an appropriate type of problem.

  2. Creating a fitness class using Creator module.

  3. Initialization of operators using toolbox operator.

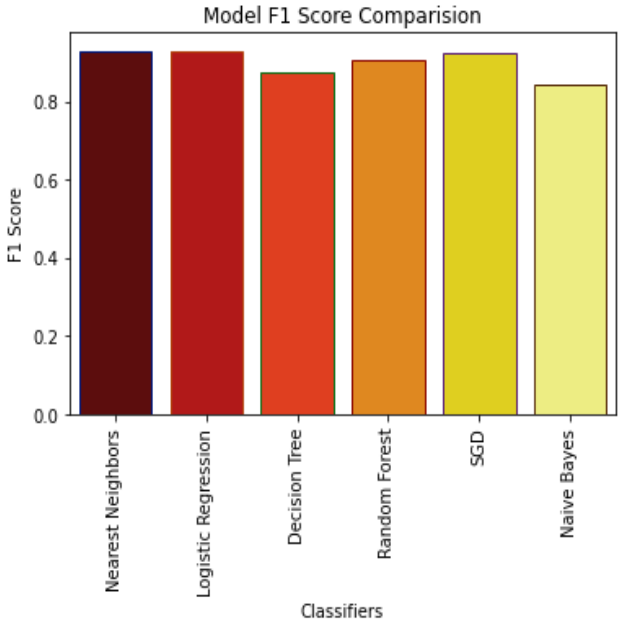
  4. Constructing the main function.



Experiments and Results

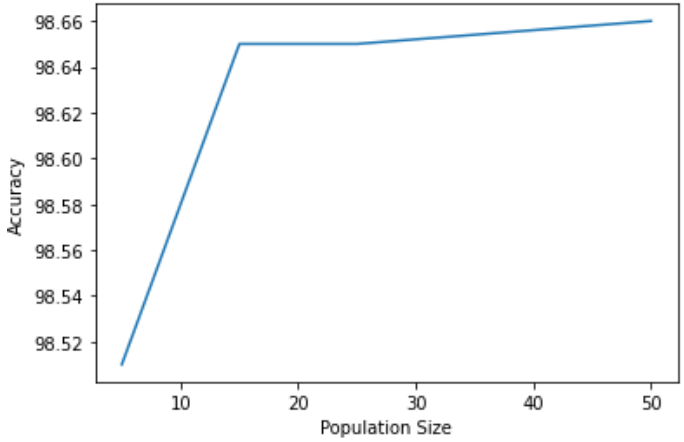
- The experiments were performed on six Machine Learning models.
- Values were evaluated using Confusion matrix.

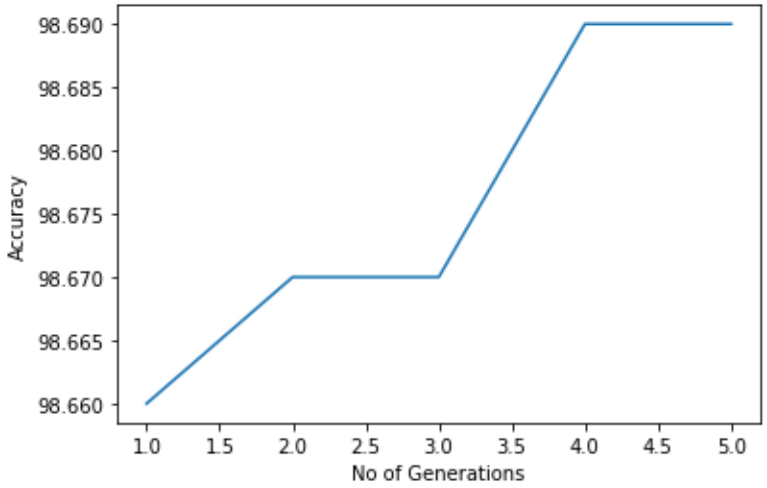
Classification Method Used	Correct Classification % Score
KNN	98.57%
Naive Bayers	96.66%
Logistic Regression	98.62%
Decision Tree	97.38%
Random Forest	98.02%
Stochastic Gradient Descent	98.55%



GP Results

- GP implementation depends on various parameters such as population size, no. of generations, crossover and mutation rate, etc.
- When the population size was considered as 50 with a crossover rate of 0.01, an accuracy of **98.67%** was obtained.
- It was observed that as the no. of generations were increased, the accuracy did not change much.





Comparative Study

Author	Model	Dataset	Accuracy
Hasanen Alyasiri John Clark and Daniel Kudenko	Cartesian Genetic Programming	Modern DDoS	97.19%
Manjula Suresh and R. Anitha	Naive Bayes	CAIDA	97.2%
Mouhammd Alkasassbeh, Ahmad B.A Hassanat, Ghazi Al-Naymat, Mohammad Almseidin	Random Forest Naive Bayes	Modern DDoS	98.02% 96.91%
Naveen Bindra, Manu Sood	Random Forest	CIC IDS 2017	96.13%
S. Umarani, D. Sharmila	Naives Bayes	1998 World Cup Website	95.95%
Proposed Model	**Genetic Programming**	**Modern DDoS**	**98.67%**

Conclusion

- This scientific analysis investigates an application of Genetic Programming (GP) for intrusion detection. For this study, the Modern DDoS dataset is used. This dataset contains contemporary threats gathered from various environments.
- The proposed GP model detects DDoS attacks with improved accuracy of **98.67%** while comparing it with six established classification models. The obtained results highlight the advantages of adopting the GP model.
- However, it was observed that adopting other approaches for operations such as mutation or crossover can result in better results. Due to limited resources, this was not tested.
- In future, this model can be investigated for other types of attacks and also to come up with a universal model to detect all kinds of well-known threats.