1. What are the challenges for network anomaly detection in compare to other domain? Which considerations/techniques as preprocessing and post processing NIDSs might be more applicable/useful?

The term anomaly-based intrusion detection in networks refers to the problem of finding exceptional patterns in network traffic that do not conform to the expected normal behavior. These nonconforming patterns are often referred to as anomalies, outliers, exceptions, aberrations, surprises.

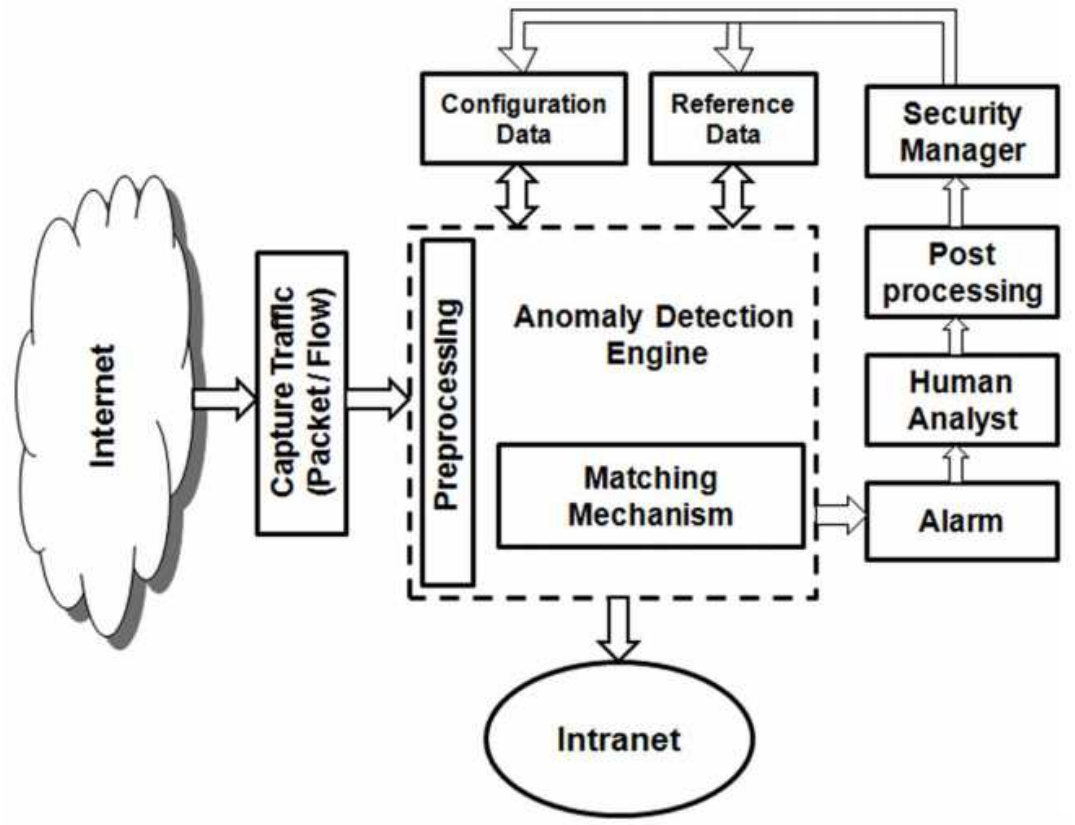

Fig. 1. A generic architecture of ANIDS

There are **several challenge**s faced by Network based Anomaly Detection in comparison to other domains like the performance metrics, datasets, robustness, scalability etc. Some challenges summarized below are:

i. Runtime limitation presents an important challenge for a NIDS. Without losing any packets, a real time IDS should be ideally able to capture and inspect each packet.
ii. Most NIDSs and network intrusion detection methods depend on the environment. Ideally, a system or method should be independent of the environment.
iii. The nature of anomalies keeps changing over time as intruders adapt their network attacks to evade existing intrusion detection solutions. So, adaptability of a NIDS or detection method is necessary to update with the current

anomalies encountered in the local network or the Internet.

iv. Ideally, a NIDS or detection method should avoid a high rate of false alarms. However, it is not possible to escape totally from false alarms, even though it needs to aim for that in any environment and facilitate adaptability at runtime. This is another challenge for the NIDS development community.
v. Dynamic updation of profiles in anomaly-based NIDSs without conflict and without compromising performance is an important task. The profile database needs to be updated whenever a new kind of attack is detected and addressed by the system.
vi. Preparing an unbiased network intrusion dataset with all normal variations in profiles is another challenging task. The number of normal instances is usually large and their proportion with attack instances is very skewed in the existing publicly available intrusion datasets. Only a few intrusion datasets with sufficient amount of attack information are available publicly. Thus, there is an overarching need for benchmark intrusion datasets for evaluating NIDSs and detection methods.
vii. Reducing computational complexity in preprocessing, training and deployment is another task that needs to be addressed.
viii. Developing an appropriate and fast feature selection method for each attack class is yet another challenge.
ix. Selection of an appropriate number of non-correlated, unbiased classifiers from a pool of classifiers by generating classifier hypothesis for building an effective ensemble approach for network anomaly detection is another challenge.

Defining a normal region which encompasses every possible normal behavior is very difficult. In addition, the boundary between normal and anomalous behavior is often not precise. Thus an anomalous observation which lies close to the boundary can actually be normal, and vice-versa.

The **main components** of the generic model of the ANIDS discussed in the paper are as follows and showcased in the figure are :

1. **Anomaly detection engine**:
   This is the heart of any network intrusion detection system. It attempts to detect occurrence of any intrusion either online or offline. However, before sending any network traffic to the detection engine, it
   needs preprocessing.
   Matching mechanism: It entails looking for a particular pattern or profile in network traffic that can be built by continuous monitoring of network behavior including known exploits or vulnerabilities.
2. **Reference data:**
   The reference data stores information about known intrusion signatures or

profiles of normal behavior. Reference data needs to be stored in an efficient manner. Possible types of reference data used in the generic architecture of a NIDS are: profile, signature and rule.

3. **Configuration data:**
   This corresponds to intermediate results, e.g., partially created intrusion signatures. The space needed to store such information can be quite large.
4. **Alarm:**
   This component of the architecture is responsible for generation of alarm based on the indication received from the detection engine.
5. **Human analyst:**
   A human analyst is responsible for analysis, interpretation and for taking necessary action based on the alarm information provided by the detection engine. The analyst also takes necessary steps to diagnose the alarm information as a post-processing activity to support reference or profile updation with the help of security manager.
6. **Post-processing**:
   This is an important module in a NIDS for post-processing of the generated alarms for diagnosis of actual attacks.
7. **Capturing traffic:**
   Traffic capturing is an important module in a NIDS. The raw traffic data is captured at both packet and flow levels.
8. **Security manager:**
   Stored intrusion signatures are updated by the Security Manager (SM) as and when new intrusions become known. The analysis of novel intrusions is a highly complex task.

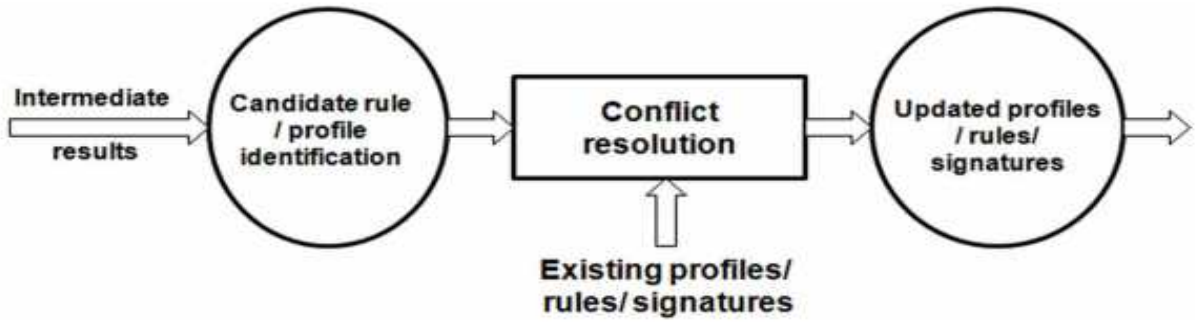

Fig. 2. Steps for updation of configuration data in ANIDS

2. In NIDS domain till now most of NIDSs are using signature based techniques, what is your opinion using ML to address this issue?

IDS has two detection approaches; anomaly-based and signature-based.
Signature based intrusion detection system (SBIDS) based on the known signature. This type of detection is more effective against known attacks, and it depends on the continuance updating signature. The main drawback of SBIDS is, it is unable to detect the unknown attacks and novel attacks, but the detection rate is higher than anomaly intrusion detection rates.

Signature analysis contains the same knowledge-acquisition approach as an expert system, but the way of knowledge acquired is different. Machine learning is a technique which can be defined as the ability of the program and/or a system to learn and improve their performance for certain tasks or a group of tasks over time. It primarily concentrates on establishing a system for improving the performance on the basis of previous result it means that machine learning has the ability to alter the execution strategy based on newly acquired information

The **characteristics that it exhibits that are not well aligned** with the requirements of machine-learning include: (i) a very high cost of errors; (ii) lack of training data; (iii) a semantic gap between results and their operational interpretation; (iv) enormous variability in input data; and (v) fundamental difficulties for conducting sound evaluation.

- Fundamentally, **machine-learning algorithms excel much better at finding similarities** than at identifying activity that does not belong there: the classic machine learning application is a classification problem, rather than discovering meaningful outliers as required by an anomaly detection system.

- In intrusion detection, t**he relative cost of any misclassification** is extremely high compared to many other machine learning applications. A false positive requires spending expensive analyst time examining the reported incident only to eventually determine that it reflects benign underlying activity. Even a very small rate of false positives can quickly render an NIDS unusable .

- Overall, an anomaly detection system faces a much more stringent limit on the **number of errors** that it can tolerate.

- Anomaly detection systems face a **key challenge of transferring their results into actionable reports** for the network operator. In many studies, we observe a lack of this crucial final step, which we term the semantic gap. Unfortunately, in the intrusion detection community we find a tendency to limit the evaluation of anomaly detection systems to an assessment of a system's capability to reliably identify deviations from the normal profile.

- **Network traffic often exhibits much more diversity** than people intuitively expect, which leads to misconceptions about what anomaly detection technology can realistically achieve in operational environments.

- **Evaluation challenges** in terms of the difficulties for (i) finding the right data, and then (ii) interpreting results, exist.

Instead of such difficulties and challenges, Machine Learning can be used if we understand the data and obtain an insight into the operation of the system in terms of its capabilities and limitations from an operational point of view.
Also, we can aim at evaluating and semantic understanding of the data and performance in a better .

REFERENCES:

1) Sommer, R., & Paxson, V. (2010). Outside the Closed World: On Using Machine Learning for Network Intrusion Detection. 2010 IEEE Symposium on Security and Privacy, 0(May), 305–316. http://doi.org/10.1109/SP.2010.25

2)Bhuyan, M. H., Bhattacharyya, D. K., & Kalita, J. K. (2014). Network Anomaly Detection: Methods, Systems and Tools. Communications Surveys & Tutorials, IEEE,16(1), 303–336. http://doi.org/10.1109/SURV.2013.052213.00046

3) García-Teodoro, P., Díaz-Verdejo, J., Maciá-Ferná Ndez, G., & Vá Zquez, E. (2009). Anomaly-based network intrusion detection: Techniques, systems and challenges. Computers & Security, 28, 18–28.
http://doi.org/10.1016/j.cose.2008.08.003