Many insider threat

detection systems, [18], [20], [21], [22], were supported by

DARPA’s project ADAMS [23]

The project aimed to “identify

patterns and anomalies in very large datasets” in order to detect

and prevent insider threats.

In [18], [22], various anomaly

detection algorithms, including Hidden Markov Models and

Gaussian Mixture Models, were employed in an ensemble on

user activity log data for identifying insider threat indicators.

A visual language for describing anomalies was proposed in

the work of [18].

In [20], Eldardiry et al. employed a hybrid

combination of anomaly detectors on several different types

of user activity logs to detect two classes of insiders – blendin

malicious insiders, and insiders with an unusual change in

behaviors. Salem et al. [24] and Toffalini et al. [25] proposed

masquerader detection approaches based on anomaly detection

in user search and file access behaviors.

With this in mind,

the moving weighted average is a common technique to

support anomaly detection under streaming conditions [17],

[28]. In [28], Tuor et al. proposed an anomaly detection

approach using a deep neural network (one model for the

organization), or recurrent neural networks (one model per

user), to produce anomaly scores. Bose et al. proposed a

system employing scalable supervised and unsupervised learning

algorithms on a fusion of heterogeneous data streams to

detect anomalies and insider threats [30]. On the other hand,

Le et al. benchmarked genetic programming algorithms under

propose a user-centered insider threat detection

system, where data analytics is employed on multiple levels

of data granularity under different training conditions. We

evaluate the proposed system on publicly available CERT

insider threat datasets [33]. To the best of our knowledge,

this is the first work comprehensively assessing the effect of

different data granularity levels and training conditions of ML

in insider threat detection. Assuming the point of view of a

security analyst in evaluating suspicious user activities, the

system provides rich information to guide the analysis, where

the results are reported by not only the correctly detected

data instances, but also the correctly identified normal and

malicious insiders, as well as insider threat scenarios. Wellknown

security metrics are adopted and new ones introduced

to show practical performances of the system, while also

presenting insider threat indicators helping to understand the

user behaviors.

the workflow for applying ML techniques is presented

(Section III-A). The workflow is designed to be modular and

easily expandable for a wide range of corporate environments,

data acquisition conditions, as well as learning and analysis

methods. Section III-B details the data collection and preprocessing

steps, where features are constructed and different

levels of data granularity are defined. Finally, Section III-C

presents the ML algorithms used in this research for data

analytics.