Real-time big data processing for anomaly detection: A Survey

Riyaz Ahamed Ariyaluran Habeeb[1], Fariza Nasaruddin[1], Abdullah Gani[2], Ibrahim Abaker Targio Hashem[2], Ejaz Ahmed[3], Muhammad Imran[4]

[1]Department of Information System, Faculty of Computer Science and Information Technology, University of Malaya, Kuala Lumpur 50603, Malaysia.
[2]School of computing & Information Technology, Taylor's University, Subang Jaya, Selangor, Malaysia.
[3]Centre for Research in Mobile Cloud Computing, University of Malaya, Malaysia.
[4]College of Computer and Information Sciences, King Saud University, Saudi Arabia

Email Address: - riyazahd@siswa.um.edu.my, fariza@um.edu.my, abdullah.gani@taylors.edu.my, ibrahimabaker.targiohashem@taylors.edu.my, ejazahmd@ieee.org, dr.m.imran@ieee.org

Abstract

The advent of connected devices and omnipresence of Internet have paved way for intruders to attack networks, which leads to cyber-attack, financial loss, information theft in healthcare, and cyber war. Hence, network security analytics has become an important area of concern and has gained intensive attention among researchers, off late, specifically in the domain of anomaly detection in network, which is considered crucial for network security. However, preliminary investigations have revealed that the existing approaches to detect anomalies in network are not effective enough, particularly to detect them in real time. The reason for the inefficacy of current approaches is mainly due the amassment of massive volumes of data though the connected devices. Therefore, it is crucial to propose a framework that effectively handles real time big data processing and detect anomalies in networks. In this regard, this paper attempts to address the issue of detecting anomalies in real time. Respectively, this paper has surveyed the state-of-the-art real-time big data processing technologies related to anomaly detection and the vital characteristics of associated machine learning algorithms. This paper begins with the explanation of essential contexts and taxonomy of real-time big data processing, anomalous detection, and machine learning algorithms, followed by the review of big data processing technologies. Finally, the identified research challenges of real-time big data processing in anomaly detection are discussed.

Keywords: Real-time, big data processing, anomaly detection and machine learning algorithms.

1. Introduction

The rapid growth of various emerging technologies, such as sensors, connected devices, smart home appliances, smart cities, 5G communication media, smartphones, mobile cloud, healthcare application, multimedia, and virtual reality, and autonomous automobiles contribute to the huge accumulation of real-time data to flow in a network. A report has predicted that, in 2020 the Internet world might witness a massive community of 50.1 billion connected devices (A. Ali, Hamouda, & Uysal, 2015). This expected growth sends alarming signals regarding network security.

The objective of this present study is to provide a comprehensive insight about the state-of-the-art real-time big data technologies, applications and existing anomaly detection techniques. Furthermore, this paper also presents the details of the comparative analysis and the relationship of three different domains, which are anomaly detection, machine learning algorithms, and real-time big data processing. It also presents a complete taxonomy based on the comparative analysis of above mentioned domains. Another motive of this paper is to identify and discuss the challenges of real-time big data processing for anomaly detection.

A research has analysed the data collected from 70 organizations located in 61 countries, and has found that the most number of network attacks has been reported from the industries such as, public, information, and financial services with total of 79,790 security incidents, and 75% of the attack spread from 0 victims to 1 victim within a given day (DBIR, 2015). Furthermore, various types of cyber-attacks have been identified in the network infrastructure. For example, spamming, search poisoning, botnets, denial of services, distributed denial of services, phishing, malware, website threats, and credential comparison (DBIR, 2015).

Of late, monitoring the networking threats has become the biggest challenge for most organizations, especially in the prominent sectors, such as government, energy, healthcare, banks, and research centres. These organizations spend huge amounts of money to protect and secure the infrastructure using various monitoring tools. However, the existing security tools and analysis of logs to detect the attackers in offline mode wear out in the long run and become obsolete, because, attackers use sophisticated techniques to penetrate the infrastructure.

Various large-scale complex cyber-attack on connected devices have been reported in 2016 (Almeida, Doneda, & de Souza Abreu, 2017). Hackers infected thousands of Internet-connected devices, such as, webcams, recorders, and other connected devices. Cyber attackers used household items and devices to launch an attack on major websites in U.S (Hunt, 2016). The growth of the Internet of Things (IoT) has opened the gate for an enormous amount of data generation and flow, which has laid a roadblock for the network infrastructure security monitoring processes (Yaqoob et al., 2017). Recently, SpiderLabs reported that massive cyber-attack in Singapore healthcare data centre has resulted in the theft of 1.5 million patients' records from the data centre. However, the data theft was detected and reported only after few days by database administrators. They detected abnormal activity on one of the SingHealth's IT databases which lead to alert and shutdown the entire network to stop the further data breach (Chow, 2018).

In order to ensure network security, anomaly detection is regarded as one of the powerful mechanisms, which helps the security analysts to identify and detect any possible threats, which may occur in the next few days or months in the network. In particular, anomaly detection leads to early detection of irrelevant pattern or unusual events in the network. It monitors critical network attributes in real-time and produces an alarm if any anomaly or abnormal trends are detected in the network.

The massive amounts of data generated in real time have inhibited the performances of network analysts in terms of scaling the abundant volume of data. Hence it is crucial to produce network security analytics performance reports in the real and near real time, and not just generated from existing monthly and weekly log data. Nonetheless, the existing anomaly detection techniques are unable to process and detect the threats in the network in real time or near real time and fail to produce adequate security analytics performance reports. Furthermore, increasing numbers of new types of attacks are found in networks, every day, but the current monitoring tools have become incompetent to detect those attacks due to the huge volume, velocity, variety, and veracity of data received for analysis.

The era of big data has posed many challenges and problems in network security that remain unaddressed in sectors, such as healthcare, manufacturing, telecommunication, banking, education, and transportation, particularly, the areas of concern are network architecture, modelling, data processing and analysis, and anomaly detection. It is crucial to immediately process the data collected to detect any potential threat in the network, however, the existing traditional monitoring tools are incapable of handling big data, and therefore struggle to continuously monitor network infrastructure and detect the anomaly behaviour and threats (Raguseo, 2018). Hence, it has become essential to overcome the bottlenecks of the existing techniques by employing a hybrid of machine learning algorithms associated with big data technologies, to efficiently process real-time big data to detect anomalies. Machine learning knowledge will help the analysis of collected data to detect and monitor the network with the use of algorithms. Furthermore, various real-time big data technologies can help to process and stream the enormous amount of network data in real time and near real time.

Figure 1 illustrates bottom up sequence of real-time big data processing for anomaly detection, where various smart devices are communicated via network technologies. Such devices generate a lot of sensor data, which are stored in cloud and other storages devices. These stored datasets collected from sensors devices are processed with big data processing technologies, such as, Hadoop, Spark, Apache storm, and the results are used for analysis and anomaly detections using machine learning algorithms.

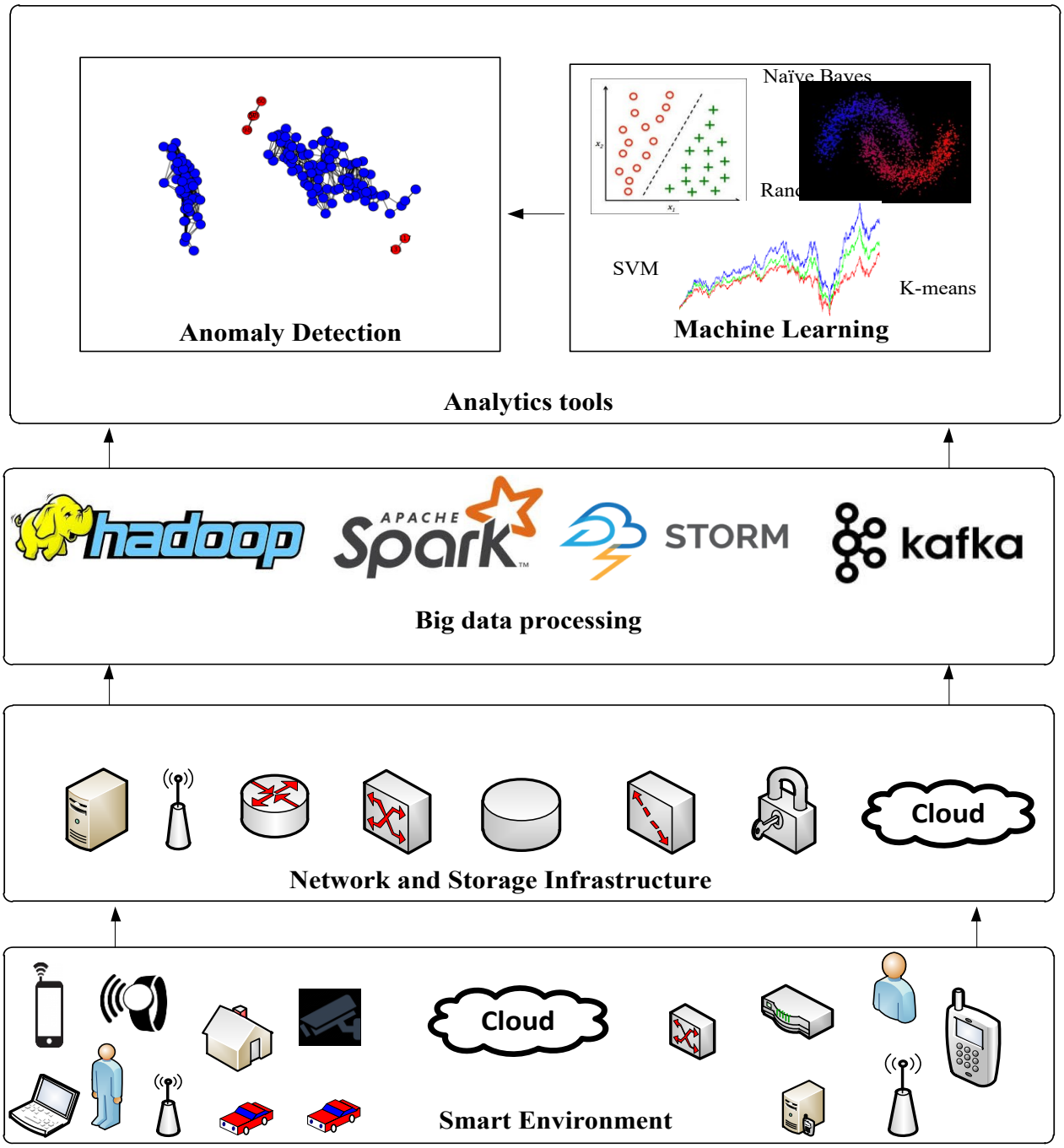

Fig. 1. The sequence of real-time big data generation, processing and anomaly detection

Numerous studies have investigated the growth and use of big data processing technologies on various industries (Hashem et al., 2015) (Gani, Siddiqa, Shamshirband, & Hanum, 2016). In which, 48 indexing techniques and performance analysis for big data processing have been investigated(Gani et al., 2016). However, these studies have not focused on real-time big data processing on network anomaly detection with machine learning. On the other hand some studies have investigated big data applications, processing, technologies, and big data lifecycle, open issues, and challenges (Chen, Mao, & Liu, 2014). Similarly, four major different types of anomaly detection techniques have also been critically analysed(M. Ahmed, Mahmood, & Hu, 2016). Likewise, various machine learning and data mining techniques for cyber analytics were surveyed and the complexity of algorithms were discussed (Buczak & Guven, 2016).

Table 1. Summarizes the main focus and objectives, limitations, how our survey differ from existing works, journal or conference and lastly number of papers covered for their studies.

Year	Focus / Objectives of Previous studies	Journal/ Conference	Limitation	Number of papers	How our study differs.	Reference
2013	Comprehensive survey on security analytics in the aspect of technologies, trends and tools.	Conference	No comparison and critical analysis were included for big data technologies.	24	Relationship among security analytics, machine learning and big data technologies has been presented.	(Mahmood & Afzal, 2013)
2014	Data generation, data acquisition, data storage, and data analysis.	Journal	Lacks critical analysis on big data technologies.	156	Relevance between big data technologies and machine learning have been emphasised.	(Chen et al., 2014)
2015	Use of big data for forecasting threats by identifying, reviewing the problems and challenges	Journal	Real-time processing has not been focused.	96	Real-time big data processing for anomaly detection is one of the main focuses.	(Hassani & Silva, 2015)
2015	Hadoop ecosystem tools and tools selection criteria against machine learning tools are reviewed.	Journal	Commercial big data technologies were not reviewed and machine learning algorithms were not focused.	93	Open source and commercial tools have been analysed. Real-time big data processing for anomaly detection.	(Landset, Khoshgoftaar, Richter, & Hasanin, 2015a)
2015	Various cloud environment used to analyse big data applications	Journal	Review on existing big data technologies such as Hadoop, spark, storm, and etc.,	140	Comparative analysis on various real-time big data technologies.	(Assunção, Calheiros, Bianchi, Netto, & Buyya, 2015)
2015	Analysis of graph based anomaly detection techniques.	Journal	Use case and algorithms were very limited.	262	Various real-time algorithms and application for anomaly detection have been reviewed.	(Akoglu, Tong, & Koutra, 2015)
2015	Review of four data mining techniques for anomaly detection used for HTTP web services.	Journal	Lot more algorithms needs to be compared against various types of data.	31	Various real-time algorithms and application for anomaly detection have been reviewed.	(Kakavand, Mustapha, Mustapha, Abdullah, & Riahi, 2015)
2016	Various machine learning and data mining methods used for cyber analytics in support of intrusion detection were reviewed.	Journal	Real-time application and algorithms were not covered.	113	Various real-time algorithms and application for anomaly detection have been reviewed.	(Buczak & Guven, 2016)

2016	Analysis of major anomaly detection techniques.	Journal	Real-time applications and algorithms were not covered.	100	Various real-time algorithms and application for anomaly detection have been reviewed.	(M. Ahmed et al., 2016)
2016	Focus on industrial viable anomaly detection and machine learning algorithm used in real-world network intrusion detection.	Conference	Minimum numbers of algorithms were covered for their study	55	Various real-time algorithms and application for anomaly detection have been reviewed.	(Gilmore & Haydaman , 2016)
2016	Review on existing and emerging technologies on real-time IoT data streams.	Journal	Lack of critical analysis on big data technologies.	64	Security analytics, machine learning and real-time big data technologies have been emphasised.	(Yasumoto , Yamaguchi, & Shigeno, 2016)
2016	Various machine learning techniques for big data processing were reviewed.	Journal	Lack of critical analysis on real-time big data processing.	129	Real-time big data processing for anomaly detection has been emphasised.	(Qiu, Wu, Ding, Xu, & Feng, 2016)
2018	Current and future trends on both big data and cognitive computing were reviewed.	Journal	Big data technologies were not covered.	95	Real-time big data processing for anomaly detection has been emphasised.	(Gupta, Kar, Baabdullah, & Al-Khowaiter , 2018)

It is significant to mention that, our study is different from the above-mentioned studies in several practices, where most of the existing analyses have generally evaluated big data processing, anomaly detection or machine learning techniques mainly focused on batching processing rather than real-time processing. In contrast, we have mainly focused on real-time big data processing technologies using machine learning for anomaly detection.

The main contributions of this paper are:

i. Surveying state-of-the-art researches focused on real-time big data processing for anomaly detection.
ii. Proposing a taxonomy to classify existing literature into a set of categories such as, big data technology, anomaly detection, machine learning techniques, modes, data, and application.
iii. Analysing existing solutions based on the proposed taxonomy.
iv. Highlighting the research challenges and prompting guidelines for future researchers.

The paper is structured as follows: Section 2 explains the research motivation and use cases for real-time anomaly detection. Section 3 introduces the anomalies detection, real-time big data processing, machine learning with their application areas. Section 4 compares the state-of-the-art real-time big data processing, analyses and synthesises the limitation of anomaly detection and machine learning algorithms. It also classifies different big data technologies used for performing security analytics, and also identifies the categories of machine learning techniques for anomaly detection. Section 5 demonstrates the taxonomy of real-time big data processing technologies for anomaly detection. Section 6 details the research challenges in real-time anomaly detection. Finally, this paper concludes by summarising and highlighting the future direction.

2. Motivation and Use cases

In the following sections we have described the motivation and few use case scenarios to motivate the survey of anomalies detection through real-time big data processing.

According to report by, Cisco ("Cisco VNI Forecast and Methodology, 2015-2020," 2016) forecasted that, 2.3 Zettabytes of Internet protocol(IP) traffic would go across the Internet in 2020, which will be 879 Exabyte more from 2015. This leads to a lapse in the existing security analytics to detect the threats in real time. Furthermore, Cisco also reported that 71 percent of total IP traffic in 2020 is expected to come from non-PC devices (smart devices) such as, tablets, smart watches, smartphones, smart bands, video game consoles, television set-top devices, smart key chains, smart bulbs, smart security cameras, smart TVs, and smart locks (Kerner, 2016). This leads to a huge volume of data to be analysed in real time with high velocity and more varieties, which fulfil the characteristics of big data V's.

Meanwhile, non-PC devices pose a huge security threat if they are not monitored in real time. In October 2016, large organizations such as, CNN, Twitter, Reddit, The Guardian, and Netflix in US and Europe were massively attacked via smart home devices (Woolf, 2016). In addition, new threats are expected to emerge in 2021 as hackers find new ways to attack smart devices and protocols (Jones, 2016).

According to Gartner (2017) the expenditures towards worldwide security spending might reach an amount of $96B by the end of 2018. Universally, organizations will become extra conscious of the security risk, due to the inefficient and inadequate protection against attacks by the existing technologies. The U.S. Federal Cybersecurity Market Forecast has been estimated to reach $22 Billion by 2022, which will constitute a steady compound annual growth rate of 4.4% (Media, 2017). Another report suggests that User and Entity Behavioural Analytics (UEBA) and Deep Learning algorithm designs are emerging as the two most prominent technologies in cybersecurity. Further, this will drive the new norms of machine learning solutions to replace Security Information and Event Management (SIEM) of traditional anti-virus within the next five years (Security, 2017). It has also been highlighted that in new generation applications, data stream processing has emerged as one of the potential research areas, in which, data continuously flows into the processing site. In data stream, real-time data constantly arrives from the external source with certain time intervals. Input data can change between a few bytes per seconds to gigabytes per second (Di Mauro & Di Sarno, 2014).

The above discussion, followed by the use case scenarios presented below have motivated this present paper to focus on real-time anomaly detection.

2.1. Modern network traffic scenario: Present IoT infrastructures use various connected and mobile devices, and its machine-to-machine communication generates large scale of sensor data every second, and the generated data are stored in the cloud. These data are heterogeneous in nature with a variety of parameters such as, IP address, data transfer speed, volume, etc. (Xie & Chen, 2017). The heterogeneous data generated from these devices should be monitored and collected in real-time to be patterned for detecting abnormal behaviour.

2.2. Mobile Cloud

Smartphones provide a wide variety of sensors integrated in one device, allowing monitoring, processing, measuring, locating the device anytime. Multimedia sensors like the microphone, dual-camera, and finger print sensors integrated in smartphone allow to employ this devices in a wide variety of applications (García, Tomás, Parra, & Lloret, 2018). All these data generated from sensors needs to be stored and processed in cloud due to resource shortage in the mobile device (Karim et al., 2017). Cloud computing technologies for mobile devices offers an innovative method of delivering IT services efficiently like IT services that are available at all times and from all location (O. Ali, Shrestha, Soar, & Wamba, 2018).

Particularly, cloud computing helps mobile health services to provide access for electronic healthcare system. For example, remote data processing and monitoring, remote consultation, and digital multimedia data (Y.

Zhao, Ni, & Zhou, 2017). Moreover, all these services contribute for the growth of big data in their services. In recent days, many of the cloud service providers like Amazon, Google cloud, Microsoft, IBM, Oracle and others major organizations are adopting big data technologies into their cloud platform for superior process, storage and analysis. All these data in the cloud needs to be monitored in real-time to detect any abnormal behaviour such as data theft, patient illness, origin or separating of new diseases from different demographic, cloud infrastructure such as memory usage, power consumption, cooling system and many more.

2.3. Autonomous vehicles scenario: These driverless vehicles redefine the automotive industry ecosystem accompanied by enormous big data generated by connected devices in vehicles(Ger, 2017). Attackers have begun to target these connected vehicles data to hack the vehicles, which can help them to control the entire system, and disable the vehicle anytime. Now, most of the modern vehicles are equipped with a number of modules, such as in-vehicle networks (IVN) including, engine controls units (ECU), body control modules (BCM), smartphone integration module, which provide critical functionalities for control and safety of the vehicles ("Symantec™ Anomaly Detection for Automotive," 2016). These modules need to be analysed in real time to detect the anomalies in the vehicles, which includes sudden increase of speed, radar sensors detection, camera sensing, abnormal petrol consumption, sudden engine failure, inappropriateness in changing lanes, and inaccurate object detection.

2.4. Healthcare scenario: Real-time anomaly detection helps in monitoring services of the patients to detect the anomalies in real-time and timely manner. This helps the hospital and caretakers to make a wise decision, especially as elderly people who are living alone are becoming a social problem for community and government(Yasumoto et al., 2016). The data can have several abnormal patient conditions or instrument errors, human errors, or focus on detecting disease outbreaks in a specific area. These records consist of various types of features such as, patient age, gender, height, sugar level, blood details, which need to be analysed with high accuracy (Chandola, Banerjee, & Kumar, 2009). In addition, growing numbers of compatible IoT devices help healthcare industry to collect and analyse massive data.

Furthermore, current medical technologies produce various types of multimedia data such as high quality videos and image, graphs and sound files. These multimedia data contain rich and complex information, which help for diagnosing and monitoring any disease. Moreover, computing requirements of multimedia solutions for healthcare have led to the employment of cloud services for e-healthcare system (García et al., 2018). On the other hand, incorporating multimedia data into electronic health records face big data challenges and these data needs to be monitored in real-time to detect any abnormal behaviour in the system.

2.5. Insider Trading Detection: In stock markets, data changes in milliseconds and anomaly detection techniques have been used to detect the insider trading early. People make illegal profits by leaking the inside information before the actual information is made available to the public. The information could be of pending merger, acquisition, a terrorist attack, judgment on the particular industry or any other relevant information that affect the stock prices of any specific industry. Insider trading can be detected by identifying anomalous trading activities in the market. It has to be detected in real time manner to prevent people from making illegal profits (Chandola et al., 2009).

2.6. Safety Critical Detection: For safety-critical system, attackers begin targeting mobile connected vehicles, and vehicle-to-vehicle communications networks to enter into the controls units and body control modules, which provide critical data about the vehicles. Besides, these mobile devices store, process and access critical data from the cloud infrastructure. Here anomalies detecting techniques can help to monitor or notify the vehicle at what time it has been attacked, or just malfunctioning.

Conversely, all the above use case scenarios reveal that there are still challenges and difficulties in using existing anomaly detection techniques. Furthermore, every day increasing numbers of new types of threat are found in the connected devices. Current monitoring technologies are challenged to detect the anomalies because of the growing volume, variety, and velocity of data received for the analysis.

3. Overview

In this section, an overview of anomalies detection, real time big data processing and machine learning is provided to offer the fundamental knowledge on these topics.

3.1. Anomalies detection

Anomaly detection is an important problem that has been researched within diverse research areas and application domains. Anomaly detection refers to the process of finding patterns in data that do not conform to expected behaviour. These non-conforming patterns are often referred to as anomalies, outliers, discordant observations, exceptions, aberrations, surprises, peculiarities, or contaminants in different applications domains. Anomaly detections find extensive use in a wide variety of applications such as, networking monitoring, healthcare, smart devices, smart cities, Internet of things, fraud detection, cloud, and much more. For example, an anomalous traffic pattern in a computer network could mean that a hacked computer is sending out sensitive data to an unauthorized destination (Chandola et al., 2009). In particular, network monitoring helps to detect threat from various network infrastructure elements, such as MAC spoofing, IP Spoofing, TCP/UDP fanout, Duplicate IP, Duplicate MAC, Virus detection, bandwidth anomaly detection and connection rate detection. Additionally, anomaly detection helps to track the profiles of normal day to day activities of every system, application, or network. Drawbacks of the existing anomaly detection techniques are possibilities of high false alarm rates based on unknown system behaviours (Buczak & Guven, 2016). There are three different categories of an anomaly (See table 2). Of which, mostly the collective category is used for the real-time or online anomaly detection.

Table 2:
Categories and description of anomaly

Categories	Description
Point Anomalies	If an individual data instance can be considered as anomalous with respect to the rest of data, then the instance is termed as point anomaly (Chandola et al., 2009).
Contextual Anomalies	If a data instance is anomalous in a specific context, then it is termed as contextual anomaly and also referred as conditional anomaly. Each data instances are classified into either contextual or behavioural attribute. (Chandola et al., 2009).
Collective Anomalies	If a collection of related data instances is anomalous concerning the entire data set, it is termed as a collective anomaly. Data instances in a collective anomaly may not be anomalies by themselves, but their occurrence together as a collection is anomalous (Chandola et al., 2009)

Various considerations can be given for anomaly detection in the system such as, timeliness, the rate of changes, scale, conciseness and definition of incidents. We have surveyed various anomaly detection techniques, which have been precisely developed for latest application domains, like, image processing, autonomous vehicles, smart homes, flight monitoring, healthcare, network monitoring, sensor networks, fraud detection for safety critical system and credit card systems, and many others. Over time, numerous anomaly detection techniques have been established by various research centres and industries, but, many factors make anomaly detection approach very challenging such as, to find the exact gap between the normal behaviour and abnormal behaviour for the different areas. Moreover, different application domain has a different variant on the data. In our research, we have critically analysed various anomaly detection techniques used for monitoring network traffic pattern. These traffic patterns have to be monitored in real-time to detect any abnormal behaviour in the network. However, with growing number of IoT devices and smart connected applications, which transmit multiple data between network with massive volume, variety and high velocity, which makes it more challenging for the existing anomalies detection techniques to detect the threat. In Figure 2 we present a timeline of the historical evolution and trends for anomaly detection techniques and big data technologies.

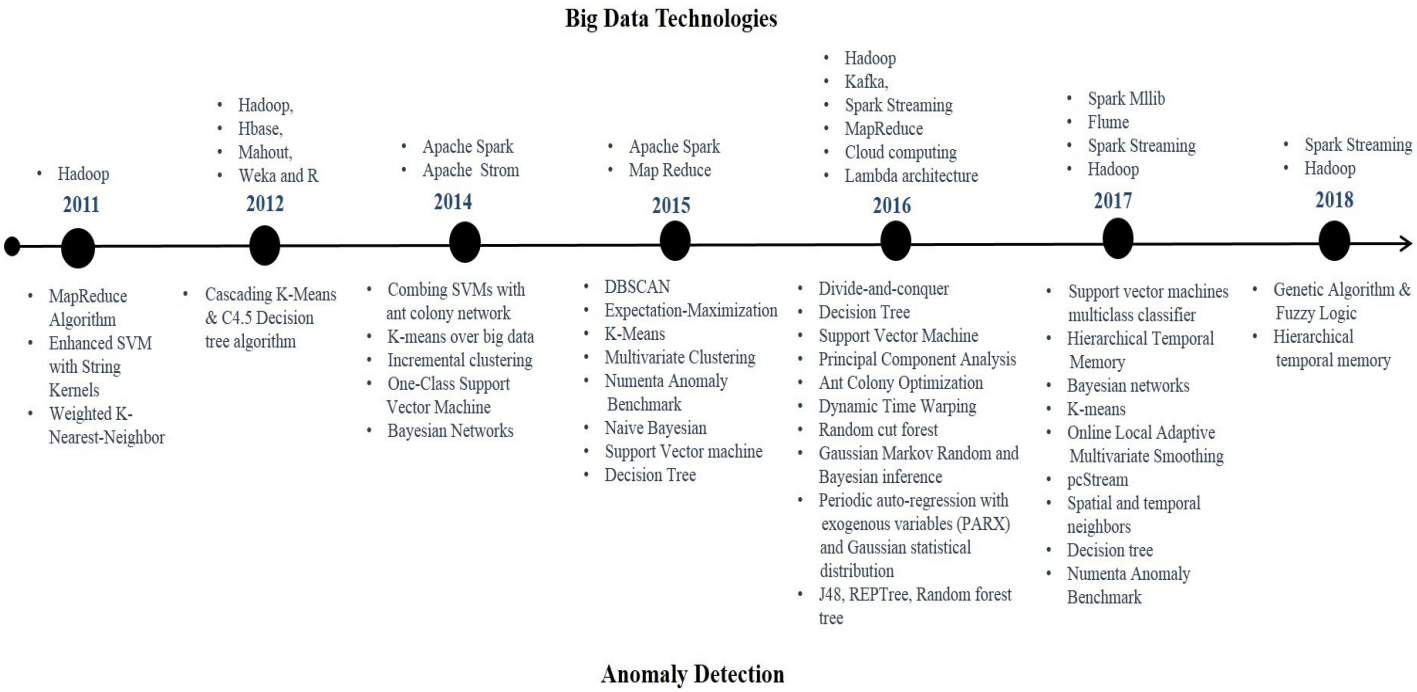

Fig.2. Historical evolution and trends of anomaly detection techniques and big data technologies.

3.2. Real-time big data processing

According to (Cloud Strategy Partners, 2015), real-time big data processing is critical than any other processing application, because it is required for the uninterrupted monitoring of event, messages, processes in network infrastructure. Further fast data are generated for network monitoring from hardware and software. For example, the log file that can be rapidly changing in-memory data set. In fast data, data dynamically changes in certain time intervals that can be seconds and milliseconds. The huge quantity of data that arrive continuously to the pipeline can be in any format such as, structured, unstructured, and semi-structured. These data contain the detailed information about the messages and events. Streamed data are positioned in the big data analytics for processing, and then big data analytics will help to make the analysis and decision for further process. Adoption of streaming architecture will guarantee the efficient and seamless communication between the sensing devices and network (Hashem et al., 2016). A Large amount of real-time data can be processed with following tools Storm, Splunk, S4, SAP Hana, Spark, which were compared with its strengths and weakness (Yaqoob et al., 2016).

With connected devices continuously collecting, processing, and storing massive amounts of data, shows that we are living in the era of big data. (Hashem et al., 2015) have defined a set of techniques and technologies that require new forms of integration to uncover large hidden values from the big data that are diverse, complex and of a massive scale. Variety and various structures of big data have been generated from online transactions, emails, audios, videos, search queries, instant messages, social networking interaction, health records, images, click streams, logs, posts, mobile phones and application scientific equipment, and sensors(A. Ahmad, Paul, & Rathore, 2016). The term big data is coined to describe information as a flood of data and their effect is being felt everywhere, from business to science, and from the government to the arts (Slavakis, Giannakis, & Mateos, 2014). Globally, the volume of data is expected to grow 40% per year, and 50% by 2020. Walmart alone processes and imports more than 1 million customer transactions into databases, estimated to be more than 2.5 PB of data each hour. According to (Yaqoob et al., 2016) the current amounts of the Internet of things data can be considered small, as of 2030, wherein the number of sensors is expected to reach 1 trillion, at the time IoT data will become big data. Many applications areas face challenges from big data properties, including, network monitoring and detection, geospatial data, vehicular traffic, market prediction, and business forecasting.

In recent years, due to the accumulation of massive volumes and variety of data from different sources, a number of industries and institutions across the world are moving towards the real-time response, which is regarded as a crucial decision. Essentially, due to the amassment of constantly moving data from connected devices and sensors, it has become crucial to make decisions in real time, rather than on weekly or monthly basis. To extract extreme value from

moving data, organizations need to process data much faster and take timely actions. The real-time event processing improves correlation and pattern detection on a scale of millions of events and constantly moving data streams at microsecond speeds. Furthermore, the technological enhancement in the connected devices has enabled them access to very high Internet speed, which has reached gigabytes and terabytes(Rathore, Ahmad, & Paul, 2016b). High-speed Internet services have elevated the reach and services in various sectors, such as healthcare, e-commerce, marketing, banking, vehicular sensors and others involved in online transactions. These industries have an enormous number of applications, which deal with real-time data that are constantly generated and keeps on changing. However, it has been an exponential increase in the availability of streaming and time series data. These real-time streaming data have to be processed and stored in time to receive value and insights from them, especially when it comes to real-time network data to detect the anomalies behaviours. Big data infrastructure needs to support the whole big data lifecycle including, data collection, filtering, processing, storing, and visualization (Larson & Chang, 2016).

Most of the organizations have difficulty in analyzing and processing the immense amount of collected data. Furthermore, traditional tools take too long time to process the data, low throughput and are also unable to come across the data that generated with velocity, variety, volume, and veracity. It will be inevitable for Small medium enterprises (SME) to process the data, as they face problems in storing the collected data. . This data storage problem within SMEs can be solved using a cloud service provider. In data centre: to monitor the parameters and to detect their issues, data needs to be collected and aggregated in real time, which helps to indicate the threats or abnormal behaviour of the centres, which incredibly challenges the real-time data sources to process (Ellis, 2014). The collection and processing of real-time data have varieties of application areas, such as operational monitoring, anomaly detection, web analytics, online advertising, social media, mobile data, smart home appliances, smart city, biometric, and much more (Ellis, 2014) (S. Ahmad, Lavin, Purdy, & Agha, 2017) as shown in Figure 3.

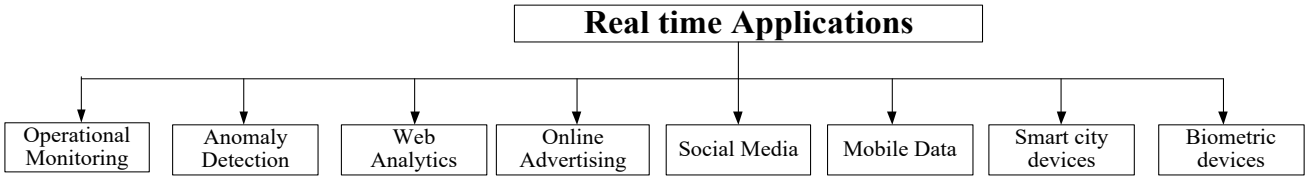

Fig. 3. List of Real-time applications areas

3.2.1. Operational Monitoring: The power of the Internet and its reach has made operational monitoring as the most important aspect of monitoring streaming data. Operational monitoring is a process that tracks the performance of any physical system connected to a network. In data centres, operational monitoring continuously captures the physical states of the systems, such as temperature, the speed of the fan, utilization of their power supplies, disk drives, processor load, network activities, and storage access time (Ellis, 2014).

3.2.2. Anomaly Detection:

The detection of anomalies in real-time data has practical and substantial applications across several industries. Detecting anomalies can give actionable information in critical scenarios, but reliable solutions do not yet exist (S. Ahmad et al., 2017). Some of the use cases were discussed in session 2 for anomalies detection applications. In real-time anomaly detection, statistical techniques are frequently used due to their computational flexibility. Some of these techniques contains sliding thresholds, outlier test, such as extreme studentized deviate and k-sigma, change point detection, statistical hypotheses testing, and exponential smoothing. The eccentricity anomaly detection techniques are more effective than other techniques, because there is no need to define parameters for analysis in the former (S. Ahmad et al., 2017).

3.2.3. Web Analytics: Growth of various commercial websites like, e-commerce, online newspapers and others demand tracking events and activities of those websites in real time, which will enable organizations to get the insight of the visitor's preference on their website, and as well as number of visitors to a specific product page and their correlation (Ellis, 2014). Most of the web analytics data are stored in the cloud and accessed by mobile devices.

3.2.4. Online Advertising: Online advertising is one of the major contributors for the real-time data. In which, administrators track various metrics, such as, number of purchases, number of clicks on any specific advertisements, time of arrival at a website via a modern advertising exchange and also which places bids on the page view. All these data are collected, moved, analysed and stored instantaneously in real time to produce better services and profit for the organization(Ellis, 2014).

3.2.5. Social Media: Social media is also termed as newspapers of the younger generation, the omnipresence of social media, such as Twitter, Facebook, Instagram to name a few, contribute to the amassing of millions of tweets, updates, and comments. These data are collected and disseminated in real time, which makes them as an important source of information for news outlets and other consumers around the world. The data involved in this kind of platforms are mostly unstructured which have to be parsed, processed and somehow understood by automated systems (Ellis, 2014).

3.2.6. Mobile Data: Mobile data or cellular data enable the smartphone users to gain access to wireless Internet. Mobile data can help to get an insight of the origin-to-destination matrix, which facilitate to figure out the origin of people and their whereabouts in a given area of the Internet. Additionally, the headcount of a population group in given location helps in deriving a density map of the mobile area (Ellis, 2014).

3.2.7. Smart city devices: Wireless communication technologies in the smart city can help to monitor physical and atmospheric conditions in real time, such as temperature pressure, light, humidity, traffic, and others (Hashem et al., 2016). It has also been highlighted that adoption of real-time architecture for the smart city will assure effective and seamless communication among sensing devices within the smart city infrastructure. It also includes the quality of services support in the network, which is extremely crucial for the real-time application for smart cities.

3.2.8. Biometric devices: Biometric devices are used to monitor patients' health data in real time to make better and timely decisions (Carvalho, Rocha, Vasconcelos, & Abreu, 2018). A study has proposed an approach to stimulate the real-time falling action of humans, which can be collected by the accelerometer and triaxial gyroscope to detect the occurrence of a fall (Zhang, Lai, Lai, Wu, & Chao, 2017). Another research has proposed an approach to transfer the information from ICU devices in real time to the doctors in a remote location (Almadani, Saeed, & Alroubaiy, 2016).

The above sub-sections had summarized the importance of real time processing and its general difficulties. Furthermore, parameter and adoption of real time architecture or approach for specific application areas had also been addressed.

3.3. Machine learning

The main objective of the machine learning is to allow a system to learn from the past or present and use the knowledge to make predictions or decisions regarding unknown future events. (Landset, Khoshgoftaar, Richter, & Hasanin, 2015b). Machine learning can be applied to different industries, like banking, autonomous car, manufacturing, retail industry, marketing, networking, and general science, including chemistry, physics, medicine, bioscience, pharmaceutical, insurance, energy, and sustainability. Several machine learning algorithms have been proposed and used for mining meaningful information from the data through preparation and validation using categorized datasets. These algorithms are classified into two major categories, such as supervised and unsupervised. In real-time applications, the machine learning algorithm needs to analyse a continuous sequence

of data occurring in real-time. When compared to batch processing, the entire dataset is not available in runtime. Furthermore, real-time application requires to process data in sequential form as they arrive, and make decisions online (S. Ahmad et al., 2017). Manipulating and modifying the existing machine learning system will satisfy the market needs to increase the conservation of energy and increase the computational cost (Al-Jarrah, Yoo, Muhaidat, Karagiannidis, & Taha, 2015). Traffic among machines has become an essential portion of today's network environment and will escalate even more in the near future. It is expected to produce exceptional traffic patterns that will challenge network administrators to learn and track the threat in the network. Existing machine learning techniques are incompatible for addressing big data classification problems, and are incapable of handling unstructured data, which are essential to produce high accuracy for high-velocity data, and ineffective for multiple learning tasks and also for computational efficiency. Computational complexity has exponentially risen in high dimensional data that fail to fulfil current needs.

In this contest, various machine learning algorithms, such as Nearest Neighbours, Bayesian Networks, Support Vector Machines, Decision Trees, Random Forest, Ant Colony Optimization, Fuzzy logic, Principal Component Analysis are discussed in this paper (see sections 4 and 5).

The above sections had discussed the importance of anomaly detection, real-time big data processing, machine learning and noted most used algorithms for anomaly detection.

4. State-of-the-Art machine learning-based big data processing technologies for anomalies detection

This section consists of four major parts, the first part presents a comprehensive review on the state-of-the-art anomalies detection techniques. In the second part a comparative analysis is provided on the state-of-art big data and anomaly detection and the third part presents details of big data and machine learning techniques. Lastly a comprehensive review is presented about anomaly detection, big data, and machine learning techniques.

4.1. Current anomalies detection techniques

This sub-section critically analyses various techniques used for anomalies detection. A combining Support Vector Machine (SVM) with ant colony network has been designed to produce a high performance intrusion detection system, which contains parallel phases to detect the anomaly in real time (Feng, Zhang, Hu, & Huang, 2014). Combining SVM with ant colony achieves better performance in accuracy rate and faster running time than other traditional methods. Nonetheless efficiency of the algorithms needs to be compared with existing techniques to show the major different between the proposed designs.

In addition, pcStream algorithm based framework has been used to evaluate three different types of attacks to detect the data leakage, malware, and device thefts. It is a stream clustering algorithm, mainly applied for dynamically detecting and managing temporal contexts. Choosing parameter can be demanding for very extensive training time. The framework uses non-exhaustive grid search over this parameter to establish network parameter (Mirsky, Shabtai, Shapira, Elovici, & Rokach, 2017). Nevertheless, there is a need to determine the proficiency of the proposed framework to be in line with current approach of anomaly detection.

Meanwhile, a robust random cut forest technique is used to perform dynamic data stream over anomalies detection, which treat different dimensions independently (Guha, Mishra, Roy, & Schrijvers, 2016). Its helps to preserve pairwise distance, which will be important for computation and likewise for anomaly detection. Consequently, the algorithm result shows great promise to fight against the false alarm. However, dataset size is the one of the constraints for robust random cut forest technique.

In one class support vector machine (OC-SVM) (She, Wen, Lin, & Zheng) high unbalanced problems of classification is used to find the positive data for detection of DDOS attack in application layer. OC-SVM abstract sorts' data from users' session and cluster these data to build user behavioural model. Meanwhile restriction of OC-SVM is in low acceptance level for feature selection.

Another one-class support vector machine algorithm (Maglaras & Jiang, 2014) is trained to trace the network offline and also detect anomalies in the real time. OC-SVM is an addition to the support vector algorithms to the

occasion of unlabelled data, exclusively for detection of outliers. It assists the map input data into a high dimensional feature space and continual finding of maximal margin hyperplane, which utmost splits training data from the original source. Yet, evaluated dataset were smaller in size as compared to the openly available larger datasets.

Further, Cluster center and nearest neighbour (CANN) algorithm is a proposed model, which combines abilities for identifying both similar and dissimilar classes for a given dataset. Moreover, it increased effective and efficient manner of anomalous detection. Its contains three stages of techniques: (i)clustering technique to extract cluster center, (ii) measure and sum the distance between all the data of the given dataset and, (iii) cluster centers (W.-C. Lin, Ke, & Tsai, 2015). On the other hand, there is a low acceptance level for feature demonstration for a better quality in the pattern detection.

Besides, Expectation maximization algorithm is utilized to model an anomaly detection using structural time series for the industrial Ethernet traffic system. The system decomposes the traffic into four components, based on a model, which has definite meaning for detection (Lai, Liu, Song, Wang, & Gao, 2016). This model helps to improve the performance of measuring abnormal and low false alarm rate. Although, updating the parameter is a complex process.

In addition, Hybrid of one-class support vector machine and deep belief nets (Erfani, Rajasegarar, Karunasekera, & Leckie, 2016) is a model for anomaly detection with high dimensional and large-scale dataset. This architecture has significant detection rate and eases the computational complexity of training and testing the model. It also helps to approach the complexity and scalability problems of one SVM. Despite the ability to provide detection in high dimension, the architecture is restricted in non-convex loss function.

Online local adaptive multivariate smoothing (Grill, Pevný, & Rehak, 2017) is to engage the high false alarm rate in network intrusion detection system. HTTP proxy logs dataset are collected from different companies' network to do the evaluation. In this technique, true anomaly score are improved concurrently detecting the time and space. However, inefficiency in selecting more accurate parameters to set for algorithm is a challenge.

Self-organizing maps (C. Yin, Zhang, & Kim, 2017) is one of the well-known clustering algorithms, which is used to detect the anomalous threat in mobile devices. The process consists of three types of evaluation benchmarks, which are accuracy rate, precision rate, and recall rate. The result proves that improved SOM can produce higher accuracy rate for openly available KDD Cup99 datasets. At the same time, it consumes longer time to find the initial weight vector and proposed setup should be evaluated with mobile devices.

Moreover, enhanced support vector (Ramamoorthi, Subbulakshmi, & Shalinie, 2011) is famous for detecting anomaly in real time applications. It provides the better classification accuracy for incoming flows, as an attack or normal flow. This model increased the classification accuracy using weight assignment for real-time instances. Nonetheless, the complexity in selecting the parameter for proposed algorithms is a bottleneck.

The hidden semi-markov model (Bang, Cho, & Kang, 2017) is used to detect anomaly in wireless sensor actuator network, based on the Long-Term Evolution (LTE) signalling traffic. This model improves the detection sensitivity result and shows more attack alarms with false positive and true negative ration. Despite that, longer time duration and unknown size for training dataset makes it complex to perform.

Lastly, Hierarchical temporal memory (Lavin & Ahmad, 2015) based model is used to test and measure open source data for anomaly detection on streaming data. In this model, Numenta anomaly benchmark broad and solid assessment tools are used for real-world anomaly detection. Regardless, the efficiency of the proposed techniques were not provided on the basis of performance and false detection.
In the above section we had discussed various anomaly detection techniques from different domains. All these techniques have their own strengths and challenges in terms of anomaly detection. Especially, clustering and support vector machine techniques provide more promising accuracy, compared to other techniques. However consuming longer time duration for training model and inefficiency in choosing the specific parameters are the challenges of the model.

Table 3: Different anomalies detection techniques

Domain	Techniques	References
Smartphone	pcStream algorithm	(Mirsky et al., 2017)
	Self-organizing maps	(C. Yin et al., 2017)
Network & Wireless network	Hybrid of One-class Support Vector Machines (1SVM) & Deep Belief Nets(DBN)	(Erfani et al., 2016)
	Combines a Support Vector Machine with an Ant Colony Network	(Feng et al., 2014)
	One-Class Support Vector Machine.	(Maglaras & Jiang, 2014)
	Block-based One-Class Neighbor Machine and Recursive Kernel-based online	(T. Ahmed, Oreshkin, & Coates, 2007)
	Online Local Adaptive Multivariate Smoothing	(Grill et al., 2017)
	Random cut forest	(Guha et al., 2016)
	Cluster center and nearest neighbor	(W.-C. Lin et al., 2015)
	One Class Support Vector Machine	(She et al.)
	Enhanced Support Vector Machine.	(Ramamoorthi et al., 2011)
	Expectation Maximization (EM) algorithm	(Lai et al., 2016)
	Hidden Semi-Markov Model	(Bang et al., 2017)
	Hierarchical Temporal Memory	(Lavin & Ahmad, 2015)
Earth Science	Expectation Maximization – Clustering	(Q. Liu et al., 2017)
Oil Production	Bayesian Gaussian Markov Random Fields	(Idé, Khandelwal, & Kalagnanam, 2016)

4.2. Anomaly detection with big data

The non-stop gathering of streaming of traffic data by the network hints the big data problems that fulfil the basic criteria of big data V's, which are volume, variety, and velocity (Suthaharan, 2014). A study has highlighted the network monitoring for security problems, associated with variety, veracity, volume, and velocity of big data 4 V's. A large amount of network data were processed efficiently in real time using big data analytics and also early detection of the various network attacks. In a traditional database, interconnections were through using data synchronization techniques which are not required in the big data analytics(Camacho, Macia-Fernandez, Diaz-Verdejo, & Garcia-Teodoro, 2014).Furthermore few other studies have addressed use cases related to big data and network security, which were malware beacon activity detection, a malware infected machine attached to an outside IP address to get instructions from the command and control centre. Nonetheless, such type connections are undetected in the huge legitimate traffic especially when the malware traffic happens over port 80, where HTTP web traffic occurs. Designing big data analytics for this scenario will help to identify what happens to the network, such as a number of bytes transferred over a certain period. Furthermore, it helps to compare between old and new pattern with months of accessed log file via graph-theoretic analysis(Swift Liu, 2015) (D. Lin, 2013).

Anomalies in data collected from a sensor network can identify which of the sensors are broken down or detecting events are fascinating for data analysts (Chandola et al., 2009). Besides, collected sensor network might contain different types of data, such as binary, discrete, continuous, audio, video, which falls under characteristics of big data. Therefore, combing big data analytics to identify the threat detection become crucial part in current environment. Table 4 shows the technologies used in different anomalous detection in big data technologies.

Table 4:
Anomaly detection and Big data Technologies

Detection	Hadoop	Storm	Spark	Flume	Kafka	Hbase	Others	References
Host misbehaviour	✓							(Gonçalves, Bota, & Correia, 2015)
Web server	✓							(Lee & Lee, 2011)
Threat Detection		✓		✓	✓	✓		(Lobato, Lopez, & Duarte, 2016)
Network & Hardware			✓		✓			(Rettig, Khayati, Cudré-Mauroux, & Piórkowski, 2015)
Power Consumption			✓				Hive	(X. Liu & Nielsen, 2016)
Hardware & resource usage		✓	✓		✓			(Solaimani et al., 2016)
Network	✓		✓					(Rathore, Ahmad, & Paul, 2016a)
	✓		✓					(Dromard, Roudière, & Owezarski, 2015)
			✓					(Casas, D'Alconzo, Zseby, & Mellia, 2016)
	✓	✓			✓			(S. Zhao, Chandrashekar, Lee, & Medhi, 2015)
Cloud log	✓							(Cui & He, 2016)
Malware in mobile			✓				MongoDB	(McNeil, Shetty, Guntu, & Barve, 2016)

4.3. Machine learning algorithms with big data

Machine learning is the data that powers the models, and the new era of big data is catapulting machine learning to the forefront of research and various industrial applications without leaving the network environment. The production of big data has forced us to rethink not just data processing frameworks, but implementations of machine learning algorithms as well.

Today, the problem of big data collections is often solved through distributed storage systems, which are designed to carefully control access and management in a fault-tolerant manner. One solution to the problem of big data objects in machine learning is through parallelization of algorithms. Data parallelism, where the data is divided into more manageable pieces, and each subset is computed simultaneously, or task parallelism, in which, the algorithm is divided into steps that can be performed concurrently. Furthermore when developing a big data processing engines various number of different tools can be used for machine learning algorithms for real-time analysis. Researchers have come up with few selection criteria for machine learning into big data processing, which are scalability, speed, coverage, usability, and extensibility of the algorithms that support well with big data processing(Landset et al., 2015b). The rapidly growing network data and traditional tools for machine learning are lacking as we move towards the distributed and real-time processing. Integrating the hybrid machine learning algorithms for the big data processing will help to increase the percentage of producing accuracy in results and also

reveal the hidden knowledge of the big data which are associations, sequences, classification, forecasting, anomalies, and clustering among the data. Table 5 shows several related studies that deal with real-time big data processing through the use of machine learning algorithms

Here, we have discussed some of the existing machine learning algorithms related to big data processing challenges. (Zhou, Pan, Wang, & Vasilakos, 2017) have presented a framework of machine learning on big data (MLBiD), which follows the phases of pre-processing, learning, and evaluation. Furthermore, it identifies the various challenges and opportunities in this domain for upcoming years. They have also proposed a taxonomy with supervised, unsupervised learning, reinforcement and data availability. Moreover, they have summarized several research issues, which includes new big data machine learning architecture that seamlessly support the real-time processing with massive volume of heterogeneous data.

(Fernández, Carmona, del Jesus, & Herrera, 2016) have addressed various issued related to the data distribution and parallelization of the present algorithms and with fuzzy representation. As well as different big data technologies challenges were discussed such Hadoop ecosystem (HDFS, HBASE, YARN,Map Reduce programming), Spark major concept resilient distributed datasets(RDD), FlinkML, including data pre-processing, supervised learning, and recommender systems.

Besides, (Suthaharan, 2014) has focused on various problems and challenges when combining big data and machine learning for network intrusion traffic. Due to time sensitive applications and prediction in network intrusion detection, it needs extremely capable big data technologies to tackle the recent problems. As well as some of the major problems, such as network topology, communication and security, associated with big data were addressed.

Table 5.Several related studies that deal with big data processing for anomaly detection through the use of machine learning.

Objectives	**Reference**
To develop and test a scalable machine learning architecture to make real-time predictions as service on big data	(Baldominos, Albacete, Saez, & Isasi, 2014)
To propose model which incorporate different soft computing techniques for large amounts of data which are characterized by spatial-temporal correlations	(Iqbal, Doctor, More, Mahmud, & Yousuf, 2016)
To design scalable architecture to process and analyse streaming data in real time fashion using machine learning algorithm for distributed environment	(Di Mauro & Di Sarno, 2014)
To investigate the existing machine learning algorithms for data distribution and parallelization problem.	(Fernández et al., 2016)
To develop a real-time remote health status prediction system using open source big data technologies and cloud	(Nair, Shetty, & Shetty, 2017)
To examine the opportunities from big data in retailing for predictive analytics with Bayesian techniques	(Bradlow, Gangwar, Kopalle, & Voleti, 2017)
To divide the big data analytics problem into six pillars and provide solutions for the problem.	(Khalifa et al., 2016)
To introduce a framework of machine learning on big data pre-processing, learning and evaluation	(Zhou et al., 2017)
To propose middleware for IoT devices which process data stream in real time.	(Nakamura, Suwa, Arakawa, Yamaguchi, & Yasumoto, 2016)

4.4. Machine learning algorithms based big data processing for anomalies detection

This sub-section critically analyses real time big data processing for anomalous detection through machine learning algorithms and their limitation.

(McNeil et al., 2016) have analysed the available tools to detect malware in the mobile devices. These tools failed to integrate group user profiling, which helps to automated behaviour driven dynamic analysis on targeted malware detection. Furthermore, they have proposed scalable real-time anomalies detection and notification of targeted malware in mobile devices (SCREDENT) architecture, to classify, detect, and predict targeted malware in real-time. Even so, evaluation of proposed architecture failed to give the promising result.

Moreover, (Lobato et al., 2016) have reviewed existing security approach, such as security information and event management(SIEM) build, to handle the data gathering and process in single point. Apart from that, it produces huge amount of false alarm. In addition, they have proposed an architecture to detect threat using steam processing and machine learning in real time. This architecture combines the benefit of real-time streaming through batch processing over a past available dataset and reduction of human involvement to the system. The proposed system also helps to detect the known and zero-day attack for attack classification and anomaly. However, the proposed system has been found to be weak on the accuracy of the dataset used for the experiment, in spite of openly available dataset, such as KDD dataset.

Meanwhile, (Gonçalves et al., 2015) have presented challenges in complex network infrastructure, which contains number of devices information stored in the vast logs file. Therefore, extracting meaningful information from that logs is demanding. The novel approach for assessing security logs of the various infrastructure devices to discover misbehaving hosts using machine learning and data mining techniques. The proposed approach has two phases. Firstly executing a set of steps for defining and configuring the detection mechanism, and secondly executing the detection mechanism in runtime. Nevertheless, the experimental setup is through on batch processing, and efficiency of the output is not accurate enough and high human intervention is also needed to automate some of the process.

Another study reveals that out of many other anomaly detection models, the machine learning has been most widely used, whereas growing number of network traffic become restriction for an existing system, since it needs to perform the complex calculation(Cui & He, 2016). In addition, model was proposed to handle better performance in detecting anomaly using Hadoop, HDFS, Mapreduce, cloud and machine learning algorithms. Further, weka interface was used in model to evaluate accuracy and efficiency with naive bayes, decision tree and support vector machine algorithm. In fact implementation of cloud infrastructure and real time input data streaming were not addressed well.

Besides, (Rettig et al., 2015) have addressed challenges in detecting anomalies in the streaming data, mainly focusing on generality and scalability. They have proposed new approach to evaluate an online anomaly detection with two metrics, using entropy and pearson correlation. Moreover, big data streaming components, such as kafka queues and spark streaming are used to assure the generality and scalability issues. Nonetheless, complex processes were limited to handle by the data and also long time duration for periodic batch processing.

Lastly, (X. Liu & Nielsen, 2016) have revealed existing anomaly detection models for smart grid were mostly based on offline and also consume huge amount of energy consumption. In fact, it should be on the other hand with real-time and minimal usage of energy. In addition, they proposed a method to detect anomaly, using in-memory distributed framework. The framework contains spark streaming and lambda system. Its major advantage is to support scalable live streaming for real-time detection. However, the framework took longer time duration to train the model. Consequently scheduling of real-time task was unknown.

All the above discussed approaches and their limitations demand reassessing the framework design to support the anomaly detection. Especially, an advance real time big data analytics for anomaly detection using machine learning will bring promising and better performance and accuracy for anomaly detection.

Table 6.Overview of big data processing technologies for anomaly detection using machine learning.

Domain	Finding	ML Techniques	Reference
	Extracting information from the security logs are trivial	Unsupervised – Clustering (Expectation- Maximization algorithm) and Supervised linear classification WEKA software used	(Gonçalves et al., 2015)
Telecommunication and Mobile network	Detecting anomalies in the streaming data	Pearson correlation pipeline	(Rettig et al., 2015)
	Existing framework do not integrate group user profiling which helps to perform targeted malware detection.	K-means clustering, Markov models	(McNeil et al., 2016)
Network traffic	Huge amount of traffic data delay the response time to detect the threats – Scalability and Accuracy - Memory consumption and searching complexity – boosting accuracy	Principal Component Analysis	(Lee & Lee, 2011)
	Current security approaches are single point, and generate huge amount of false alarm.	Decision trees algorithms, Artificial Neural Network, Support Vector Machine	(Lobato et al., 2016)
Smart grid	Detection models are based on offline. Huge amount of training data are required.	Periodic autoregression with an exogenous variable.	(X. Liu & Nielsen, 2016)
Cloud environment	Increasing network traffic data is a bottleneck for an existing system, which has to perform the complex calculation.	Naïve Bayes, Decision Tree and Support Vector Machine (Weka Interface)	(Cui & He, 2016; Shirdastian, Laroche, & Richard, 2017)

Table 7 describes the number of commercial platforms, which have integrated machine learning, big data technologies into their anomalous detection.

Table 7.
Summary of commercial platform and solution for big data streaming analytics.

Product name	Description	Architecture Components	Machine Learning	Source
Anodot	Anodot platform automatically selects the appropriate algorithm to exhibit the data pattern from available options and adjust it over time, based on real-time, mainly for anomaly detection	Hadoop, Spark, Hive, anomaly detection engine.	Yes	(Dror, 2017)
Numenta	Numenta platform can work with both, predictable and highly unpredictable platform. The algorithm works with continuously learning algorithm, so that data are automatically handled without human intervention.	No access to architecture component of the system.	Yes	(Lavin & Ahmad, 2015)
Microsoft azure stream analytics	Azure stream analytics is real-time analytic computations on streaming data to provide multiple solutions. It combines azure streaming analytics and apache storm on azure HDinsight, using PaaS solution.	Kafka, RabbitMQ, ActiveMQ Apache storm, azure stream analytics Hbase, HDFS.	Yes	(Branscombe, 2015)
WSO2 analytics	WSo2 analytics platform an one stop centre, which is capable to collect and analyse various IoT sensor data, which do real-time and batch threat analysis using a machine learning algorithm	Event receivers, Siddhi event, Apache spark, Caasandra, Hbase	No	("Introducing WSO2 Data Analytics Server," 2015)
Striim	Striim is end-to-end in-memory streaming platform used for infrastructure critical application.	Data lake, Kafka, NoSQL, Hadoop, Hbase	Yes	(Wilkes, 2016)
Tibco streamBase	TIBCO streamBase is the event processing platform, which develop, host, execute, and integrate the predictive analytics in big time real time.	Hadoop, spark. Kafka, Flume, casssandra,Hbase.	No	("TIBCO StreamBase and the TIBCO Accelerator for Apache Spark," 2017)

Industries might face more challenges with emerging new set of technologies, such as cloud to the edge, data from IOT, smart devices, intelligent things, block chains, connected home, virtual reality, 5G, quantum computing and serverless paas. However, sophisticated advance anomalous detection techniques using machine learning and big data should be adequate to handle those challenges.

5. Taxonomy of Real-time big data processing for anomaly detection.

This section highlights and proposes a taxonomy for the anomaly detection, big data, and machine learning. A taxonomy of real-time big data processing for anomaly detection is classified into different categories, which are, techniques, application, anomalies, modes, data, big data processing, and the record categories. Figure 4 shows the classified taxonomy of real-time big data processing for anomaly detection, based on the set of parameters found in majority of the literature review.

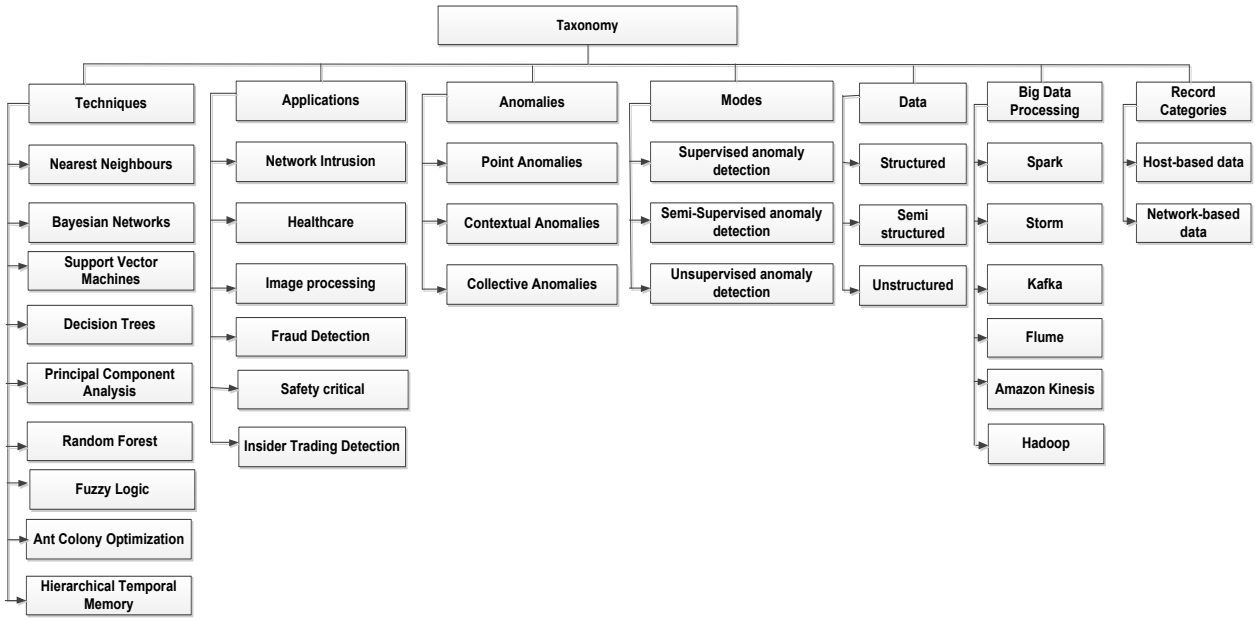

Fig. 4. The process of real time big data processing technologies for anomaly detection

5.1. Techniques

Among a vast group of anomaly detection techniques accessible, we have examined six major techniques as follows: Nearest Neighbours (Wauters & Vanhoucke, 2017), Bayesian network (Mascaro, Nicholso, & Korb, 2014), support vector machine (Aburomman & Reaz, 2017), decision trees (Muniyandi, Rajeswari, & Rajaram, 2012), Random forest (Farnaaz & Jabbar, 2016), Fuzzy logic (Hamamoto, Carvalho, Sampaio, Abrão, & Proença, 2017) principal component analysis (Hamamoto et al., 2017).

5.1.1. Nearest Neighbours (NN) depends on the utilization of distance measures. NN methods embraces the entire sampling set and incorporates the information in the set, and as well as the coveted grouping for respective item. The distance between each item in the sampling set must be processed for the purpose of classifying the points, wherein, the k closet passages in the sampling set is considered as the point in far distance. However, the shortcoming of the NN is the similarity measure, which leads to misclassification of points because of its inefficiency in accurately calculating distance between them, particularly while classifying small subset of the features (Su, 2011). Historical data is employed to identify nearest neighbours at a given data point. The applications of the Nearest Neighbour method usually revolves around the similar domain of artificial intelligence methods, which are classification and prediction. The initial k-Nearest Neighbour has been deployed as a predictor and allowed to benchmark its accuracy and scalability. Nonetheless, algorithm have outflows in concerning memory requirements and computation complexity in terms of anomaly detection (Wauters & Vanhoucke, 2017).

5.1.2. Bayesian Networks (BN) have been broadly utilized for grouping issues. BN grouped into qualitative and quantities model. The qualitative piece of the system is spoken to by a coordinated non-cyclic graph, whose nodes denote to the random factors in the problem domain and whose edges systematize significance relations between the factors they interface. The quantitative piece of the model comprised of an arrangement of probability distributions on each node (Cozar, Puerta, & Gamez, 2017). Bayesian Networks can be used for detecting anomalies within vessel tracks. Dynamic and static network are produced by using Bayesian Networks learner on an Automated Identification System data, which has been supplemented with real-world data, which proved distinct and complementary strengths in identifying anomalies. (Mascaro et al., 2014)

5.1.3. Support Vector Machine (SVM) has been used to construct a decision boundary, which has the most extreme edge between the typical data set and the source (S. Yin, Zhu, & Jing, 2014). SVM is a classifier in light of finding a separating hyperplane in the feature space between two classes is such a path, to the point that, the distance between the hyperplane and the open data points of each class is increased. The approach depends on minimized classification threat as opposed to on optimal classification. Especially helpful when the number of features, m, is high and the number of data points, n, is low (m>>n) (Buczak & Guven, 2016). Moreover, SVM is a robust classifier, used for many classification tasks, such as network intrusion detection. The main focus is extending two-class SVM classifiers to multi-class classifiers. This method is the most effective for classifying the samples from the NSL-KDD dataset, such a classifier is crucial in intrusion detection system (Aburomman & Reaz, 2017).

5.1.4. Decision tree is a tree-like structure that has leaves, which denote to groupings and divisions, which thusly state to the conjunctions of highlights that prompt those classification. The outstanding known techniques for naturally fabricating decision tress are the ID3 and C4.5 algorithms. The two algorithms formulate decision tress from an arrangement of training data utilizing the idea of data entropy (Buczak & Guven, 2016). In addition, a decision tree refines its decision boundaries on each cluster by acquiring knowledge from the subgroups within the cluster. The results from the decision trees in each cluster are exploited and a final conclusion is made decision trees (Muniyandi et al., 2012). Clustering helps to group the normal and abnormal data points in anomaly detection.

5.1.5. Random Forest (RF) is a collaborative classifier that is used to improve the accuracy. It consists of two stages feature selection and classification. (Farnaaz & Jabbar, 2016) . Random forest is to generate multiple decision trees from random subsets of data. Each of them capturing different regularities since random subset of the instances are in the interest (Promod & Jacob, 2016).One of the major advantages of the Random forest is that it yields low classification errors when compared with other traditional classification algorithms. The Random forest can be used to detect four types of attacks like DOS, probe, U2R and R2L (Farnaaz & Jabbar, 2016). Likewise, RF has been used in botnet detection problem, which due to its high accuracy in prediction, ability to handle diverse bots, based on data characterized by a very large number and diverse types of descriptors (Singh, Guntuku, Thakur, & Hota, 2014). Yet, when working with large dataset and complex estimation procedures, the RF consumes lot of computational time (Genuer, Poggi, Tuleau-Malot, & Villa-Vialaneix, 2017)

5.1.6. Fuzzy Logic algorithm is capable of making rational decisions in an environment that is imprecise, uncertain, and incomplete. It uses time interval to detect an anomaly in the network. Moreover, exponentially weighted moving average techniques are applied to calculate threshold values, which represent more recent higher weight (Hamamoto, Carvalho, Sampaio, Abrão, & Proença Jr, 2018).

5.1.7. Principal Component Analysis creates a digital signature for traffic characterization of a network segment. It is a statistical procedure used to decrease those dimensional multivariate issues by examining those difference for every variable among all response measurement. Furthermore, the response information can be symbolised by a reduced fixed set of dimensions without much loss of information (Hamamoto et al., 2017). Besides, It analyses noteworthy information data from logs to finds the important activity time intervals among the data set, and afterward diminish them, so this new set can professionally characterize to the consistent conduct of a network segment (Fernandes, Rodrigues, & Proença, 2015).

5.1.8. Ant Colony Optimization is inspired by the ability of ants to find the shortest path between their colonies to the food sources. Ant Colony Optimization is where a population of agents competing and globally asynchronous, cooperate with one another to find an optimal solution. Ant Colonization Optimization and Dynamic Time Wrapping methods have been used in the environment of pattern recognition and anomaly detection (Fernandes, Carvalho, Rodrigues, & Proença, 2016).

5.1.9. Hierarchical Temporal Memory (HTM) is capable of predicting the flow of data in real-time, based on the condition of the precedence in learning. Moreover, HTM has been widely employed for detecting anomalous in distributed real-time systems, mainly due its capability in yielding highly accurate detection. This techniques comprise of a solid framework which helps in better estimation, categorisation, and generating continuous time based data sequence, online. The HTM is also used to detect anomalous sequences on the vehicular controller area network bus, which sends alarm signal during abnormal conditions (C. Wang et al., 2018). Furthermore, HTM is employed to robustly detect anomalies on a variety of data streams, which mainly deals with noisy data and to minimize false positives (S. Ahmad et al., 2017).

5.2. Applications

In this sub-section, we have listed major application scenarios in the area of real-time big data processing for anomalous detection. The growth of IOT contributes for different number of applications with sensors that produce important data that changes over time (E. Ahmed et al., 2017). In fact, detecting anomaly in that data can help to overcome many challenges for organizations.
Interestingly, some of the applications like network intrusion, healthcare, image processing, fraud detection, safety critical applications, and insider trading detection application challenges and limitation were discussed in section 2.0.

5.3. Anomalies

This sub-section briefly overviews three different categories of anomaly such as point, contextual and collective anomalies for real time anomalous detection.

5.3.1. Point anomalies happen when data points are observed upside or downside of normal points in the available dataset (Hayes & Capretz, 2015). For example, User-to-root (U2R) and remote-to-local (R2L) attacks are best known for point anomaly. In fact data point helps to point the unauthorized access from a remote and local access privileges (M. Ahmed et al., 2016).

5.3.2. Contextual anomalies discover association in datasets and detect differences in their external behaviour characteristics, and will label anomalous effects in the data. For example power consumption of an office building was found to be far greater during midday, and during a work day, as against night, and during weekends (Hayes & Capretz, 2015).

5.3.3. Collective anomalies is a group of related data instances, which act differently corresponding to the whole dataset, In fact, these set of data instances is termed a collected anomaly. For instance, DoS attack

best suits the collective anomaly, these attack produces various connection request to web server, out of that, only a single request is reliable (M. Ahmed et al., 2016).

5.4. Anomaly Detection Modes

This section discusses different types of modes that are commonly used in anomaly detection.

5.4.1. Supervised anomaly detection techniques detect anomaly based on generating a set of grouping rules that aid in predicting future data. An example of supervised anomaly detection is classification-based anomaly detection (Kakavand et al., 2015).

5.4.2. Semi-supervised anomaly detection is an approach that models only the normal records. The other records are labelled as outliers in the testing phase (Kakavand et al., 2015).

5.4.3. Unsupervised anomaly detection focuses on the data that does not contain any labelling information and it does not need a separate training and testing phase. A general example is clustering based anomaly detection (Kakavand et al., 2015).

5.5. Data

This section details the different types of data used in anomaly detection.

5.5.1. Structured data is tabular data found in spreadsheets or relational databases (Gandomi & Haider, 2015). Structured data can be processed more efficiently for anomaly detection compared to semi structured or unstructured data.

5.5.2. Semi structured data fall between the structured and unstructured data groups. These data do not conform to strict standards. An example of semi structured data is XML (Gandomi & Haider, 2015). Compared to structured data, the semi structured data lacks proper formatting and is time consuming for anomaly detection.

5.5.3. Unstructured data does not follow any strict format or sequences. Examples of unstructured data are social media contents, images, audio, IoT sensor data, and video (Gandomi & Haider, 2015). The processing of unstructured data for anomaly detection is time and memory consuming, and require lot of resources.

5.6. Big Data processing

The state-of-the-art technologies utilize different types of big data tools in a variety of domains, this section discuss those big data tools used for anomaly detection.

5.6.1. Spark is an open source big data tool used for data processing (Solaimani, Iftekhar, Khan, Thuraisingham, & Ingram, 2014). Spark receives data from kafka and processes it in real-time, using machine learning algorithms for anomaly detection.

5.6.2. Storm is a data application programming framework that is used to write applications for rapidly processing large amounts of data (Ranjan, 2014). Storm is a tool similar to spark that processes data in real time, they both have their own pros and cons.

5.6.3. Kafka is mainly utilized for constructing data pipelines in real-time. It is also useful for developing online streaming applications. This technology can be horizontally gauged, fool proof, and rapidly swift, hence it is mostly popularly used in number of sectors. (Assunção et al., 2015).

5.6.4. Flume is a distributed service used for real-time data collection, temporary storage, and delivery of data to a target (Birjali, Beni-Hssane, & Erritali, 2017).

5.6.5. Amazon Kinesis is a distributed message queuing framework (Ranjan, 2014). It is capable of handling large data size and large pipelines, furthermore, the output generated by Kinesis can be applied to machine learning algorithms.

5.6.6. Hadoop is one of the most popular big data technology frameworks, which helps to solve this scalability problem by the Hadoop distributed file system (HDFS). It has been used to store large amounts of data across multiple nodes of commercial hardware (Landset et al., 2015a).

5.7. Record categories

In this section, we have elaborated different types of record categories that have been used in anomaly detection.

5.7.1. Host-based data contains incoming and outgoing network traffic of an individual device on a network (Sahasrabuddhe, Naikade, Ramaswamy, Sadliwala, & Futane, 2017). The host based data can be processed for anomaly detection using different machine learning algorithms in a single host (L. Wang & Jones, 2017).

5.7.2. Network-based data contains network traffic to and from all devices in a network (Sahasrabuddhe et al., 2017). It can be processed for anomaly detection using different machine learning algorithms all the devices connected network. The network based data facilities detection of anomalies based on the pattern generated from the sensor data (L. Wang & Jones, 2017).

The above sections had thoroughly discussed, the various existing anomaly detection techniques, applications, categories and modes of anomaly detection, types of data, big data processing technologies, and record categories.

6. Research challenges

This section emphasizes the most important research challenges in real-time big data processing technologies for anomaly detection. State-of-the-art techniques in anomalous detection, real-time big data processing, and machine learning have been surveyed to identify the research challenges, recommendation and future research directions.

6.1. Redundancy: Dealing with large amount of data generated from various network sensors in real-time is a critical factor in big data management, particularly due to the constant repetition of previously generated data. Even though, the existing big data processing technologies, such as Hadoop and spark frameworks have been developed for handling data replication across multiple clusters, still these technologies are inadequate in addressing the challenges related to data redundancy, data quality, inconsistency and cost of maintaining storage (Bhadani & Jothimani, 2016). Moreover, these technologies lack schema to minimize redundancy and are not capable enough to store huge amount of data (Hashem et al., 2015). Hence, it becomes essential to design a framework that is capable of addressing and minimising redundancy issues aimed at catering present and future needs.

6.2. Computational cost: a number of studies have focused on merging or incorporating several techniques to increase the performance of anomaly detection, which leads to increase in computation cost (W.-C. Lin et al., 2015). Moreover, high dimensionality combined with large sample size creates issues, such as heavy computational cost and algorithmic instability (Fan, Han, & Liu, 2014). Therefore, using big data technologies along with cloud will address the computational cost issue, by incorporating parallel and distributed processing, which helps to build multiple clusters leading to minimization of the computation cost. The mass production of high chips and processors has reducing their costs, hence utilization of these hardware will increase the power of systems that helps to process huge volume of data in real-time, resulting in reducing in computational cost.

6.3. Nature of Input data: In an aspect of any model built, the first thing to look into is the nature of input data. Input data is the collection of data instances, such as object, record, point, vector, pattern, event, case, sample, observation, entity. They are various set of attributes for each data instances, such as variable, characteristics, features, field, and dimension. It has two different types of attributes such as binary, categorical or continuous. Each data instance mostly falls under the category of either, univariate or multivariate. The diverse nature of input data makes the anomaly detection techniques struggle in selecting appropriate algorithm to handle that specific data. Basically, anomaly detection techniques will vary based on nature of attributes in that application (Chandola et al., 2009). This issues will be addressed by developing hybrid unsupervised machine learning algorithm.

6.4. Noise and missing value: The streaming data in network sensor consist of different types of data, such as binary, discrete, continuous, audio, video, and image. These data collected from various deployed sensors via a communication channel includes noise and missing values, due to the incoming speed of data (Chandola et al., 2009). Noise and missing values can produce high probabilities for rising the false positive alarm in anomalous detection. Huge quantity of unrelated features produce noise in the input data, which bypass the true anomalies (Erfani et al., 2016). These issue will be addressed by incorporating auto noise cleansing module in the detection framework. The auto cleansing module will also address the missing value issue by adding NA to datasets.

6.5. Parameters Selection: Finding the appreciate parameters for any machine learning algorithms can be challenging(Mirsky et al., 2017). Especially when dealing with real-time anomaly detection, it is essential to consider single, multiple and hyper parameters before choosing them. In addition, a set of parameters that works well at the early stages of the evolution process may not perform well at the later stages and vice versa (Sarker, Elsayed, & Ray, 2014). Parameters were one of the major contributors that decide the performance of the algorithms. Further, it can give huge impact or delay to training the model. Alternatively, we can work on the parameter-free algorithm in identifying the node partitions in streaming, directed, bipartite graphs, and monitor their evolution over time to detect events (Akoglu et al., 2015). Employing the likes of eccentricity techniques will address this challenge because it will minimise parameter selection.

6.6. Inadequate Architecture: The existing architecture are capable of handling anomaly detection in batch processing and less volume of data, however, they are incapable of handling big data in real-time. Organizations are working to produce the big data architecture to perform better, but when it comes to real-time data it is fundamentally a different architecture than big data. Components of the real-time architecture have to merge application and analytics to propose the new way of working environment that achieves the needs of both, data in motion (fast) and data at rest (big). Big data architecture is inefficient when it is not being integrated with existing enterprise data; the same way an analysis cannot be completed until big data correlates it (Katal, Wazid, & Goudar, 2013). Incorporating various big data technologies with hybrid machine learning algorithms will address the architectural issues.

6.7. Data visualizations: Processed and analysed data or report needs to be visualized by the user as well as must provide insight from the report. Nevertheless, challenge lies in selecting appropriate visualization techniques, for the anomalies detection from the various connected devices. Multiple visualization techniques are used in the design of anomalous detection visualization from simple graph to 2D, and 3D views. Heat maps, scatter plots, parallel coordinates, and node-link graphs are easy to showcase the output when it comes to 2D and 3D. The 3D interaction requires users to understand the data completely to rotate a zoom the displays (Shiravi, Shiravi, & Ghorbani, 2012). Embedding the available open source visualization techniques in framework can address this problem, furthermore, the framework enable the system to automatically select the appropriate visualization technique.

6.8. Heterogeneity of data: Unstructured data represents almost every kind of data being produced, like social media interactions, to recorded meetings, to the handling of PDF documents, fax transfers, to emails, and more. Structured data is always organized into highly mechanized and manageable way. It shows good integration with the database, but unstructured data is completely raw and unorganized. Working with unstructured data is cumbersome and of course costly too. Converting all this unstructured data into structured one is also not feasible. The employment of unsupervised hybrid machine learning algorithms will address the heterogeneous data issue. The incorporation of hybrid machine learning algorithms and real-time big data

technologies will help to cluster the incoming data into different categories, which eventually will helps easily identify data types, whereby addressing heterogeneity issue.

6.9. Accuracy: Even though the existing technologies are capable of detecting anomalies, still the dependency of the outcome is unreliable due to the accuracy issues. In some cases, better accuracy is produced at the cost of high computational processing and time. This issue will be addressed by incorporating real time big data technologies with hybrid machine learning algorithms, which emerge as an alternative powerful meta-learning tool to accurately analyse the massive volume of data generated by modern applications, with less memory and power consumptions.

Table 8: Summary of recommendation for research challenges and future research directions.

Challenges	Recommendations	Future Research Directions	Reference
Redundancy	Employing filtering module into framework will help to reduce the redundancy data.	Data Deduplication. Dimension Reduction. Network Theory.	(ur Rehman et al., 2016)
Computational cost	Including modern technologies like virtualization, cloud, edge and fog computing might help to reduce computation cost. Moreover, big data processing can use multiple clusters assisted by parallel and distributed cloud architecture.	Resources on-demand with costs proportional to the actual usage. Cloud-supported analytics.	(Assunção et al., 2015)
Nature of Input data	Developing hybrid unsupervised machine learning algorithms	Hybrid unsupervised machine learning. Deep learning.	(Weston, 2015)
Noise and missing value	Incorporating auto noise cleaning module framework will address noise and missing value problem.	Eefficient, robust, scalable, and optimized pre-processing techniques for both, historical and streaming big data.	(Assunção et al., 2015)
Parameters Selection	Eccentricity techniques to minimise parameter selection.	Developing dynamic learning algorithm provides better efficiency in parameters selection.	(Mukherjee, Shu, & Wang, 2018)
Inadequate Architecture	Integrating various modern technologies like big data, cloud, fog and edge computing technologies with hybrid or enhanced machine learning algorithms will address these issues.	AWS, open source cloud, Microsoft, IBM. In-memory architecture will be more capable for real-time analytics. Spark, H2O.	(Qiu et al., 2016) (Landset et al., 2015a)
Data visualizations	Implanting the available open source visualization techniques in framework and also spontaneous selection for appropriate visualization technique.	GraphX, HadoopR, Python, Lightning data visualization server.	(Qiu et al., 2016) (Landset et al., 2015a)
Heterogeneity of data	Combing hybrid or enhanced machine learning algorithms and real-time big data technologies will help to cluster the incoming data into different categories such as data types, size and others.	Data cleansing and data curation. The incorporation of hybrid machine learning algorithms and real-time big data technologies.	(Anagnostopoulos, Zeadally, & Exposito, 2016)

Accuracy	Embed the evaluation techniques and meta-learning tool to hybrid machine learning algorithms for accurate analyse.	The reduction of memory and power consumptions when processing large amount of data.	(Qiu et al., 2016) (Landset et al., 2015a)

In addition to the above summarized future research directions on research challenges and recommendations in table 8, we have identified forthcoming research directions for research communities to develop an adoptable and responsive model for real-time big data processing, which can help to collect data with labelling and converting unstructured data into semi-structure data which will be easier to label in run-time for processing. Likewise, model should support for flexible select specific feature and extract parameters for analysis. These selected parameters can be used for benchmarking various types of threat and real-time processing. Similarly, the proposed model should be more competent and timely to train as well as retrain the model more efficiently. Many of the existing works lack in retraining the model for processing, which will be remarkably beneficial in real-time. Furthermore, retraining the model should contain modules for offline and online analysis.

Similarly, the future model should comprise fast, hybrid and incremental learning algorithms for modern incoming real-time data, which can facilitate selection of right time windowing for online analysis. Besides, multiple level visualization techniques improve understanding of processed and analysed data. Furthermore, these techniques can support security analytics, incorporating recent visualization technologies such as, 3D, 4D and augmented and virtual reality for visualizing complex processed data. Lastly, constructing new dataset comprising present structure and unstructured data from various recent technologies like IoT, 3D printing, smart cities, and other connected devices. Developed datasets should be validated with openly available existing datasets to handle multiple distributed threat all around the world.

7. Conclusion and future remarks

In this paper, we had investigated real time big data processing and machine learning with the possibility of anomalous detection. We had examined recent works in real time and anomalous detection from a use cases perspective. Examination of these use case had helped us to identify the challenges associate with anomaly detection in real-time. We had elaborated the shortcomings and challenges of the existing approaches in detecting the anomalous threat in the specified domain. Moreover, we had also developed a taxonomy of our approach. Furthermore, we had identified that state-of-the-art approaches that pose research challenges in detecting anomaly in real-time. This study had presented a layer diagram that helps to comprehend the anomaly detection process, will might lead to the proposal of a framework for real-time big data processing to detect anomaly in the future. The framework to be proposed shall be implemented in real-time analytics using IOT data sources, gateway, network infrastructure, streaming, clustering/classification algorithms, big data processing technologies, analytics for anomaly detection, and visualization. Furthermore, the study will also evaluate the performance of the proposed framework, in terms of accuracy and efficiency in the detection rate with algorithm as against the existing technologies.

8. References

Aburomman, A. A., & Reaz, M. B. I. (2017). A Novel Weighted Support Vector Machines Multiclass Classifier Based on Differential Evolution for Intrusion Detection Systems. *Information Sciences*.

Ahmad, A., Paul, A., & Rathore, M. M. (2016). An efficient divide-and-conquer approach for big data analytics in machine-to-machine communication. *Neurocomputing, 174*, 439-453.

Ahmad, S., Lavin, A., Purdy, S., & Agha, Z. (2017). Unsupervised real-time anomaly detection for streaming data. *Neurocomputing, 262*, 134-147.

Ahmed, E., Yaqoob, I., Hashem, I. A. T., Khan, I., Ahmed, A. I. A., Imran, M., & Vasilakos, A. V. (2017). The role of big data analytics in Internet of Things. *Computer Networks, 129*, 459-471.

Ahmed, M., Mahmood, A. N., & Hu, J. (2016). A survey of network anomaly detection techniques. *Journal of Network and Computer Applications, 60*, 19-31.

Ahmed, T., Oreshkin, B., & Coates, M. (2007). *Machine learning approaches to network anomaly detection.* Paper presented at the Proceedings of the 2nd USENIX workshop on Tackling computer systems problems with machine learning techniques.

Akoglu, L., Tong, H., & Koutra, D. (2015). Graph based anomaly detection and description: a survey. *Data Mining and Knowledge Discovery, 29*(3), 626-688.

Al-Jarrah, O. Y., Yoo, P. D., Muhaidat, S., Karagiannidis, G. K., & Taha, K. (2015). Efficient machine learning for big data: A review. *Big Data Research, 2*(3), 87-93.

Ali, A., Hamouda, W., & Uysal, M. (2015). Next generation M2M cellular networks: challenges and practical considerations. *IEEE Communications Magazine, 53*(9), 18-24.

Ali, O., Shrestha, A., Soar, J., & Wamba, S. F. (2018). Cloud computing-enabled healthcare opportunities, issues, and applications: A systematic review. *International Journal of Information Management, 43*, 146-158.

Almadani, B., Saeed, B., & Alroubaiy, A. (2016). Healthcare systems integration using Real Time Publish Subscribe (RTPS) middleware. *Computers & Electrical Engineering, 50*, 67-78.

Almeida, V. A., Doneda, D., & de Souza Abreu, J. (2017). Cyberwarfare and Digital Governance. *IEEE Internet Computing, 21*(2), 68-71.

Anagnostopoulos, I., Zeadally, S., & Exposito, E. (2016). Handling big data: research challenges and future directions. *The Journal of Supercomputing, 72*(4), 1494-1516.

Assunção, M. D., Calheiros, R. N., Bianchi, S., Netto, M. A., & Buyya, R. (2015). Big Data computing and clouds: Trends and future directions. *Journal of Parallel and Distributed Computing, 79*, 3-15.

Baldominos, A., Albacete, E., Saez, Y., & Isasi, P. (2014). *A scalable machine learning online service for big data real-time analysis.* Paper presented at the Computational Intelligence in Big Data (CIBD), 2014 IEEE Symposium on.

Bang, J.-h., Cho, Y.-J., & Kang, K. (2017). Anomaly detection of network-initiated LTE signaling traffic in wireless sensor and actuator networks based on a Hidden semi-Markov Model. *Computers & Security, 65*, 108-120.

Bhadani, A. K., & Jothimani, D. (2016). Big Data: Challenges, Opportunities, and Realities. *Effective Big Data Management and Opportunities for Implementation, IGI Global, Pennsylvania, USA*, 1-24.

Birjali, M., Beni-Hssane, A., & Erritali, M. (2017). Analyzing Social Media through Big Data using InfoSphere BigInsights and Apache Flume. *Procedia computer science, 113*, 280-285.

Bradlow, E. T., Gangwar, M., Kopalle, P., & Voleti, S. (2017). The role of big data and predictive analytics in retailing. *Journal of Retailing, 93*(1), 79-95.

Branscombe, M. (2015). How Microsoft's fast track Azure will help businesses conquer IoT. Retrieved from http://www.techradar.com/news/internet/cloud-services/how-microsoft-s-fast-track-azure-will-help-businesses-conquer-iot-1291025

Buczak, A. L., & Guven, E. (2016). A survey of data mining and machine learning methods for cyber security intrusion detection. *IEEE Communications Surveys & Tutorials, 18*(2), 1153-1176.

Camacho, J., Macia-Fernandez, G., Diaz-Verdejo, J., & Garcia-Teodoro, P. (2014). *Tackling the big data 4 vs for anomaly detection.* Paper presented at the Computer Communications Workshops (INFOCOM WKSHPS), 2014 IEEE Conference on.

Carvalho, J. V., Rocha, Á., Vasconcelos, J., & Abreu, A. (2018). A health data analytics maturity model for hospitals information systems. *International Journal of Information Management*.

Casas, P., D'Alconzo, A., Zseby, T., & Mellia, M. (2016). *Big-DAMA: Big Data Analytics for Network Traffic Monitoring and Analysis.* Paper presented at the LANCOMM@ SIGCOMM.

Chandola, V., Banerjee, A., & Kumar, V. (2009). Anomaly detection: A survey. *ACM computing surveys (CSUR), 41*(3), 15.

Chen, M., Mao, S., & Liu, Y. (2014). Big data: a survey. *Mobile Networks and Applications, 19*(2), 171-209.

Chow, A. (2018). SingHealth Data Breach – An Analytical Perspective. *SpiderLabs.* Retrieved from https://www.trustwave.com/Resources/SpiderLabs-Blog/SingHealth-Data-Breach-%E2%80%93-An-Analytical-Perspective/#disqus_thread

Cisco VNI Forecast and Methodology, 2015-2020. (2016). Retrieved from http://www.cisco.com/c/en/us/solutions/collateral/service-provider/visual-networking-index-vni/complete-white-paper-c11-481360.html

Cloud Strategy Partners, L. (2015). Cloud Based Solutions for Big Data (Course Note). (ISBN: 978-1-4799-8513-5). Retrieved 09/09/2015, from IEEE Educational Activities and IEEE Cloud Computing http://ieeexplore.ieee.org/courses/details/EDP405?tag=1

Cozar, J., Puerta, J. M., & Gamez, J. A. (2017). An Application of Dynamic Bayesian Networks to Condition Monitoring and Fault Prediction in a Sensored System: a Case Study. *International Journal of Computational Intelligence Systems, 10*(1), 176-195.

Cui, B., & He, S. (2016). *Anomaly detection model based on Hadoop platform and Weka interface.* Paper presented at the Innovative Mobile and Internet Services in Ubiquitous Computing (IMIS), 2016 10th International Conference on.

DBIR. (2015). *2015 Data breach investigations report*. Retrieved from http://www.verizonenterprise.com/DBIR/2015/

Di Mauro, M., & Di Sarno, C. (2014). *A framework for Internet data real-time processing: A machine-learning approach.* Paper presented at the Security Technology (ICCST), 2014 International Carnahan Conference on.

Dromard, J., Roudière, G., & Owezarski, P. (2015). *Unsupervised network anomaly detection in real-time on big data.* Paper presented at the East European Conference on Advances in Databases and Information Systems.

Dror, Y. (2017). Practical Elasticsearch Anomaly Detection Made Powerful with Anodot. Retrieved from https://www.anodot.com/blog/practical-elasticsearch-anomaly-detection-made-powerful-with-anodot/

Ellis, B. (2014). *Real-time analytics: Techniques to analyze and visualize streaming data*: John Wiley & Sons.

Erfani, S. M., Rajasegarar, S., Karunasekera, S., & Leckie, C. (2016). High-dimensional and large-scale anomaly detection using a linear one-class SVM with deep learning. *Pattern Recognition, 58*, 121-134.

Fan, J., Han, F., & Liu, H. (2014). Challenges of big data analysis. *National science review, 1*(2), 293-314.

Farnaaz, N., & Jabbar, M. (2016). Random Forest Modeling for Network Intrusion Detection System. *Procedia Computer Science, 89*, 213-217.

Feng, W., Zhang, Q., Hu, G., & Huang, J. X. (2014). Mining network data for intrusion detection through combining SVMs with ant colony networks. *Future Generation Computer Systems, 37*, 127-140.

Fernandes, G., Carvalho, L. F., Rodrigues, J. J., & Proença, M. L. (2016). Network anomaly detection using IP flows with principal component analysis and ant colony optimization. *Journal of Network and Computer Applications, 64*, 1-11.

Fernandes, G., Rodrigues, J. J., & Proença, M. L. (2015). Autonomous profile-based anomaly detection system using principal component analysis and flow analysis. *Applied Soft Computing, 34*, 513-525.

Fernández, A., Carmona, C. J., del Jesus, M. J., & Herrera, F. (2016). A view on fuzzy systems for big data: progress and opportunities. *International Journal of Computational Intelligence Systems, 9*(sup1), 69-80.

Gandomi, A., & Haider, M. (2015). Beyond the hype: Big data concepts, methods, and analytics. *International Journal of Information Management, 35*(2), 137-144.

Gani, A., Siddiqa, A., Shamshirband, S., & Hanum, F. (2016). A survey on indexing techniques for big data: taxonomy and performance evaluation. *Knowledge and Information Systems, 46*(2), 241-284.

García, L., Tomás, J., Parra, L., & Lloret, J. (2018). An m-health application for cerebral stroke detection and monitoring using cloud services. *International Journal of Information Management*.

Genuer, R., Poggi, J.-M., Tuleau-Malot, C., & Villa-Vialaneix, N. (2017). Random forests for big data. *Big Data Research, 9*, 28-46.

Ger, M. (2017). Autonomous Vehicles Silicon Valley 2017 – The Future Of Transportation Will Drive Huge Growth In Data. Retrieved from https://hortonworks.com/blog/autonomous-vehicle-silicon-valley-2017-findings-future-transportation-will-drive-huge-growth-data/

Gilmore, C., & Haydaman, J. (2016). *Anomaly Detection and Machine Learning Methods for Network Intrusion Detection: an Industrially Focused Literature Review.* Paper presented at the Proceedings of the International Conference on Security and Management (SAM).

Gonçalves, D., Bota, J., & Correia, M. (2015). *Big Data Analytics for Detecting Host Misbehavior in Large Logs.* Paper presented at the Trustcom/BigDataSE/ISPA, 2015 IEEE.

Grill, M., Pevný, T., & Rehak, M. (2017). Reducing false positives of network anomaly detection by local adaptive multivariate smoothing. *Journal of Computer and System Sciences, 83*(1), 43-57.

Guha, S., Mishra, N., Roy, G., & Schrijvers, O. (2016). *Robust random cut forest based anomaly detection on streams.* Paper presented at the International Conference on Machine Learning.

Gupta, S., Kar, A. K., Baabdullah, A., & Al-Khowaiter, W. A. (2018). Big data with cognitive computing: A review for the future. *International Journal of Information Management, 42*, 78-89.

Hamamoto, A. H., Carvalho, L. F., Sampaio, L. D. H., Abrão, T., & Proença Jr, M. L. (2018). Network Anomaly Detection System using Genetic Algorithm and Fuzzy Logic. *Expert Systems with Applications, 92*, 390-402.

Hamamoto, A. H., Carvalho, L. F., Sampaio, L. D. H., Abrão, T., & Proença, M. L. (2017). Network Anomaly Detection System using Genetic Algorithm and Fuzzy Logic. *Expert Systems with Applications*.

Hashem, I. A. T., Chang, V., Anuar, N. B., Adewole, K., Yaqoob, I., Gani, A., . . . Chiroma, H. (2016). The role of big data in smart city. *International Journal of Information Management, 36*(5), 748-758.

Hashem, I. A. T., Yaqoob, I., Anuar, N. B., Mokhtar, S., Gani, A., & Khan, S. U. (2015). The rise of "big data" on cloud computing: Review and open research issues. *Information Systems, 47*, 98-115.

Hassani, H., & Silva, E. S. (2015). Forecasting with big data: A review. *Annals of Data Science, 2*(1), 5-19.

Hayes, M. A., & Capretz, M. A. (2015). Contextual anomaly detection framework for big sensor data. *Journal of Big Data, 2*(1), 2.

Hunt, S. T. a. E. (2016). Cyber attack: hackers 'weaponised' everyday devices with malware. Retrieved from https://www.theguardian.com/technology/2016/oct/22/cyber-attack-hackers-weaponised-everyday-devices-with-malware-to-mount-assault

Idé, T., Khandelwal, A., & Kalagnanam, J. (2016). *Sparse Gaussian Markov Random Field Mixtures for Anomaly Detection.* Paper presented at the Data Mining (ICDM), 2016 IEEE 16th International Conference on.

Introducing WSO2 Data Analytics Server. (2015). Retrieved from https://docs.wso2.com/display/DAS300/Introducing+DAS

Iqbal, R., Doctor, F., More, B., Mahmud, S., & Yousuf, U. (2016). Big data analytics: computational intelligence techniques and application areas. *International Journal of Information Management*.

Jones, N. (2016). Gartner Identifies the Top 10 Internet of Things Technologies for 2017 and 2018. Retrieved from http://www.gartner.com/newsroom/id/3221818

Kakavand, M., Mustapha, N., Mustapha, A., Abdullah, M. T., & Riahi, H. (2015). A survey of anomaly detection using data mining methods for hypertext transfer protocol web services. *Journal of Computer Science, 11*(1), 89.97.

Karim, A., Siddiqa, A., Safdar, Z., Razzaq, M., Gillani, S. A., Tahir, H., . . . Imran, M. (2017). Big data management in participatory sensing: Issues, trends and future directions. *Future Generation Computer Systems*.

Katal, A., Wazid, M., & Goudar, R. (2013). *Big data: issues, challenges, tools and good practices.* Paper presented at the Contemporary Computing (IC3), 2013 Sixth International Conference on.

Kerner, S. M. (2016). Cisco VNI: 2.3 Zettabytes of IP Traffic in 2020. Retrieved from http://www.enterprisenetworkingplanet.com/netsp/cisco-vni-2.3-zettabytes-of-ip-traffic-in-2020.html

Khalifa, S., Elshater, Y., Sundaravarathan, K., Bhat, A., Martin, P., Imam, F., . . . Statchuk, C. (2016). The six pillars for building big data analytics ecosystems. *ACM computing surveys (CSUR), 49*(2), 33.

Lai, Y., Liu, Z., Song, Z., Wang, Y., & Gao, Y. (2016). Anomaly detection in Industrial Autonomous Decentralized System based on time series. *Simulation Modelling Practice and Theory, 65*, 57-71.

Landset, S., Khoshgoftaar, T. M., Richter, A. N., & Hasanin, T. (2015a). A survey of open source tools for machine learning with big data in the Hadoop ecosystem. *Journal of Big Data, 2*(1), 24.

Landset, S., Khoshgoftaar, T. M., Richter, A. N., & Hasanin, T. (2015b). A survey of open source tools for machine learning with big data in the Hadoop ecosystem. *Journal of Big Data, 2*(1), 1.

Larson, D., & Chang, V. (2016). A review and future direction of agile, business intelligence, analytics and data science. *International Journal of Information Management, 36*(5), 700-710.

Lavin, A., & Ahmad, S. (2015). *Evaluating Real-Time Anomaly Detection Algorithms--The Numenta Anomaly Benchmark.* Paper presented at the Machine Learning and Applications (ICMLA), 2015 IEEE 14th International Conference on.

Lee, Y., & Lee, Y. (2011). *Detecting ddos attacks with hadoop.* Paper presented at the Proceedings of The ACM CoNEXT Student Workshop.

Lin, D. (2013). Big Data Analytics for Network Security Monitoring. Retrieved from http://blog.pivotal.io/data-science-pivotal/products/big-data-analytics-for-network-security-monitoring

Lin, W.-C., Ke, S.-W., & Tsai, C.-F. (2015). CANN: An intrusion detection system based on combining cluster centers and nearest neighbors. *Knowledge-based systems, 78*, 13-21.

Liu, Q., Klucik, R., Chen, C., Grant, G., Gallaher, D., Lv, Q., & Shang, L. (2017). Unsupervised detection of contextual anomaly in remotely sensed data. *Remote Sensing of Environment*.

Liu, X., & Nielsen, P. S. (2016). Regression-based Online Anomaly Detection for Smart Grid Data. *arXiv preprint arXiv:1606.05781*.

Lobato, A., Lopez, M. A., & Duarte, O. (2016). An accurate threat detection system through real-time stream processing. *Grupo de Teleinformática e Automaçao (GTA), Univeridade Federal do Rio de Janeiro (UFRJ), Tech. Rep. GTA-16-08*.

Maglaras, L. A., & Jiang, J. (2014). *Intrusion detection in scada systems using machine learning techniques.* Paper presented at the Science and Information Conference (SAI), 2014.

Mahmood, T., & Afzal, U. (2013). *Security analytics: Big data analytics for cybersecurity: A review of trends, techniques and tools.* Paper presented at the Information assurance (ncia), 2013 2nd national conference on.

Mascaro, S., Nicholso, A. E., & Korb, K. B. (2014). Anomaly detection in vessel tracks using Bayesian networks. *International Journal of Approximate Reasoning, 55*(1), 84-98.

McNeil, P., Shetty, S., Guntu, D., & Barve, G. (2016). SCREDENT: Scalable Real-time Anomalies Detection and Notification of Targeted Malware in Mobile Devices. *Procedia Computer Science, 83*, 1219-1225.

Media, M. R. (2017). U.S. Federal Cybersecurity Market Forecast 2017-2022. Retrieved from https://www.marketresearchmedia.com/?p=206.

Mirsky, Y., Shabtai, A., Shapira, B., Elovici, Y., & Rokach, L. (2017). Anomaly detection for smartphone data streams. *Pervasive and Mobile Computing, 35*, 83-107.

Mukherjee, M., Shu, L., & Wang, D. (2018). Survey of Fog Computing: Fundamental, Network Applications, and Research Challenges. *IEEE Communications Surveys & Tutorials*.

Muniyandi, A. P., Rajeswari, R., & Rajaram, R. (2012). Network anomaly detection by cascading k-Means clustering and C4. 5 decision tree algorithm. *Procedia Engineering, 30*, 174-182.

Nair, L. R., Shetty, S. D., & Shetty, S. D. (2017). Applying spark based machine learning model on streaming big data for health status prediction. *Computers & Electrical Engineering*.

Nakamura, Y., Suwa, H., Arakawa, Y., Yamaguchi, H., & Yasumoto, K. (2016). *Design and Implementation of Middleware for IoT Devices toward Real-Time Flow Processing.* Paper presented at the Distributed Computing Systems Workshops (ICDCSW), 2016 IEEE 36th International Conference on.

Promod, K., & Jacob, B. (2016). Mining a Ubiquitous Time and Attendance Schema Using Random Forests for Intrusion Detection. *Procedia Technology, 24*, 1226-1231.

Qiu, J., Wu, Q., Ding, G., Xu, Y., & Feng, S. (2016). A survey of machine learning for big data processing. *EURASIP Journal on Advances in Signal Processing, 2016*(1), 67.

Raguseo, E. (2018). Big data technologies: An empirical investigation on their adoption, benefits and risks for companies. *International Journal of Information Management, 38*(1), 187-195.

Ramamoorthi, A., Subbulakshmi, T., & Shalinie, S. M. (2011). *Real time detection and classification of DDoS attacks using enhanced SVM with string kernels.* Paper presented at the Recent Trends in Information Technology (ICRTIT), 2011 International Conference on.

Ranjan, R. (2014). Streaming big data processing in datacenter clouds. *IEEE Cloud Computing, 1*(1), 78-83.

Rathore, M. M., Ahmad, A., & Paul, A. (2016a). Real time intrusion detection system for ultra-high-speed big data environments. *The Journal of Supercomputing, 72*(9), 3489-3510.

Rathore, M. M., Ahmad, A., & Paul, A. (2016b). Real time intrusion detection system for ultra-high-speed big data environments. *The Journal of Supercomputing*, 1-22.

Rettig, L., Khayati, M., Cudré-Mauroux, P., & Piórkowski, M. (2015). *Online anomaly detection over big data streams.* Paper presented at the Big Data (Big Data), 2015 IEEE International Conference on.

Sahasrabuddhe, A., Naikade, S., Ramaswamy, A., Sadliwala, B., & Futane, P. (2017). Survey on Intrusion Detection System using Data Mining Techniques.

Sarker, R. A., Elsayed, S. M., & Ray, T. (2014). Differential Evolution With Dynamic Parameters Selection for Optimization Problems. *IEEE Trans. Evolutionary Computation, 18*(5), 689-707.

Security, H. N. (2017). Machine learning in cybersecurity will boost big data, intelligence, and analytics spending. Retrieved from https://www.helpnetsecurity.com/2017/01/31/machine-learning-cybersecurity/

She, C., Wen, W., Lin, Z., & Zheng, K. APPLICATION-LAYER DDOS DETECTION BASED ON A ONE-CLASS SUPPORT VECTOR MACHINE.

Shiravi, H., Shiravi, A., & Ghorbani, A. A. (2012). A survey of visualization systems for network security. *IEEE Transactions on visualization and computer graphics, 18*(8), 1313-1329.

Shirdastian, H., Laroche, M., & Richard, M.-O. (2017). Using big data analytics to study brand authenticity sentiments: The case of Starbucks on Twitter. *International Journal of Information Management*.

Singh, K., Guntuku, S. C., Thakur, A., & Hota, C. (2014). Big data analytics framework for peer-to-peer botnet detection using random forests. *Information Sciences, 278*, 488-497.

Slavakis, K., Giannakis, G. B., & Mateos, G. (2014). Modeling and optimization for big data analytics:(statistical) learning tools for our era of data deluge. *IEEE Signal Processing Magazine, 31*(5), 18-31.

Solaimani, M., Iftekhar, M., Khan, L., Thuraisingham, B., Ingram, J., & Seker, S. E. (2016). Online anomaly detection for multi-source VMware using a distributed streaming framework. *Software: Practice and Experience, 46*(11), 1479-1497.

Solaimani, M., Iftekhar, M., Khan, L., Thuraisingham, B., & Ingram, J. B. (2014). *Spark-based anomaly detection over multi-source VMware performance data in real-time.* Paper presented at the Computational Intelligence in Cyber Security (CICS), 2014 IEEE Symposium on.

Su, M.-Y. (2011). Real-time anomaly detection systems for Denial-of-Service attacks by weighted k-nearest-neighbor classifiers. *Expert Systems with Applications, 38*(4), 3492-3498.

Suthaharan, S. (2014). Big data classification: Problems and challenges in network intrusion prediction with machine learning. *ACM SIGMETRICS Performance Evaluation Review, 41*(4), 70-73.

Swift Liu, H. (2015, 25/02/2015). How Do Big Data Analytics Enhance Network Security? Retrieved from http://www.forbes.com/sites/huawei/2015/02/24/how-do-big-data-analytics-enhance-network-security/

Symantec™ Anomaly Detection for Automotive. (2016). Retrieved from https://www.symantec.com/content/dam/symantec/docs/data-sheets/anomaly-detection-for-automotive-en.pdf

TIBCO StreamBase and the TIBCO Accelerator for Apache Spark. (2017). Retrieved from https://www.tibco.com/sites/tibco/files/resources/wp-streambase-accelerator-apache-spark.pdf

ur Rehman, M. H., Liew, C. S., Abbas, A., Jayaraman, P. P., Wah, T. Y., & Khan, S. U. (2016). Big data reduction methods: a survey. *Data Science and Engineering, 1*(4), 265-284.

Wang, C., Zhao, Z., Gong, L., Zhu, L., Liu, Z., & Cheng, X. (2018). A Distributed Anomaly Detection System for In-Vehicle Network Using HTM. *IEEE ACCESS, 6*, 9091-9098.

Wang, L., & Jones, R. (2017). Big data analytics for network intrusion detection: A survey. *International Journal of Networks and Communications, 7*(1), 24-31.

Wauters, M., & Vanhoucke, M. (2017). A Nearest Neighbour extension to project duration forecasting with Artificial Intelligence. *European Journal of Operational Research, 259*(3), 1097-1111.

Weston, C. Y. (2015). On the K-Nearest Neighbor approach to the generation of fuzzy rules for college student performance prediction.

Wilkes, S. (2016). Making In-Memory Computing Enterprise Grade – Overview - Striim. Retrieved from http://www.striim.com/blog/2016/06/making-in-memory-computing-enterprise-grade-overview/

Woolf, N. (2016). DDoS attack that disrupted internet was largest of its kind in history, experts say. Retrieved from https://www.theguardian.com/technology/2016/oct/26/ddos-attack-dyn-mirai-botnet

Xie, S., & Chen, Z. (2017). Anomaly Detection and Redundancy Elimination of Big Sensor Data in Internet of Things. *arXiv preprint arXiv:1703.03225*.

Yaqoob, I., Ahmed, E., ur Rehman, M. H., Ahmed, A. I. A., Al-garadi, M. A., Imran, M., & Guizani, M. (2017). The rise of ransomware and emerging security challenges in the Internet of Things. *Computer Networks, 129*, 444-458.

Yaqoob, I., Hashem, I. A. T., Gani, A., Mokhtar, S., Ahmed, E., Anuar, N. B., & Vasilakos, A. V. (2016). Big data: From beginning to future. *International Journal of Information Management, 36*(6), 1231-1247.

Yasumoto, K., Yamaguchi, H., & Shigeno, H. (2016). Survey of real-time processing technologies of iot data streams. *Journal of Information Processing, 24*(2), 195-202.

Yin, C., Zhang, S., & Kim, K.-j. (2017). Mobile Anomaly Detection Based on Improved Self-Organizing Maps. *Mobile Information Systems, 2017*.

Yin, S., Zhu, X., & Jing, C. (2014). Fault detection based on a robust one class support vector machine. *Neurocomputing, 145*, 263-268.

Zhang, C., Lai, C.-F., Lai, Y.-H., Wu, Z.-W., & Chao, H.-C. (2017). An inferential real-time falling posture reconstruction for Internet of healthcare things. *Journal of Network and Computer Applications*.

Zhao, S., Chandrashekar, M., Lee, Y., & Medhi, D. (2015). *Real-time network anomaly detection system using machine learning.* Paper presented at the Design of Reliable Communication Networks (DRCN), 2015 11th International Conference on the.

Zhao, Y., Ni, Q., & Zhou, R. (2017). What factors influence the mobile health service adoption? A meta-analysis and the moderating role of age. *International Journal of Information Management*.

Zhou, L., Pan, S., Wang, J., & Vasilakos, A. V. (2017). Machine learning on big data: Opportunities and challenges. *Neurocomputing, 237*, 350-361.