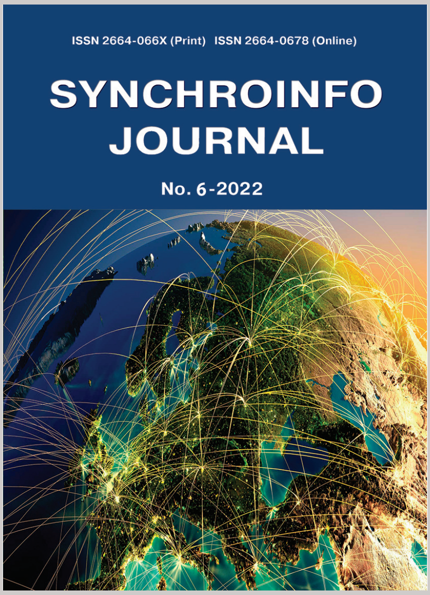
CONTENT

Vol. 8. No. 6-2022



ISSN 2664-066X (Print) ISSN 2664-0678 (Online)

SYNCHROINFO JOURNAL

No. 6-2022

Published bi-monthly since 2015.

ISSN 2664-066X (Print)
ISSN 2664-0678 (Online)

Publisher
Institute of Radio and
Information Systems (IRIS),
Vienna, Austria

Address:
1010 Wien, Austria,
Ebendorferstrasse 10/6b
media-publisher.eu/
synchroinfo-journal

VIRTUALIZATION SECURITY: A GATEWAY TO SECURE CLOUD COMPUTING

Ranabhat Saugatdeep
Moscow Institute of Physics and Technology (MIPT), Moscow, Russia, ranabkhat.s@phystech.edu

Alexey N. Nazarov
Federal Research Center Computer Science and Control of Russian Academy of Sciences, Moscow, Russia, a.nazarov06@bk.ru

Tuleun Terhemen Daniel
Moscow Institute of Physics and Technology (MIPT), Moscow, Russia, tuleun.t@phystech.edu

DOI: 10.36724/2664-066X-2022-8-6-2-9

ABSTRACT

Lately, the most efficient solution for an organization and individual of internet based distributed computing platforms is cloud computing. Processing, storage, and data management are just a few of the useful services that end users can get from cloud computing. However, it raises a lot of issues with privacy and security which needs to be resolved. Major component of cloud computing is virtualization technology, which helps to develop complex cloud services based on the externalization and composition of these resources. The primary concern with cloud computing security is virtualization security. The collective measures and methods that ensure the safety of a virtualization environment are referred to as virtualization security. It tackles the security risks that virtualized components encounter, as well as techniques for mitigating or preventing them. In recent days, attacks are mainly aimed either to the virtualization architecture or the infrastructure. We prefer to examine the security holes associated with virtualization in this paper and how concentrating on virtualization security might ultimately result in safeguarding cloud computing platform. Through this paper we tend to highlight the recent progress and an overview of the existing works performed on various dimensions of the cloud security. We conclude our paper by making several recommendations with respect to the attack vector for supporting cloud protection with a mathematical LP model. .

KEYWORDS: *Secure Cloud Computing, Virtualization Technology, Vir-tualization Security, Virtual Machine (VM), Data Privacy and Integrity.*

1. ntroduction

The provision of on-demand computer services, including applications, storage, and processing power, often over the internet on a pay-as-you-go basis, is known as cloud computing (CC). The location of service and other factors, like the operating system and underlying hardware on which it is running, are key concept in cloud computing, which are mostly irrelevant to the user. As an evolving computing model based on the Internet, it offers consumers quick and convenient data storage and network service, with the ability for computing resources to be dynamically deployed and shared in a scalable data center. CC is clearly superior to traditional self-hosting approach in terms of cost, speed of deployment, and flexibility; as a result, it is well suited to provide infrastructure services, platform services, software and storage services, among many other services. One of the primary concerns with CC has always been data privacy because the data needs to be secure from outside parties or malicious actors. Since cloud computing is used to share data, data theft is still a major danger that affects both consumers and cloud service providers (CSP) and is highly widespread.

Virtualization is a major technology for cloud computing services, facilities, and the consolidation of several standalone systems onto a single hardware platform by virtualizing computational resources (e.g., network, CPUs, memory, and storage). Virtualization is made possible by hardware abstraction, which conceals the complexity of controlling the physical computing platform and facilitates computing resource scaling. Hypervisors are used to implement it. A hypervisor is in charge of isolating Virtual Machines (VMs) so that they do not directly access the virtual disks, memory, or programs of other VMs on the same host. These two features are important aspects of cloud computing because they enable resource sharing and pooling in order to improve agility, flexibility, lower costs, and increase business value. CSP make substantial efforts to secure virtualization techniques by attempting to eliminate or reduce vulnerabilities, threats, and attacks. However, given the features of virtualization, cloud computing is today facing many new security concerns which can jeopardize personal privacy, company interests, and even state security. The cloud computing instances, such as Amazon's Simple Storage Service's service outrage and Google customers' documents leakage, have prompted concerns about the security of CC technology. In reality, the security issue has become a major impediment to the growth of cloud computing. Only by building a comprehensive information security defense system based on a detailed examination of all types of security problems can we ensure the healthy and sustainable development of CC. The most significant distinction between cloud computing and traditional IT environments is the virtual computing environment, which makes security more difficult. Through virtualization technology, different levels of application systems, such as server, software, data, network, and storage, can be segregated from one another, removing the reliance on physical devices for traditional IT architecture and transforming infrastructure into virtual resources that can be flexibly adjusted on demand. However, typical security tactics and strategies begin to fail in the virtual environment, resulting in new security challenges such as attacks between virtual machines or between the VM and the host, DDoS attacks, anti-virus, data or application security isolation, and so on.

This article presents the widely used virtualization technology, analyses the security risks associated with virtualization in cloud computing and based on their categories, recommends appropriate mitigation strategies.

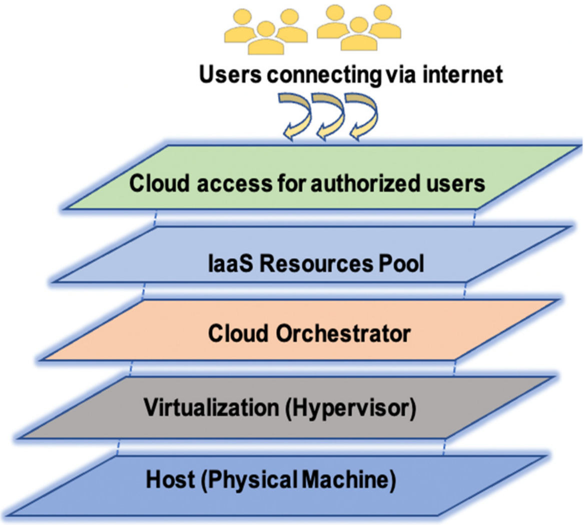

Users connecting via internet
Cloud access for authorized users
IaaS Resources Pool
Cloud Orchestrator
Virtualization (Hypervisor)
Host (Physical Machine)

Fig. 1. Cloud computing stack

2. Virtualization components

By adding a second layer of execution, including their own administrator job, virtualization techniques and virtualized architectures call for effective management and security protection. This layer consists of a number of different aspects, each of which contributes a specific part in the virtualization process as shown in Table 1 and thus could be a fresh target for malicious attack. The attack surface is increased by the virtualization layer's introduction of a complex mix of software elements at various levels of the computing architecture (such as the operating system, communication, administration, and interface), each with its own administrator credentials. While there is no doubt that virtualization approaches improve security, they also come with higher security concerns that must be considered while implementing a robust and safe virtualized infrastructure.

Table 1

Core components of virtualization

No.	Virtualization Components	Purpose
1	Host Machine	The physical device hosting the virtualized environment and its host operating system. The hypervisor operates on the host operating system, which also controls the physical hardware underlying the virtualized environment.
2	Hypervisor	Serves as a bridge between virtual machines and the underlying physical equipment. By sharing physical resources, it mediators all hardware requests made by virtual machines to physical hardware.
3	Virtual Machine Monitor	It records the actions taken by virtual machines. hardware requests are forwarded to actual resources and resource sharing across various virtual machines is supported.
4	Management Server	In charge of controlling virtual machines directly, combining services, assigning resources, moving virtual machines, and ensuring high availability.
5	Management Console	The core element of the virtualization software that gives users access to a management interface for setting up and controlling virtual machines.
6	Guest Machine	Utilizing the hardware abstraction offered by the virtual machine monitor, instantiate the virtualized (encapsulated) system composed of the operating system and applications, commonly known as virtual machines.
7	Network Component	Facilitate the creation of virtual networks where virtual network equipment (such as switches and routers) are entirely managed by software and the network stack and protocols are emulated to closely resemble actual ones.
8	Virtual storage	Provides every element needed to abstract physical storage in a single storage device that may be accessed directly or across a network.

3. Survey of related work

According to Sumter, L.Q. et al., [1] the creation of a system that tracked the movement and processing of data stored in the Cloud. There is a need for a security captured device on the cloud to ensure that users' data is secure and safe from security threats and attacks. Based on a case study, the proposed implementation was incorporated in a small Cloud computing environment. This proposal is only for the small-scale cloud framework not for large scale cloud domain. It is definitely used for provide security assurance to the users while communicating through the cloud domain or environment.

Hocenski [2] gave a broad overview of the security issues, requirements, and complexities that many CSP face. They characterized twenty recommended security approaches and their Cloud computing requirements that CSP should strongly consider as they develop or refine their compliance programmes.

All cloud computing services are linked to the virtualization technique. Iaas, PaaS, SaaS, NaaS (network as a service), and one more service that is entirely dependent on virtualization is known as virtualization as a service. The author of [3] stated that the security challenges how to effect cloud virtualization and what kind of security issues users and customers of cloud environment face. The security of virtual machines has already been prioritized and prioritized. The risk is a critical asset in cloud computing virtualization, and its mitigation at that virtualized level is being monitored by those people surveyed. In her paper, various security flaws are identified as existing or occurring now, and different algorithms by protocol and rules are applied to provide security to the virtualization layer.

Two types of controls in the cloud computing domain: physical control and virtual control were put forward in [4]. Encryption techniques must be implemented and used in the cloud. Encryption techniques can be used to encrypt data or information and secure and protect communication media that exists between the client and the cloud server. One main benefit of encrypted data is that it prevents data loss. We can use prevention policies to prevent data loss. Access control policies are also available to verify user authentication and protect against unauthorized access. Allow for the safe migration of data from a private cloud environment to a public CSP.

The fundamental problem, problems that users have while attempting to use the cloud computing virtualization application and how cloud computing has been successful in dealing with some issues such as data loss, data access, and data storage are explored in [5]. CC is the use of scientific knowledge to the reduction of some critical assets like cost reduction, hardware reduction, and money reduction. Only these notions are used in CC. Virtualization is mostly used in CC to share resources and create an isolated image, often known as a virtual image because it is not real. Another reason why virtualization is employed in CC is for dynamic resource allocation. The author discussed several virtualization challenges such as an infected application, data accuracy and mass data loss in cloud computing in this study. These challenges can be addressed by utilizing scientific knowledge that is virtualized.

Sabahi, F. et al., [6] discussed about Hyper-visor which helps to safeguard the environment of cloud computing technology and proposed a virtualized architecture for cloud security as well. The author attempted to explain various terminologies that aid in the security of virtualization in CC such as, reduce workload, decentralized security tasks between hypervisor and virtual machines and distribute the centralized security system When compared to a decentralized system, a centralized system has fewer flaws. One advantage is that defining the distributed system is beneficial for fault tolerance. According to the author, anything built by humans in this digital age is potentially breakable. However, the essential idea or concept is that decentralized application access presents a security concern in the cloud computing environment, which can cause a huge issue or problem while moving data in the cloud. So, from a security standpoint, it is necessary to control the security of the cloud computing environment as well as virtualization in the cloud environment.

With the use of Security Oriented Architecture, the challenges to networking environment in cloud computing domain can be evaluated [7]. The Security Oriented Architecture domain can carry out a number of tasks. As a result, it reduces the operation of information technology, saves money for cloud users giving a higher quality of flexibility and delivers needed services. This paper discussed the dangers of utilizing virtualization in cloud computing.

Security metrics are proposed in [9] that providers could use to evaluate the security of their services describing necessary policy rules and required procedures. In general, they emphasized to systematize in great detail, structure them according to the proposed levels and summarize the threats and vulnerabilities of the most common SaaS cloud model.

4. Security risk in virtualization

A hypervisor is used to virtualize resources, which allows for the dynamic allocation of resources to virtual computers (VMs). In cloud virtual systems, cryptography is used to safeguard data in transit from user to the cloud, but data must be decrypted in memory. This opens the door to privacy violations because virtualization improves an instance's memory pages virtually transparently, allowing a hostile provider to get data.

Table 2

Category of virtualization security risk

No.	Risk Type	Description
1	Virtual machine migration	VMs can be moved to another physical machine via "VM hot migration," which is done without shutting down the virtual machine. VM migration can be done for a variety of purposes, including load balancing or preserving fault tolerance. Data privacy and integrity may be lost as a result of the transfer, which exposes the VM’s contents to the network.
2	Virtual machine escape	Users can achieve mutual isolation and resource sharing on a virtual computer. In a virtual computer, a software typically should not interfere with other virtual machines. VM escape is the term for when applications operating in a virtual machine are able to get beyond isolation limits and execute directly on the host machine in some circumstances due to technological restrictions and flaws in virtualization technologies. If the virtual machine escape attack is successful, the host and the hypervisor are both in extreme threat.
3	Rookit attack	A specific type of malware called a rootkit conceals itself on the installation target and targets certain files, processes, and network links. What is more typical is that rootkits are frequently used in conjunction with other harmful software like Trojans and backdoors. Rootkits conceal information by loading unique drivers and changing the kernel of the operating system.
4	Denial of service attack	VM uses shared resources from a physical machine. A denial-of-service attack will occur in the virtual environments if an attacker uses one virtual machine to take full control of the host computer's resources, causing other virtual machines' resources to be disturbed or even crash due to a lack of resources.
5	Virtual machine monitor problem	The key component of virtualization is the virtual machine monitor, which generates or manages virtual resources as well as manages and isolates virtual machines. If the virtual machine monitor is penetrated, the attacker will have access to all other virtual machines it maintains as well as the metadata it stores about those virtual machines.
6	Decoupling attacks on virtualization platforms	In the conventional working context, just the physical machine with the system vulnerability is vulnerable, and the attack reach is very constrained. A single virtual machine can be used to attack the entire virtual vulnerability platform by several rent users using a cloud computing virtualization platform to share information across the entire system.

5. Counter measures

This section may each be divided by subheadings or may be combined. Following the classification of security risks, we provide several countermeasures and recommendations. The overall virtualization security protection model is shown in Figure 2.

a. Virtualization deployment planning, design, management

Administration of the infrastructure should be more rigorously organized than that of a physical one because it is far more important. In fact, the virtual infrastructure should be seen as a full machine in and of itself, housed in a single physical location with strict security constraints. Planning contributes to assuring security and adherence to all essential organizational policies. To maximize security while minimizing costs, organizations should address security from the outset of the systems development life cycle. Organizations must follow basic ICT security recommendations and practice such as keeping software up to date using host-based firewalls, antivirus, and IDS, to mention a few.

b. Restric and protect administrative access

The security of the entire virtualization infrastructure is dependent on the virtualization management system, which manages the hypervisor and supports operations on guest operating systems as well as other administrative tasks. Organizations should limit access to the virtualization management system or any other console interface, allowing only authorized administrators to access the hypervisor.

c. C trols against Hypervisor-based Attacks

Some of the strategies used to protect against hardware-based assaults are also successful in protecting hypervisors. Hardware-assisted Virtualization (HaV) is an efficient method that provides hypervisor security. HaV protects hypervisor integrity and improves isolation of system hardware resources. Hardware physical memory virtualization is possible with HaV; memory addresses are transformed from guest virtual to guest physical, then to system physical [8]. The system also includes a secure Input Output Memory Management Unit (IOMMU), which enables virtual machines to access peripheral devices directly. As a result, cyberattacks from malicious devices that could compromise the hypervisor's integrity are avoided.

On top of virtual machine migration, software-based security solutions secure virtual machines where tenants live. These solutions safeguard virtual computers both internally (at the operating system level) and externally (hypervisor level). Memory isolation, device isolation, and network isolation can all help to secure VMware's hypervisor.

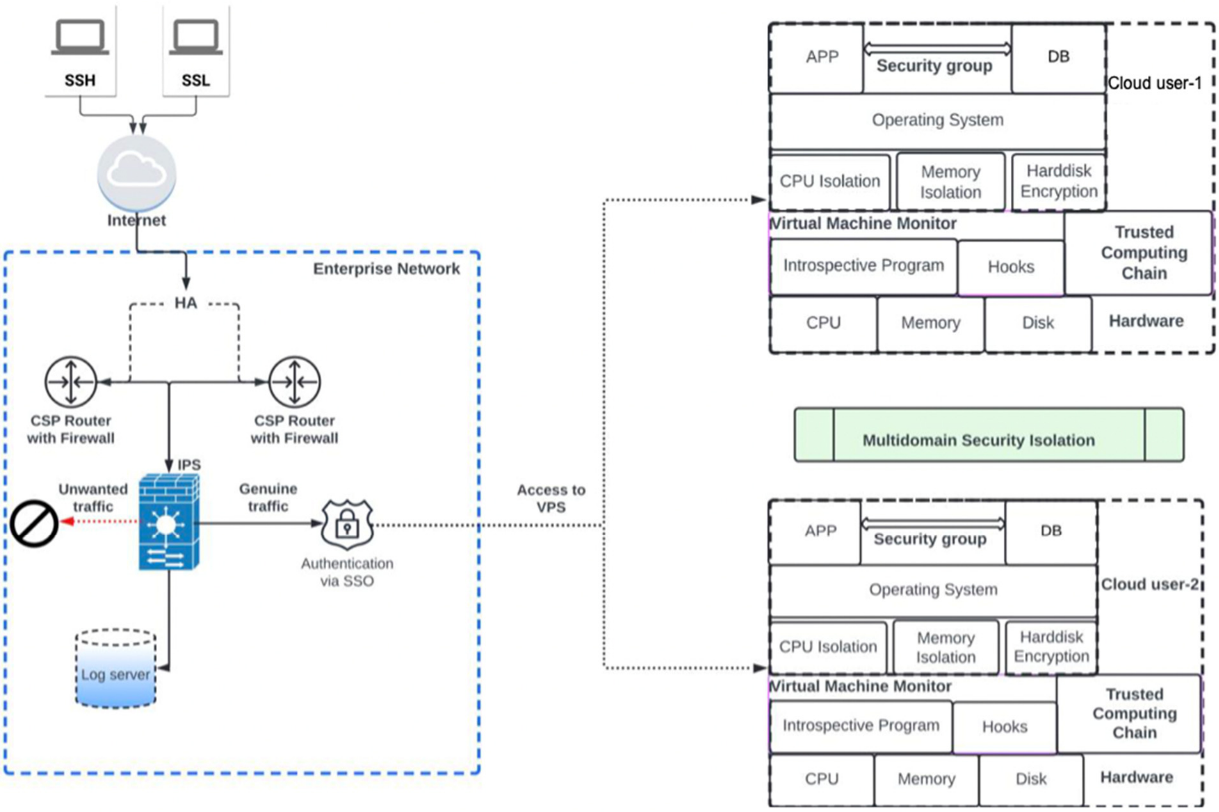


Fig. 2. Architecture of virtualization and cloud computing security

d. Access control for virtual network

The virtual network access control method can be accomplished using the Open vSwitch (OVS) technique, which inspects network data packets moving in and out through the security agent. The kernel driver layer's security agent enables virtual ports in the data flow route to load control function and accompanying security policy. This security agent mode can perform operations like security control optimization and policy synchronization. Furthermore, the Open vSwitch approach can be used to achieve several functionalities such as access control between vLANs, security domains, or virtual machines, as well as dynamic construction and migration of security policies.

e. S pshots and image management

Passwords and personal information may be concealed in images and snapshots. A company gains considerable security and operational benefits from proper image management. Images must be properly safeguarded against unwanted access, alteration, and replacement. Keeping a limited known good images of guest OSs is a recommended approach. Snapshots are more dangerous than images since they carry the RAM memory state. This could contain critical information that was not even kept on the drive itself, such as clear text passwords.

f. C trol the access to VMs

Limited access to virtual environments and effective privilege locking are required to reduce code exploitation via malicious software.

g. Isolate guest OS and perform partitioning

Policies for physical and logical partitioning should also be implemented for hypervisors. Partitioning prevents unauthorized access and reduces the possibility of code injection from one VM into another, as well as the risk of denial of service due to resource saturation. Side-channel attacks are also mitigated by VM separation. Isolation strategies have been used in this context to limit:

- access to VMs;
- communications between VMs
- communications from VMs to hypervisor.

h. Monito the resources

The hypervisor or VMM can be configured to monitor running virtual machines, network traffic, memory, and processes (introspection). Introspection also has auditing features as well as security controls like firewall policy, intrusion detection, and access control. Network-based security policies should not hinder traffic in a typical network arrangement. Active monitoring detects changes in system status and allows hook functions to be embedded in the monitored virtual machine.

Enhancing the virtualization security protection in cloud computing through the aforementioned features in Figure 2 will effectively boost cloud computing virtualization security, hence improving the security and reliability of cloud computing services.

6. Security-policy based model

A security policy based on rules (criteria) is proposed based on the proposed security measures related to the above-mentioned threats and protection measures. The considered measures and rules are essentially protection functions, used both to identify and prevent attacks and to assess the risk of attacks and intelligent monitoring of information security of modern info communications, including on a cloud model. The types of information attacks listed in the table are due to the technological features of VMs and the cloud resources used in specific systems. In general, for the calculation justification of cloud resources through which these attacks can be carried out, you can use the following methodological results.

Based on the analysis of [10] known cloud deployment models and the IaaS dynamic traffic model, the IaaS resource allocation model was developed. It is used to evaluate resources and tasks for the entire IaaS structure. The model can be used as a base for automating practical methods for multivariate calculation justification of necessary investments in IaaS design solutions with optimization of information flows. Developed on the basis of the entropy approach, the IaaS model of the system" information sources - boundary nodes " describes the most likely distribution of data from information sources over IaaS boundary nodes fixed in space, and all its coefficients have a certain physical meaning. The model allows you to get the values of the upper and lower estimates of the number of virtual connection (VC) as a function of the bit rate values of information sources. Based on the analysis of a stepwise bit dynamic model of content traffic of different categories of IaaS, resource allocation model of IaaS, a method has been developed for distributing the required resource of an IaaS fragment between all content categories in the interests of all resource tenants. The problem of conditional optimization in the "entropy setting" is formulated and solved by the method of undetermined Lagrange multipliers to find the distribution matrix of virtual connections between all tenants of the IaaS fragment resource, which are supplied with different categories of content.

Based on the results of solving this problem of conditional optimization, an optimization model for the distribution of virtual connections from content sources of different categories by service tenants in the IaaS fragment was obtained. The model reduces to solving a complete system of linear equations, and all its coefficients have a certain physical meaning. The model allows you to get the upper and lower estimates of the number of virtual connections as functions of the values of bit or packet speed of content sources of different categories. The conditions of model applicability, accuracy and stability of decisions on estimating the number of virtual connections in the IaaS fragment are determined.

The results obtained in [10] can be useful in solving the problem of structural-topological synthesis of IaaS at the pre-project stage.

Risk- an object of system of virtualization (SV) are being attacked by the intruder, consists of two components:

a) The probability of failure counter attack against the user (hereinafter - the failure of the object) or the probability of a successful attack.

b) Evaluation (e.g., financial, material, time to repair the damage, etc.) scale consequences (damage) of a successful attack.

The object of risk is considered to be sufficiently protected if given the opportunity to overcome potential barriers probability of a successful attack (the probability of the risk, the probability of failure or vulnerability of the object of risk) $P_A^Y = \left(1 - P_З^Y\right)$ than minimum value $P^Y_{А-ДОП}$, i.e., $P_З^Y \geq 1 - P^Y_{А-ДОП}$ – the condition of the feasibility, (1) where $P_З^Y$ - the probability of a successful counter attack (immunity, the success of the object of risk) subject to risk.

For any object risk of SV in the general case there is a complete system (list) security functions or attributes, each of which is in Table 3, denote the binary logic variable with the appropriate subscript.

Table 3

Security Features

Designation of security functions	**Appointment of security functions**
X_1	Preventing the occurrence of conditions conducive to the generation of (occurrence) destabilizing factors Preventing the occurrence of conditions conducive to the generation of (occurrence) destabilizing factors
X_2	Warning immediate manifestations of destabilizing factors
X_3	Detection manifested destabilizing factors
X_4	Prevention of exposure to risk in the manifested and revealed destabilizing factors
X_5	Prevention of exposure to risk on the manifest, but the undetected destabilizing factors
X_6	Detecting the impact of destabilizing factors on the subject of risk
X_7	Localization (restriction) found the impact of destabilizing factors on the subject of risk
X_8	Localization of undetected exposure to risk by destabilizing factors
X_9	Dealing with the consequences of the localized impact of the detected object on the destabilizing factors risk
X_{10}	Dealing with the consequences of undetected localized exposure to risk by destabilizing factors

The result of each of the Security Functions, or the outcome is a random event and can take two values – success or failure. It is assumed that a binary logical variable X_j, $j = 1 \div n$, $n = 10$ is equal to 1 with probability P_j if the execution of the $j-$ Security Functions has led to the failure risk of the object, and this binary logical variable equal to 0 with a probability $Q_j = 1 - P_j$ otherwise.

In general, the logic function (L-function), the success of the attack, realizing the impact of destabilizing factors as [11] and [12], $Y = Y\left(X_1, \ldots, X_n\right)$, and the probability function (P-function, P-polynomial) the risk of failure of the object – $P\left(Y = 1/X_1, \ldots, X_n\right) = \Psi\left(P_1, \ldots, P_n\right) = PY$.

According to the general case of [11, 12], L-function of the success of an attack is:

$$Y = X_1 X_2 \left(\overline{X_3} X_4 \vee X_3 X_5\right) \times$$
$$\times \left(\overline{X_6 X_7} X \ \vee X_6 \overline{X_8} X_{10} \vee \overline{X_6} X_7 \vee X_6 X_8\right) \qquad (2)$$

and the probability of success of an attack can be calculated using the B-polynomial

$$PY = PY\left(P_1, P_2, \ldots, P_{10}\right) = P_1 P_2 \left[\left(1 - P_3\right) P_4 + P_3 P_5\right] \times$$
$$\times \left[\left(1 - P_6\right)\left(1 - P_7\right) P_9 + P_6 \left(1 - P_8\right) P_{10} + \left(1 - P_6\right) P_7 + P_6 P_8\right]. \qquad (3)$$

The barriers a-h that are created to counteract the negative effects of destabilizing factors on the subject of risk, perform certain security functions that prevent the implementation of the attacks on the subject of risk in SV. Such changes are methodically quite easily and flexibly taken into account by introducing gradations of protection functions (1).

By analogy with the above, we assume that the binary logical variable X_{ij}, *i=1,2,...,n, n=10* corresponding to the *j-th* gradation of the known *j-th* security function, is equal to 1 with probability P_{ij} , if because of it the execution of the known *j*- th security function led to the failure of the risk object, and is equal to 0 with probability $Q_{ij} = 1 - P_{ij}$ otherwise.

In view of above specific a-h barriers to develop constructive solutions to counter its list of security functions in Table 3 appropriate to expand according to Table 4.

New security functions for SV are the logical components of the functions of the Tables 2 and 3:

$$X_i^{SV} = \begin{cases} X_i \vee X_{ij}, \ \forall X_{ij} \neq \Phi, \ i=1,2,\ldots,n, \ n=10, \ j=a,b,\ldots h, \\ X_i, \ \forall X_{ij} = \emptyset, \ i=1,2,\ldots,n, \ n=10, \ j=a,b,\ldots h, \end{cases} \qquad (4)$$

where $\emptyset -$ empty set.

Table 4

Security functions for SV

Designation of security functions ***i*=1,2,…10**	**Appointment of security functions**
X_{ia}	Virtualization deployment planning, design and management
X_{ib}	Restrict and protect administrative access
X_{ic}	Controls against Hypervisor-based Attacks
X_{id}	Access Control for Virtual Network
X_{ie}	Properly manage images and snapshots.
X_{if}	Control the access to VMs
X_{ig}	Isolate guest OS and perform partitioning.
X_{ih}	Monitor the resources

Substituting (3) into (1) obtained the L-polynomial for SV

$$Y^{SV} = X_1^{SV} X_2^{SV} \left(\overline{X_3^{SV}} X_4^{SV} \vee X_3^{SV} X_5^{SV}\right) * \\ * \left(\overline{X_6^{SV} X_7^{SV}} X_9^{SV} \vee X_6^{SV} \overline{X_8^{SV}} X_{10}^{SV} \vee \overline{X_6^{SV}} X_7^{SV} \vee X_6^{SV} X_8^{SV}\right) \qquad (5)$$

To assess the probability $Q_i^{SV}\ \forall X_{ij} \neq \Phi,\ i = 1,2,\dots,n,$ $n = 10,\ \ j = a, b, \dots h,$ let 's apply the Bayes formula,

$$Q(X_{jr}^{SV} / X_j^{SV}) = \frac{Q(X_{jr}^{SV}) Q(X_j^{SV} / X_{jr}^{SV})}{\sum_{\substack{r=a \\ \forall X_{jr} \neq \emptyset}}^{f} Q(X_{jr}^{SV}) Q(X_j^{SV} / X_{jr}^{SV})}, \forall X_{jr} \neq \emptyset. \qquad (6)$$

Using (6), the expression (3) for the B-polynomial and the formula (1) for a new kind of reachability condition are refined. Formula (6) can be used for interactive training (setup, identification) of the B-model (5) based on statistical data in order to clarify the current risk value.

Conclusion

As cloud computing technology advances, security risks and threats will increase in the future. Virtualization, the foundation of cloud computing is vulnerable to security risks, which has slowed down cloud computing's expansion and widespread adoption. However, despite the fact that many features are created with these objectives in mind, it poses a risk to secrecy, integrity, authenticity, and non-repudiation.

To secure user's application and their cloud account, users must enable all risk mitigation rules because security issues can take many different forms. Analysing the primary security risks that virtualization technology faces, enhancing virtualization's security defenses are very important. It is challenging to comprehend all the security concerns and privacy threats with the development of effective solutions. Work must be done to support virtualization and cloud computing, as well as to comprehend the difficulties posed by cloud security issues.

This paper also intends to give this field of study a new direction and assist researchers in identifying potential countermeasures to such risks and threats. Based on new security features that take into account the technological features of the general case of cloud virtualization, a new Security Protection Model has been developed. The proposed LP-models (LP-polynomials) for the general case of an arbitrary risk object SV create prerequisites for assessing the fulfilment of the reachability condition for each specific risk object.

References

[1] R. La‘Quata Sumter, “Cloud Computing: Security Risk Classification". *ACMSE Oxford. USA*, pp. 15-17, Apr 2010. https://doi.org/10.1145/1900008.1900152

[2] Kresimir Popovic et.al. “Cloud Computing Security Issues and Challenges” Proc. 33rd *Int’l Convention on Information and Comm. Technology. Electronics and Microelectronics* (MIPRO 10). IEEE Press, pp. 344-349, Jun 2010.

[3] H.M. Anitha, P. Jayarekha, “Security Challenges of Virtualization in Cloud Environment”. *International Journal of Scientific Research in Computer Science and Engineering*. Vol. 6, Issue. 1, pp.37-43, Feb 2018. https://doi.org/10.26438/ijsrcse/v6i1.3743

[4] T. Swathi, K. Srikanth, S. Raghunath Reddy, “Virtualization in Cloud Computing”. *International Journal of Computer Science and Mobile Computing*, IJCSMC, Vol. 3, Issue. 5, pp. 540-546, May 2014.

[5] L. Malhotra, D. Agarwal and A. Jaiswal, “Virtualization in Cloud Computing”. *Journal of Information Technology and Software Engineering,* Dec 2014, 4:2. DOI: 10.4172/2165-7866.1000136

[6] Farzad Sabahi, “Secure Virtualization for Cloud Environment Using Hypervisor-based Technology”. *International Journal of Machine Learning and Computing*. Vol. 2. No. 1, Feb 2012. https://doi.org/10.7763/ijmlc.2012.v2.87

[7] Mladen A. Vouch, “Cloud Computing Issues. Research and Implementations”. *Journal of Computing and Information Technology*. pp. 235-246, Jun 2008. https://doi.org/10.1109/ITI.2008.4588381

[8] Z. Gilani, A. Salam, and S. Ul Haq, "Deploying and managing a cloud infrastructure: real world skills for the CompTIA cloud+ certification and beyond," *Wiley*, Jan. 2015.

[9] D. Georgiou, and C. Lambrinoudakis, “Compatibility of a security policy for a cloud-based healthcare system with the EU general data protection regulation (GDPR),” Information, 11(12), p. 586, Dec 2020. Available at: https://doi.org/10.3390/info11120586

[10] A.N. Nazarov et al, Alireza Nik Aein Koupaei, Anshita Dhoot, Asyraf Azlan & Seyed Milad Ranaei Siadat, “Mathematical modelling of infrastructure as a Service,” *Systems of Signals Generating and Processing in the Field of on Board Communications*, Mar 2020. Available at: https://doi.org/10.1109/ieeeconf48371.2020.9078629

[11] A.N. Nazarov, & M.M. Klimanov, “Estimating the informational security level of a typical corporate network”. *Automation and Remote Control*, no, 71(8), pp. 1550-1561, Aug 2010. https://doi.org/10.1134/s0005117910080059

[12] A. Nazarov, M. Klimanov, “Characteristic analysis of logic and probabilistic model of information security”. Sofia, Bulgaria, *Proceedings of International Workshop on Distributed Computer and Communication Computer and Communication Networks*, pp. 154-164. Published by Research and Development Company “Information and Networking Technologies”. Oct 2009.

ANALYSIS OF THE PROBLEM OF MULTIVALUED OF CLASS LABELS ON THE SECURITY OF COMPUTER NETWORKS

D. I. Rakovskiy,
Moscow Technical University of Communications and Informatics (MTUCI), Moscow, Russia
Prophet_alpha@mail.ru

DOI: 10.36724/2664-066X-2022-8-6-10-17

ABSTRACT

Modern computer networks have a complex infrastructure that requires constant monitoring to detect anomalous conditions that can cause malfunctions, which is unacceptable for large-scale distributed networks. An important problem in the intelligent processing of syslog data is the existence of multi-label datasets. Among the Russian-language scientific publications, the problem under consideration in the context of information security of computer networks is not presented. The purpose of the research work is to increase the security of computer networks through the use of multi-label learning methods in solving the problem of classifying system log class labels. In this paper, a comparative analysis of single-label and multi-label classifiers in a computational experiment on the Mean accuracy metric was carried out. According to the results of the analysis, 80% of single-label classifiers were inferior in classification accuracy according to the Mean accuracy multi-label metric to their counterparts, which may indicate a strong influence of multi-label class labels on the models under consideration. The considered structure of experimental data in a tabular form is influenced by the multi-label problem much more strongly than it can be estimated by a standard frequency check, which actualizes further research in this direction.

KEYWORDS: *supervised learning, multi-label classification, multiclass classification, information security, multi-label learning*

I. Introduction

Modern computer networks (CNs) have a complex infrastructure that requires constant monitoring to detect anomalous conditions that can cause malfunctions, which is unacceptable for large-scale distributed CNs [1].

CN security can be achieved through the use of classical measures to prevent traffic interception – the installation of software and hardware information protection [2]; intrusion detection and prevention systems [3, 4]; antivirus software [5] and other solutions [6].

Works devoted to the development of software solutions for detecting and leveling cyber threats in CN are relevant [7]. Known works on the estimation and prediction of the state of complex objects: application for information security [8, 9].

An important problem in the intelligent processing of syslog data is the existence of datasets containing entries with multiple class label associations. That is, the class associated with an object is characterized by a set of labels.

A dataset suitable for classification typically contains a set of features and an associated set of class labels. The goal of classification is a trained model capable of assigning an appropriate class to an unknown object (records in "historical data").

Works, one way or another, exploring the problems of multi-label, are united by the term: Multi-Label Learning, MLL [10]. MLL generalizes the notion of data analysis to the realm of tasks, in which multiple labels can be associated with each object. Among these articles, a cluster of works on the analysis of text corpora [11] and the tone of messages in social networks [12] stands out.

Domestic works devoted to the analysis of data sets generated by CN with multi-label and class labels are currently not presented. Existing works, for example, [13, 14], are devoted to aspects of fuzzy classification. Fuzzy classification belongs to the field of fuzzy logic, which is part of the multi-class learning methods.

MLL is indirectly related to the concept of Misclassification. The term is currently used to label works devoted to solving problems of data mislabeling [15] and improving classification accuracy [16].

Information security characterizes the preservation of the properties of confidentiality, integrity and availability of information [17]. It should be noted that the analysis of the impact of multi-label class labels on CN security must be carried out in a certain terminological context. GOST R ISO/IEC 27000, from which the above definition of information security derives, as well as GOST R ISO/IEC 12207, was chosen as such a context. According to the referenced document, paragraph 3.25, "Security: The ability of a computer system to protect information and data so as to prevent their unauthorized reading or modification by other systems and individuals, and so that systems and individuals admitted to them do not receive failures."

It is necessary to clarify the security of information circulating in the CN: "The security of information is the maintenance at a given level of those parameters of the information located in the automated system that characterize the established status of its storage, processing and use" [18]. It follows from the two definitions that the security of information circulating in the CN is related to the security of the supporting infrastructure [19].

Within the framework of this work, we will analyze the influence of multi-label on the accuracy of the classification of CN states that are directly related to the profile of the normal functioning of CN [20].

The aim of the work is to increase the security of computer networks through the use of multi-label learning methods in solving the problem of classifying system log class labels.

II. Generation of CN class labels

The CN can be represented as a set of M sets of values of discretely changing attributes of "historical data" of the CN:

$$\boldsymbol{A} \subseteq A_{first} \cup A_{second} = \{A_{first\,1} \times A_{first\,2} \times \ldots \times A_{first\,len_1}\} \cup \\ \cup \{A_{second\,1} \times A_{second\,2} \times \ldots \times A_{second\,len_2}\}; \qquad (1)$$

In this equations, $A_m = \{a_{mn}; m = \overline{1,M}, n = \overline{1,N}\}$, $A_m \subset \boldsymbol{A}$, $M = len_1 + len_2$.

Attributes in (1) can be divided into two types: primary $\{A_{first\,k_1}; k_1 = \overline{1,len_1}\}$ and secondary $\{A_{second\,k_2}; k_2 = \overline{1,len_2}\}$.

The primary attributes are obtained directly from the system sensors installed inside the CN. Secondary attributes are obtained as a result of processing primary attributes. Examples of secondary attributes can be, for example, the average signal delay time in the CN, the number of lost packets in the CN for a particular host, and so on.

To describe CN, we introduce a set of class labels of categorical type - S, which we will call "CN states". The CN states can also be entered as a set:

$$\boldsymbol{S} = \{S_1, S_2, \ldots, S_M\} \cup \{s_{normal}\}; \; S_m = \{s_i; i = \overline{1,I}\} \qquad (2)$$

where S_m – m-th subset of the CN states associated with the corresponding A_m attribute of CN. Power of the subset S_m has an upper bound equal to I. In practice, the subsets included in S may have different powers. An example of elements with different power is the inequality $|S_1| \neq |S_2|$.

In the case of $\forall S_m = \varnothing$ status s_{normal} is entered, characterizing the normal functioning of the CN.

To automate the process of determining the states of the CN we introduce a set of rules

$$\boldsymbol{METARULES} = \{RULE_1, RULE_2, ..., RULE_M\},$$
$$RULE_m = \{r_{mj}; j = \overline{1, |S_m|}\} \quad (3)$$

Each subset - $RULE_m$ is associated with the corresponding subset of CN states by the m-th attribute S_m. The power of a subset $RULE_m$ depends on the power of the corresponding subset S_m. The iteration variable j is introduced to account for the difference in power of different subsets S_m. If all subsets are identical S_m, $j = \overline{1, |S_m|} \equiv i = \overline{1, I}$, the upper boundary will be identical to I.

Decisive rules are proposed to be selected based on the individually entered Service Level Objectives, SLO, based on the technical and operational characteristics of the CN.

Consider the process of labeling a set of attributes corresponding to n observations of historical data (*n* rows in the table of historical data) – $\{a_{1n}, a_{2n}, ..., a_{Mn}\}$. The specified string is an argument of the labeling function $mark(\{a_{1n}, a_{2n}, ..., a_{Mn}\})$, and forms a set of labels set_n, corresponding to the *n*-th line.

The set_n is formed by checking each attribute of the *n* string $\{a_{1n}, a_{2n}, ..., a_{Mn}\}$ – for compliance with the rules of the corresponding set $RULE_m$ (3) – r_{mj}. If the rule r_{mj} is fulfilled, then in the set of labels set_n element s_{mj} is added, where $j = \overline{1, |S_m|}$.

The labeling process can be formalized as:

$$mark : \{a_{1n}, a_{2n}, ..., a_{Mn}\} \to set_n;\, set_n \subseteq \boldsymbol{S}, \text{ where}$$
$$mark(\{a_{1n}, a_{2n}, ..., a_{Mn}\}) = \begin{cases} set_n, \text{ if } set_n \neq \varnothing \\ s_{normal}, \text{ otherwise} \end{cases},$$
$$\text{where } set_n = \left\{ \begin{matrix} s_{mj} \in S_m \mid r(a_{mn}, j) = 1, \\ j = \overline{1, |S_m|},\, m = \overline{1, M} \end{matrix} \right\}, \quad (4)$$
$$\text{where } r(a_{mn}, j) = \begin{cases} 1, \text{ if rule } r_{mj} \in RULE_m \text{ is followed} \\ 0, \textit{otherwise} \end{cases}$$

If none of the rules are satisfied, then $set_n = \left\{ s_{mj} \in S_m \mid r(a_{mn}, j) = 1,\ j = \overline{1, |S_m|},\, m = \overline{1, M} \right\} = \varnothing$.

This means that the result of marking will be a predetermined state of the CN - s_{normal}. Each item r_{mj}, is a freely defined verbal-logical rule introduced for a particular CN.

The rules can be paired with the security policy relevant for the CN: with the threat model; with the SLO service level indicators; with other methods of security and service quality assessment.

When marking rules affect the data of the CN, each record (string – $\{a_{1n}, a_{2n}, ..., a_{Mn}\}$) is assigned either a set of states set_n, according to relation (4), or the state s_{normal}.

The labeling of "historical data" about the behavior of the CN can be represented as a table of size M columns by N rows:

$$D_N = \{(\{a_{1n}, a_{2n}, ..., a_{Mn}\}, set_n); m = \overline{1, M}, n = \overline{1, N}\},$$

where the *n*-th row of record attribute values $\{a_{1n}, a_{2n}, ..., a_{Mn}\}$ the state of the CN and the set of labels set_n.

Although not the only way of marking experimental data but is the most convenient in terms of organizing the processing and analysis of data by specialized software tools.

III. Structure and description of the studied network infrastructure

Research to evaluate network performance was carried out on a CN consisting of 6 hosts forming a cluster managed by Rancher (Fig. 1) [21]. The host interaction architecture of the studied CN is based on the principle of virtualization and interaction between Docker containers; services managed by an Apache Spark cluster; databases (PostgreSQL; Apache Ignite; Apache Cassandra; Redis); Apache Ignite cluster; software based on microservice architecture and other auxiliary modules.

Technical characteristics of distributed CN host machines are given in Table. 1. Machines #1 - #3 form the physical topology of the distributed CN; machines No. 4-6 operate through virtualization by the VMware ESXI operating system based on machines No. 1-3.

To collect data on 6 CN machines, special software for obtaining information from system sensors was used: *packetbeat* (aggregates HTTP and DNS request protocol traffic); *metricbeat* (aggregates data on CPU usage, disk usage, memory usage, network usage, system processes); *filebeat* (aggregates message log data); *execbeat* (aggregates the execution of specialized scripts and sending the result of their execution).

To collect indicators related to SLO, the CN under consideration implements a system for synchronous monitoring of all hosts. The scheme for collecting indicators is shown in Figure 2. The received data is aggregated in a centralized storage managed by Apache Cassandra.

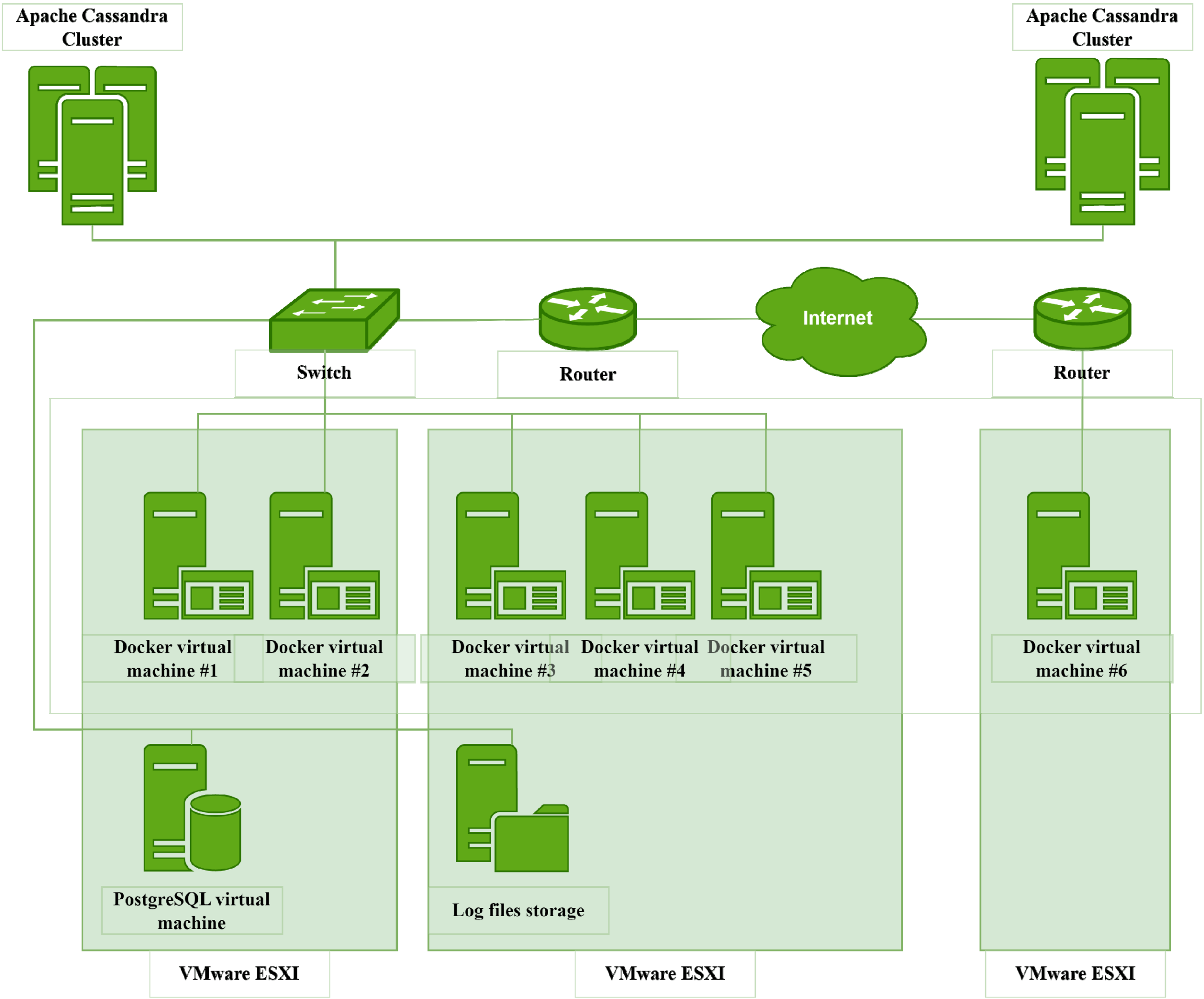

Apache Cassandra Cluster
Apache Cassandra Cluster
Switch
Router
Internet
Router
Docker virtual machine #1
Docker virtual machine #2
Docker virtual machine #3
Docker virtual machine #4
Docker virtual machine #5
Docker virtual machine #6
PostgreSQL virtual machine
Log files storage
VMware ESXI
VMware ESXI
VMware ESXI

Fig. 1. Scheme of the studied network infrastructure

Table 1

CN configuration

No	Host mapping to a virtual machine	Operating system	Numbe r of Cores	RAM , GB	Hard disk capacity (total), GB	Processor Model
1	server3-20 (physical machine from the cluster Apache Cassandra)	CentOS Linux 7	4	64	1524	Intel(R) Xeon(R) CPU E3-1220 v6 @ 3.00GHz
2	server3-21 (physical machine from the cluster Apache Cassandra)	CentOS Linux 7	4	64	1524	Intel(R) Xeon(R) CPU E3-1220 v6 @ 3.00GHz
3	server3-22 (physical machine from the cluster Apache Cassandra)	CentOS Linux 7	4	64	1524	Intel(R) Xeon(R) CPU E3-1220 v6 @ 3.00GHz
4	server24- 384-1 (Docker virtual machine №1; Docker virtual machine №2)	Ubuntu 18.04.1 LTS	5	50,05	68	Intel(R) Xeon(R) CPU E5-1650 v4 @ 3.60GHz
5	server24- 384-2 (Docker virtual machine №3; Docker virtual machine №4; Docker virtual machine №5)	Ubuntu 18.04.1 LTS	6	48,61	265	Intel(R) Xeon(R) CPU E5-2420 v2 @ 2.20GHz
6	server24- 384-3 (Docker virtual machine №6)	Ubuntu 18.04.1 LTS	8	60	285	Intel(R) Xeon(R) CPU E5-2420 v2 @ 2.20GHz

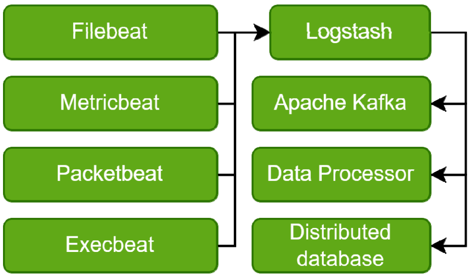


Fig. 2. Diagram of using the SLO-related metrics aggregation system

On the hosts given in Table 2, in accordance with the diagram in Figure 2, aggregators of software and hardware metrics are installed: *Packetbeat* (host network activity aggregator – traffic monitoring, HTTP protocol and DNS queries) [22]; *Metricbeat* – an aggregator of indicators associated with the operating system and host hardware devices – the use of CPU, memory, disks, running processes; *Filebeat* is a syslog aggregator; *Execbeat* – software for CN tests by generating and sending arbitrary scripts for execution. *Execbeat* was used to send ICMP requests (ping requests) to determine network latency and send GET requests using it to determine the server's response time to a sent request.

Each of the four types of aggregators sends data to the central point CN – the *Logstash* log aggregator, which converts all incoming information into JSON files. The choice of format is due to the generally accepted notation of the structure of the structure of JSON files. The described stack of aggregators is widely used in the construction of information processing systems in the field of information security [23-26].

After conversion, the JSON file is sent to the Apache Kafka message handler, which performs a buffering function between the large input stream and the distributed database. Hyperlinks to said software are provided in footnote [1].

IV. The problem of primary (first) and secondary (second) attributes

Let us consider an illustrative example of the results of single-valued classifiers - binary and multi-class - with multi-label experimental data on the basis of the results given in [15]. A fragment of this data is given in Table I.

An actual applied problem is the determination of CN states without knowledge of secondary attributes. In this case, the labels of the SLO classes are determined only on the basis of the primary data of the system sensors under conditions of partial uncertainty of the remaining parameters.

Consider two cases:

1. Complete a priori certainty of both primary and secondary attributes of CN at each moment of time;

2. Partial uncertainty of CN secondary attributes that are either unknown or computed with a long delay.

With full attribute information (A_{first} and A_{second},), due to total dependency set_n on A_{second}, the task of classifying the CN state is performed by a multi-label classifier with an accuracy close to ideal, i.e. without mistakes. An obstacle to such an ideal classification is the identification of direct transformation rules A_{second} to set_n ($A_{second} \to set_n$).

If the rules are represented by trivial logical conditions "if … then …", then the classification accuracy of many rule-based classifiers (for example, decision trees or neural networks) will be close to ideal. If the secondary attributes are unknown, but the primary attributes and the corresponding CN states are known, the secondary attributes will be a latent variable. In the absence of information about secondary attributes, the single-label mapping of primary attributes to CN states is not guaranteed, since secondary attributes become latent variables. However, the fundamental possibility of displaying primary attributes in CN states is still possible.

V. Computational Experiment

To compare two classification methods - the "classic" single-label and multi-label - let's conduct a computational experiment in Python with the following input data. It is proposed to consider the single-label approach to classification using the example of multiclass algorithms selected according to two criteria:

– openness of the source code (the library that implements this algorithm is in the public domain);

– availability of multi-label implementation of this algorithm. According to the established criteria, the following algorithms were selected from the open scikit-learn library of the Python programming language [27]:

Tree.DecisionTreeClassifier – classifier generated on the basis of the "Decision Tree" algorithm (non-parametric supervised learning method);

Tree.ExtraTreeClassifier – A classifier generated on the basis of the "Extra Decision Tree" algorithm (non-parametric supervised learning method). When searching for the best split to split the node samples into two groups, for each of the randomly selected attributes, the best split is selected according to the specified criterion;

Ensemble.ExtraTreesClassifier – classifier generated on the basis of the "Extra Decision Tree" algorithm (ensemble implementation);

Neighbors.KNeighborsClassifier – Classifier generated based on the voting algorithm "K-Neighbors";

[1] - PACKETBEAT. Lightweight shipper for network data // Elastic URL: https://www.elastic.co/beats/packetbeat
- METRICBEAT. Lightweight shipper for metrics. // Elastic URL: https://www.elastic.co/beats/metricbeat
- FILEBEAT. Lightweight shipper for logs. // ElasticURL: https://www.elastic.co/beats/filebeat
- Elastic beat to call commands in a regular interval and send the result to Logstash // Elasticsearch URL: https://github.com/christiangalsterer/execbeat
- Apache Kafka. A distributed streaming platform. // Apache Kafka URL: https://kafka.apache.org/

Ensemble.RandomForestClassifier - A classifier generated on the basis of the "Random Forest" algorithm (ensemble implementation).

The indicators of the SLO service level (decision rules) and the corresponding CN states associated with secondary attributes are obtained, formed in the form of thresholds that determine the categorical markers CN states:

– If none of the service level objectives has been violated, then the CN state is equal to the normal marker.
– If the signal delay time to the test server (*ping_avg*) > 5 ms, then CN state is equal to the *signal_delay* marker.

– If the response time of the test server (*server_response_timetotal*) > 1.5 s, then the CN state is equal to the *server_response_delay* marker.

- If the number of packets lost during transmission to the test server (*network_outdropped*) > 0, then the CN state is equal to the *packets_dropped* marker.

– If the time to process a request by the disk of the host machine (*disk_ioreadmergespersec*) > 2 s, then the CN state is equal to the *disk_iowriteawait* marker.

The number of decision rules and the CN attributes affected by them can be increased depending on the task, but 5 class labels are sufficient for illustration.

Consider the distribution of experimental data over the number of simultaneously violated service level indicators. The initial distribution is shown in Figure 3.

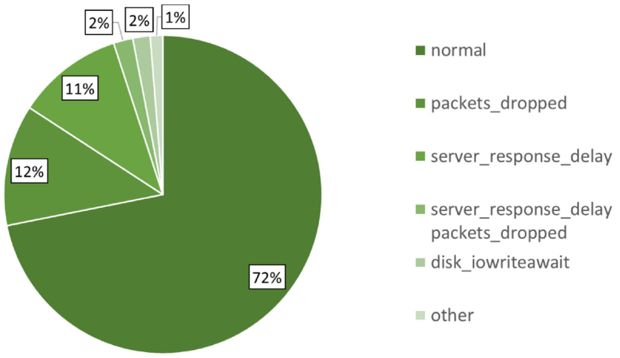


Fig. 3. Distribution of experimental data by the number of simultaneously violated service level indicators

As can be seen from the diagram, 72% of the experimental data is occupied by the CN state of the normal functioning of the CN, which gives rise to the problem of class imbalance. For the computational experiment, the first 200 thousand records of the initial experimental data were taken [21]. The amount of experimental data was chosen based on the available computing resources.

The specified attributes – *ping_avg, network_outdropped, disk_ioreadmergespersec, server_response_timetotal* – are converted to the corresponding CN states and excluded from further analysis. Thus, the specified secondary attributes CN become latent variables.

The following attributes are chosen as primary attributes for illustrative purposes: *disk_await, disk_writebytes, network_outbytes, network_inbytes, ping_max.*

Since one record can be associated with several CN states at the same time, the method of reducing multi-label class labels to a single-label form was chosen - *Label Powerset* (LP, [28]), generates a new class for each possible combination of labels by unitary coding of the alphabet of all possible combinations of CN states and then solves the multi-label analysis problem as a single-label multi-class analysis problem.

To improve the objectivity of the classification, the accuracy was assessed by cross-validation: the sample was divided into 10 equal parts; alternately one of the parts became a test. The classification efficiency metric is Mean accuracy (which is the standard metric for all algorithms provided by the scikit-learn.org library).

The experiment was carried out with standard hyperparameters set for the default algorithms. Hyperparameter optimization was not performed. For pairs "single-label classification algorithm X – multi-label classification algorithm X", the same hyperparameters were set.

The results of the computational experiment are given in Table 2. The table shows the name of the algorithm, the results for the single-label and multi-label cases. The cell with the highest value of the Mean accuracy metric among all types of classification is highlighted in light color.

Table 2

Comparative analysis of single-label and multi-label classifiers in a computational experiment

Name of the classification algorithm	Mean accuracy metric value for single-label case of multiclass classification	Mean accuracy metric value for multi-label classification case
Tree.DecisionTreeClassifier	0,52	0,75
Tree.ExtraTreeClassifier	0,66	0,69
Ensemble.ExtraTreesClassifier	0,64	0,81
Neighbors.KNeighborsClassifier	0,64	0,91
Ensemble.RandomForestClassifier	0,70	0,13

As can be seen from the table, 80% of single-label classifiers were inferior in classification accuracy according to the Mean accuracy multi-label metric to their counterparts, which may indicate a strong influence of multi-label class labels on the models under consideration. Despite the fact that multi-label plots are only 3% (see Table 2), the gain in accuracy reaches 23% in terms of the Mean accuracy metric for MLL algorithms.

The conducted experiment allows us to form the following conclusions. The LP method used to markup single-label data leads to high classification errors for boosting algorithms during cross-validation.

The data structure of [21] is affected by the multi-label problem much more than can be estimated by the standard frequency check performed in Table 1, 2. One of the possible reasons for such a strong influence is the use of primary attributes as arguments that are not directly related to the classified CN states.

Since the predictive power of frequency testing of the effect of multi-label class label results on the classification results of single-label classifiers is low, further research on this topic is planned. Conducting research in the field of multi-label analysis can lead to an increase in the accuracy of both static and dynamic fault detection in CN and network attacks [29].

Conclusion

The results of the study of estimating the characteristics of CN states of a distributed computer system consisting of six hosts for given indicators of the service level SLO are analyzed.

Class labels (CN states) generated as a result of CN operation, in the general case, are multi-label as a result of the removal and analysis of information on several CN attributes (from several system sensors). The nature of multi-label CN states is different from the nature of multi-label occurrence in the analysis of text corpora or social network data.

Anomalies associated with violation of the established SLO thresholds regularly occur simultaneously for several analyzed attributes. The results of the computational analysis made it possible to judge the nonlinear dependence of the frequency distribution of multi-label class labels on the degree of influence of multi-label on the classification results, which, in turn, directly affects the security of information circulating in the CN.

In connection with the results obtained, if there is a priority in the classification of certain class labels (which is important for information security tasks), multi-label classifiers are proposed for use.

References

[1] A. Kuznetsov, V. Babenko, K. Kuznetsova, S. Kavun, O. Smirnov, O. Nakisko “Malware correlation monitoring in computer networks of promising smart grids”, *Proceedings of the IEEE 6th International Conference on Energy Smart Systems, ESS 2019*, 2019, pp. 347-352. DOI: 10.1109/ESS.2019.8764228

[2] A.S. Bol’shakov, D.I. Rakovskii “An efficient multiple-criteria decision analysismethod in the field of information security”, *Legal Informatics*, 2020, no 4. pp. 55-66. DOI 10.21681/1994-1404-2020-4-55-66.

[3] I.V. Kotenko, S.S. Khmyrov “Analysis of models and techniques used for attribution of cyber security violators in the implementation of targeted attacks”, *Voprosy kiberbezopasnosti*, 2022, vol 50, no 4, pp. 52-79. DOI 10.21681/2311-3456-2022-4-52-79.

[4] D.A. Gaifulina, I.V. Kotenko “Application of deep learning methods in cybersecurity tasks”, *Voprosy kiberbezopasnosti*, 2020, vol 37, no 3. pp. 76-86. DOI 10.21681/2311-3456-2020-03-76-86.

[5] M. Alrammal, M. Naveed, S. Rihawi “Using heuristic approach to build anti-malware”, Proceedings of the ITT 2018 - Information Technology Trends: Emerging Technologies for Artificial Intelligence. 5, *Emerging Technologies for Artificial Intelligence*, 2019, pp. 191-196. DOI: 10.1109/CTIT.2018.8649499.

[6] A.S. Bol'shakov, D.I. Rakovskii “Software for modelling information security threats in information systems”, *Pravovaya informatika*, 2020, no 1, pp. 26—39. DOI: 10.21681/1994-1404-2020-1-26-39. E.Y. Pavlenko, N.V. Gololobov, D.S. Lavrova, A.V. “Kozachok Recognition of cyber threats on the adaptive network topology of large-scale systems based on a recurrent neural network”, *Voprosy kiberbezopasnosti*, 2022, vol. 52, no 6, pp. 93 – 98. DOI:10.21681/2311-3456-2022-6-93-99

[7] K.E. Izrailov, M.V. Buinevich, I.V. Kotenko, V.A. “Desnitsky Assessment and prediction of the complex objectsstate: applicatioin for information security”, *Voprosy kiberbezopasnosti*, 2022, vol 52, no 6, pp. 2 – 21. DOI:10.21681/23113456-6-2022-2-21

[8] O.I. Sheluhin, A.V. Osin, D.I. Rakovsky “New Algorithm for Predicting the States of a Computer Network Using Multivalued Dependencies”, *Automatic Control and Computer Sciences*, 2023, vol. 57, no 1, pp. 48–60. DOI: 10.3103/S0146411623010091

[9] E. Gibaja, S. Ventura “A Tutorial on Multi-Label Learning”, *ACM Computing surveys*, 2015, no 47, pp. 1-40. DOI: 10.1145/2716262

[10] A.C.E.S. Lima, L.N. de Castro “A multi-label, semi-supervised classification approach applied to personality prediction in social media”, *Neural Networks*, 2014, vol. 58, pp. 122-130.

[11] S.N. Karpovich “Multi-Label Classification of Text Documents using Probabilistic Topic Model ml-PLSI”, *Trudy SPIIRAN*, 2016, vol 47, no 4, pp. 92-104 DOI: 10.15622/sp.47.5

[12] I.V. Kotenko, I.B. Saenko, A.A. Branitsky, I.B. Paraschuk, D.A. Gayfulina “Intelligent system of analytical processing of digital network content for its protection from unwanted information”, *Informatics and automation*, 2021, vol. 20, no 4, pp. 755-784

[13] G.G. Kulikov, V.V. Antonov, Antonov D.V. “Analysis of the possibility of analytical knowledge extraction of a formal model of subject domain information system by neural network methods”, *Neurocomputers,* 2013, no 3, pp. 12-16.

[14] M. Azad, M. Moshkov “A Bi-criteria Optimization Model for Adjusting the Decision Tree Parameters”, *Kuwait Journal of Science*, 2022, vol. 49, no 2, pp. 1–14. DOI 10.48129/kjs.10725

[15] A. Niemistö , O. Yli-Harja, I. Shmulevich, V.V. Lukin, A.N. Dolia “Correction of misclassifications using a proximity-based estimation method”, *Eurasip Journal on Applied Signal Processing*, vol. 2004, no 8, pp. 1142-1155. DOI: 10.1155/S1110865704402145

[16] A.S. Markov “Cybersecurity and information security as nomenclature bifurcation scientific specialties (Russian text)”, *Voprosy kiberbezopasnosti*, 2022, vol 47, no 1, pp. 2-9. DOI 10.21681/2311-3456-2022-1-2-9

[17] Lovtsov D. “Principles of ensuring information security in ergasystems”, *Legal Informatics*, 2021, no 1, pp. 36-50. DOI 10.21681/1994-1404-2021-1-36-50

[18] A. S. Bolshakov, R. V. Khusainov, A.V. Osin “Traffic anomaly detection using a neural network to ensure information protection”, *I-methods*, 2021, vol. 13, no 4, pp. 1 – 15.

[19] O.I. Sheluhin, D.I. Rakovskiy “Prediction of the profile functioning of a computer system (network) based on multivalued patterns”, *Voprosy kiberbezopasnosti*, 2022, no 6, pp. 28-45 (in Russian) DOI:10.21681/2311-3456-2022-6-53-70

[20] O.I. Sheluhin, D.I. Rakovsky “Selection of metric and categorical attributes of rare anomalous events in a computer system using data mining methods”, T-Comm. 2021, vol. 15, no. 6, pp. 40-47. (in Russian) DOI: 10.36724/2072-8735-2021-15-6-40-47

[21] B. Raja, K. Ravindranath, B. “Jayanag Monitoring and analysing anomaly activities in a network using packetbeat”, International Journal of Innovative Technology and Exploring Engineering, 2019, Vol. 8, No 6, Pp. 45-49.

[22] I.V. Kotenko, A.A. Kuleshov, I.A. Ushakova “System for collecting, storing and processingsecurity information and events based on elasticstack tools”, Informatics and Automation (SPIIRAS Proceedings), 2017, vol. 54, no 5, pp. 5-34. DOI 10.15622/sp.54.1(in Russian)

[23] V.V. Petrov, K.V. Bryukhanov, E.Y. Avksentieva “Network monitoring: network traffic analysis using ELK”, In Modern Science: actual problems of theory & practice, 2020, no 5, pp. 102-105. DOI 10.37882/2223-2966.2020.05.34. (in Russian)

[24] G. Calderon, G. Del Campo, E. Saavedra, A. Santamaria “Management and Monitoring IoT Networks through an Elastic Stack-based Platform”, Proceedings of 2021 International Conference on Future Internet of Things and Cloud, FiCloud 2021. Virtual, Online, 2021, Pp. 184-191. DOI 10.1109/FiCloud49777.2021.00034.

[25] I.V. Kotenko, A.A. Kuleshov, I.A. Ushakov “Aggregation of elastic stack instruments for collecting, storing and processing of security information and events”, Proceedings of the 2017 IEEE SmartWorld, Ubiquitous Intelligence & Computing, Advanced & Trusted Computed, Scalable Computing & Communications, Cloud & Big Data Computing, Internet of People and Smart City Innovation (SmartWorld/SCALCOM/UIC/ATC/CBDCom/IOP/SCI). California, USA: Institute of Electrical and Electronics Engineers, 2017, pp. 1 – 8. DOI 10.1109/UIC-ATC.2017.8397627.

[26] U. Chaudhuri, S. Dey, B. Banerjee, A. Bhattacharya, M. Datcu “Interband Retrieval and Classification Using the Multilabeled”, Sentinel-2 BigEarthNet Archive. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 2021, vol. 14, pp. 9884-9898. DOI 10.1109/JSTARS.2021.3112209

[27] L. Maltoudoglou, A. Paisios, H. Papadopoulos, L. Lenc, J. Martínek, P. Král “Well-calibrated confidence measures for multi-label text classification with a large number of labels”, Pattern Recognition, 2022, vol. 122, pp. 108271. DOI: 10.1016/j.patcog.2021.108271

[28] O.I. Sheluhin, S.Yu. Rybakov, A.V. Vanyushina “Modified Algorithm for Detecting Network Attacks Using the Fractal Dimension Jump Estimation Method in Online Mode”, Proceedings of Telecommunication Universities, 2022, vol. 8, no 3, pp. 117-126. (in Russian) https://doi.org/10.31854/1813-324X-2022-8-3-117-126

PRINCIPAL COMPONENT ANALYSIS FOR MACHINE LEARNING

Polina Shulpina,
MTUCI, Moscow, Russia
polli-lionet@yandex.ru

V. A. Dokuchaev,
MTUCI, Moscow, Russia
v.a.dokuchaev@mtuci.ru

DOI: 10.36724/2664-066X-2022-8-6-18-24

ABSTRACT

Training a Supervised Machine Learning model involves several stages. In the first stage, the data is passed via model, creating predictions (forecasts). The next stage is to compare these forecasts with factual values (ground truth). The final stage is optimizing the model by minimizing a certain cost function. The model improves that way. Occasionally, an input sample contains many columns. Using each column in a model leads to problems, the curse of dimensionality. At that rate, it is necessary to be selective about functions. We will embrace Principal Component Analysis (PCA), that is one of the main ways to reduce the dimensionality of data, losing the least amount of information.

KEYWORDS: *principal component analysis, PCA, machine learning, deep learning, feature scaling, feature extraction, data preprocessing*

Information about authors:
Polina Shulpina, *Network Information Technologies and Services, MTUCI, Moscow, Russia*
V.A. Dokuchaev, *DSc, Prof., Network Information Technologies and Services, MTUCI, Moscow, Russia*

Introduction

The digital transformation currently underway affects all aspects of human life: economic, political, social, etc. [1-3]. In the conditions of robotization, automation and informatization of technological processes and business, the correctness of decisions made and the reduction of the time required for this is becoming increasingly important. Therefore, the creation and implementation of decision support systems using the AI concept becomes relevant. At the same time, the problem of Big Data arises – the necessity to handle a big quantity of various data in order to make the right decision, taking into account new potential risks [4-7]. One way to meet the challenge is to use Principal Component Analysis.

The principal component analysis is a linear algebraic dimensionality reduction method. First, it is necessary to consider why the specifics of PCA can be beneficial for machine learning (ML) projects. Thereby, the first thing to pay attention to is the "curse of dimensionality" [8] that is particularly powerful with older ML algorithms (Naive Bayes, K-Nearest Neighbour). Let’s discuss the distinction between Feature Extraction (FE) and Feature Selection (FS) on to PCA and what we can do to struggle against dimensionality reduction [9].

When training a Supervised Machine Learning model, we follow a three-stage iterative process, which is depicted in Figure 1.

The first stage in model training is feeding samples to the model. The model has just been initialized. Predictions are generated for every sample and they probably are incoherent.

The next stage is to compare forecasts to ground truth. The simile manufactures loss value or an error that shows how inadequate the model satisfies.

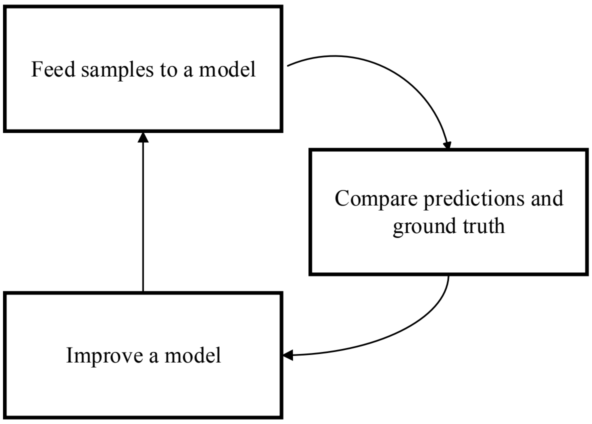


Fig. 1. Three-stage iterative process for model training

The third stage is improving the model. Optimization occurs differently. Gradients are computed using back propagation and then optimizers change the model's internal components. In addition, it is possible to change weights via minimization just a single function.

After optimizing the model, the predictions get a little better. Iterating is needed to be kept until we are contented with the outcome. After this, we finish the learning process.

I. Underfitting and overfitting a model

At the initial steps of the process of learning, a model is probably not suitable to grab samples in a dataset (Fig. 2). The decision is elementary: we should continue learning until we reach the correct conformity for the dataset (Fig. 4). We can't keep training forever. To the contrary, we will face the problem of overfitting. Overfitting is the phenomenon when the quality of the model on the training dataset significantly exceeds the quality of the model on the test set.

As a result, a model acclimatized to a concrete dataset, visible in Figure 3. Both an underfitted and overfitted model cannot generalize well to the test data. Therefore, we should find a balance between these two aspects.

The curse of dimensionality is one of the biggest problems in Machine Learning, which states that the higher the dimensionality, the more sparse the data. In other words, as the number of features grows, the amount of data we need to generalize grows exponentially.

It is worth distinguishing between feature extraction and feature selection. Feature selection is the selection of a sub-set of features from the available ones without transforming the object. Feature extraction is the transformation of an object, i.e. transformation of the selected features into a lower-dimensional space. Let’s discuss them [10].

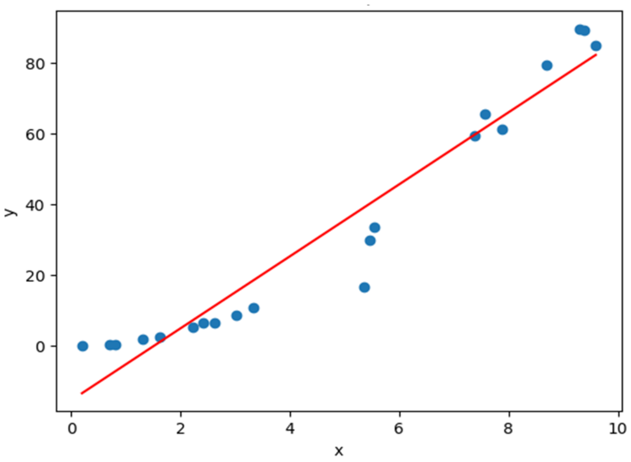


Fig. 2. Model that does not catch a pattern in the dataset

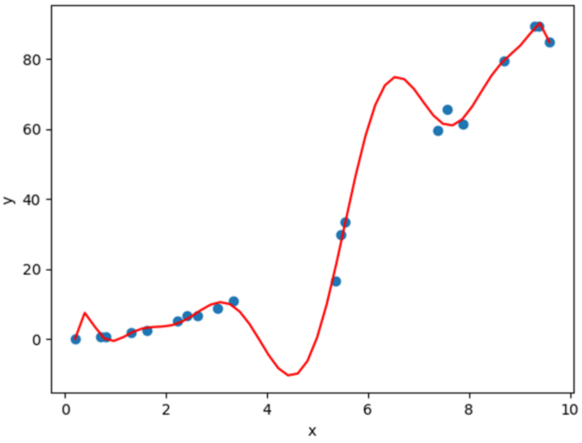


Fig. 3. Model that fits the dataset too well

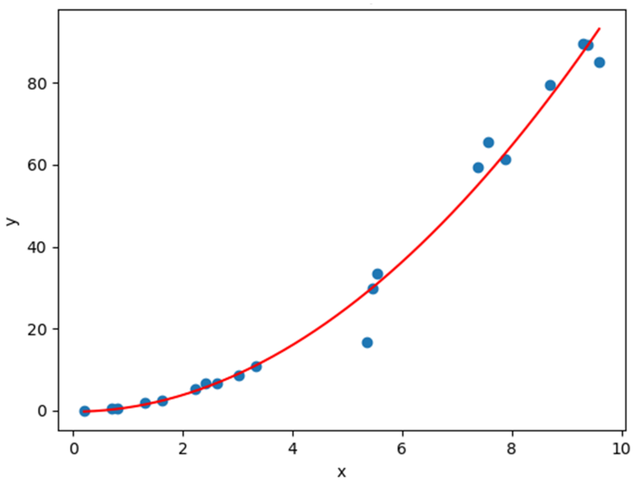


Fig. 4. Model that fits well to describe the dataset

The problem of too many features is known as the Curse of Dimensionality. When the number of features (data dimensionality) grows, it becomes increasingly difficult for the model to learn the correspondence between features and goals. In the case of Dimensionality Reduction, there are two approaches: Feature Selection and Feature Extraction.

- Feature Selection. Each object is characterized by a set of features. An attribute of an object (image, text, sound) can be any quantitative characteristic; the main thing is that this characteristic is unique for this type of objects. The more features, the longer the algorithm works. Feature Selection helps to solve the following problems: simplify the model in order to better understand its operation, reduce training time, avoid the "curse of dimensionality", and reduce overfitting.
- Although feature selection is formally a special case of feature extraction, feature selection uses its own unique algorithms and methods. Feature selection techniques can be divided into three methods: wrapper methods, filter methods, and embedded methods. Feature Selection is an integral portion of the training process, so you need to do Feature Selection independently for the training and test sets. If this is not done and Feature Selection is carried out for the entire database, then the data will inadvertently be distorted, which will lead to overfitting.
- Feature Extraction. Feature Extraction is associated with a decline in the dimension of the feature space. Feature Extraction is a set of methods that map input functions to new output functions. Feature Extraction accelerates the convergence of machine learning algorithms, making them applicable in practice.
- It is reduction the dimension of the feature space by applying several mapping functions. It retrieves the most informatory features based on the selected metrics. In contrast to Feature Selection, Feature Extraction modifies the master features. The main part of Feature Extraction is the mapping function.

II. Introduction to Principal Component Analysis

The principal component method is a way to decrease the dimensionality of the information while losing the least quantity of data. In other words, the sense of this method is that every principal component is connected with a specific ratio of the overall variance (dispersion) of the master dataset (a burden). The dispersion, that is a gauge of the volatility of the data, can represent the rate of information content of the data [9].

The task of reducing the dimensionality of the dataset is to describe the data points using a number smaller than the dimensionality of the space.

The task of PCA: to create a novel feature space of lower dimensionality whose variance between the axes are reallocated to increase the variance for every axe.

There are the following steps of PCA:

1. The overall dispersion of the master feature space is rated. It cannot be made by merely folding the dispersion for every variable, since they are, in most cases, not autonomous. Hence, it is necessary to summarize the reciprocal dispersions of the variables, which are defined of the covariance matrix.

2. The eigenvalues and eigenvectors of the covariance matrix are calculated.

3. Dimensionality reduction is performed. Diagonal principles of the covariance matrix indicate the dispersion under the master system of axes and its eigenvalues under the novel coordinate system. By separating the dispersion connected with every principal component via the amount of the dispersions of each component, we get the ratio of dispersion connected with every component. Then, a great deal of principal components are ejected, the fraction of the remaining components is 80-90%.

There are two methods to decompose the dataset into two vectors. The first method is decomposing the covariance matrix into eigenvectors. The second method is decomposing the covariance matrix into singular values [11].

Consider we create a dataset based on two overlapping patches that are parts of only one dataset. The spread of the dataset is shown in Figure 5.

The dataset mostly spreads out in two directions. These are the directions from the upper right-hand corner to the lower left-hand corner and from the bottom right middle to the top left middle.

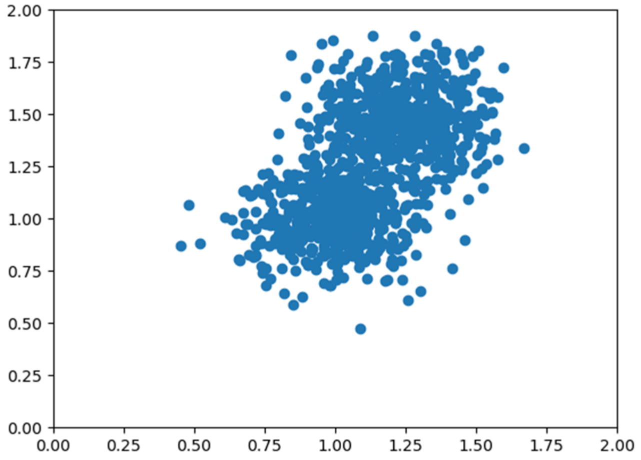


Fig. 5. Spread of the dataset

Now it is possible to depict trends as a pair of two vectors (principal directions of the data). The number of principle directions is equal to the number of measurements. We have two dimensions. The scatter of the dataset as a pair of two vectors is shown in Figure 6.

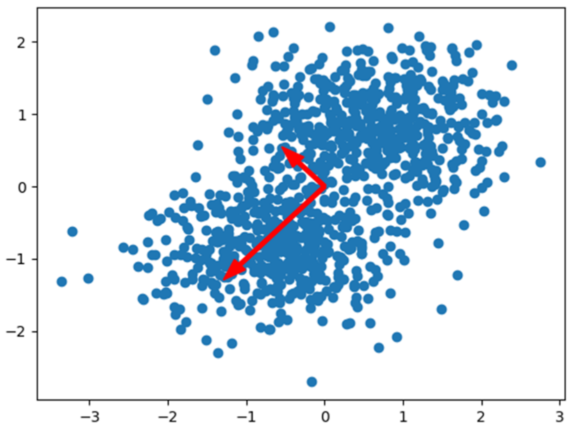


Fig. 6. Scatter expression of the dataset

These vectors are called eigenvectors. Their length is represented by the so-called eigenvalue. The data along new feature axes are explained by eigenvalues, so we can determine two trends and amount of the variance in the dataset [12].

Since the axes and vectors are orthogonal with one another, we can transform the dataset by making the directions of the axes equivalent to the directions of the eigenvectors.

Vector sorting and dimension reduction takes place before the dataset is projected onto the principal components. The eigenvectors show the direction of our projection. The corresponding eigenvalues, in turn, show the important principal direction in interpretation the dispersion of the dataset, so that we can discard insignificant directions.

One of our objectives is to retain the main directions that contribute a lot to the spread in the dataset. We can achieve this by reducing the dimensionality. Then let's proceed to sort the eigenvectors and eigenvalues in descending order of importance.

When performing sorting, make sure that the sorted list with eigenvectors picked out likewise as our sorted list with eigenvalues. The highest eigenvalues should be at the top of the list, as they tell us the largest scatter in the dataset.

Figure 6 shows the scatter expression of the dataset. The eigenvalue of the down- targeted eigenvector transcends the eigenvalue of the up- targeted vector. For our example, the complete deposit of the eigenvectors to the variance is as follows:

$$\left[0.76318124; 0.23681876\right]$$

In the dataset, 76.3% of the changes are explained by the first eigenpair, while the second explains just 23.7%. Together these eigenpairs describe 100% of the scatter. Applying PCA to reduce dimension gives a chance to go ahead only with vectors with the highest contribution [11].

In case it is necessary to reduce the dimensionality to unity, it would be necessary to go further and take for the data projection an eigenvector with a contribution of 0.763. This means a loss of 0.237 data about the scatter. However we would gain a smaller amount of measurements in return.

In order to project our data onto the principal components, we will change the axes by equating them to eigenvectors.

In Figure 7, we have projected the data onto a single eigenvector. After projecting, just the x-axis has values, so the feature space has shrunk to one measurement. Thus, we reduced the dimension by using PCA [11].

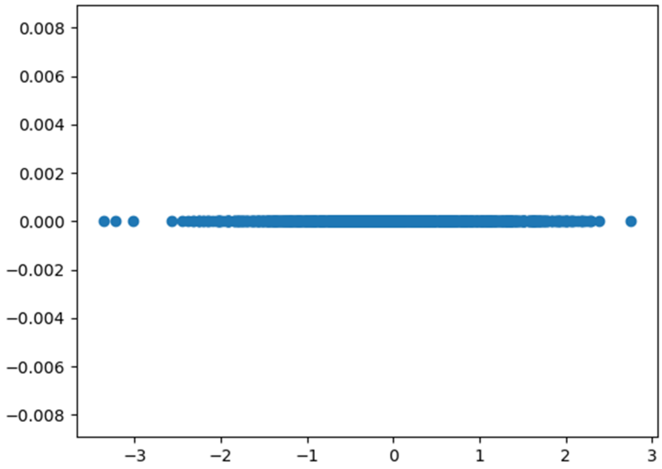


Fig. 7. Projecting a dataset onto an eigenvector

III. EIGENVECTOR DECOMPOSITION OR SINGULAR VALUE DECOMPOSITION

Two methods are used to scatter data using eigenpairs: Eigenvector Decomposition ("EIG") and Singular Value Decomposition ("SVD") [13]. Matrix decomposition is a beneficial implement for diminishing a matrix to its composite portions with an eye to streamline a number of more complicated procedures. The decomposition, which decomposes the matrix into eigenvectors and eigenvalues, is the most popular form of matrix decompositions. A singular decomposition is a decomposition of a real matrix with the aim to reduce it to a canonic form. Singular value decomposition is an easy way for operating with matrices. It displays the geometrical pattern of a matrix and allows depicting the affordable data.

A. Eigenvector Decomposition

If the vector v is an eigenvector of a square matrix A, it must be expressed in the following form: $Av = \lambda v$. At this time, λ is called the eigenvalue corresponding to the eigenvector v, and the group of eigenvectors of the matrix is a group of orthogonal vectors. The eigenvalue decomposition consists in decomposing the matrix into the following form: $A = Q\Sigma Q^{-1}$, where Q is a matrix made up of the eigenvectors of this matrix A, Σ is a diagonal matrix, and each element on the diagonal is an eigenvalue.

The decomposed Σ -matrix is a diagonal matrix. The eigenvalues are ordered from larger to smaller. The eigenvectors corresponding to these eigenvalues describe the direction of change of the matrix – from major to minor changes.

When a matrix is multidimensional, then that matrix is a linear transformation in a multidimensional space. This linear change cannot be represented by images, but it is possible that this transformation also has multiple transformation directions. The first N eigenvectors obtained by value decomposition correspond to the most important N directions of change of this matrix. We can approximate this matrix (transformation) using the first N changing directions.

So, eigenvalue expansion allows you to get eigenvalues and eigenvectors. The eigenvalue indicates how important the feature is, and the eigenvector indicates what the feature is. Each eigenvector can be understood as a linear subspace. You can do a lot with these linear subspaces. However, eigenvalue decomposition also has many limitations: for example, the transformed matrix must be a square matrix.

Let's work with the popular Iris dataset, part of the free scikit-learn machine learning library. This set is often used for classification and clustering tasks. There are four sample features in this dataset: tepal length (0), tepal width (1), petal length (2) and petal width (3).

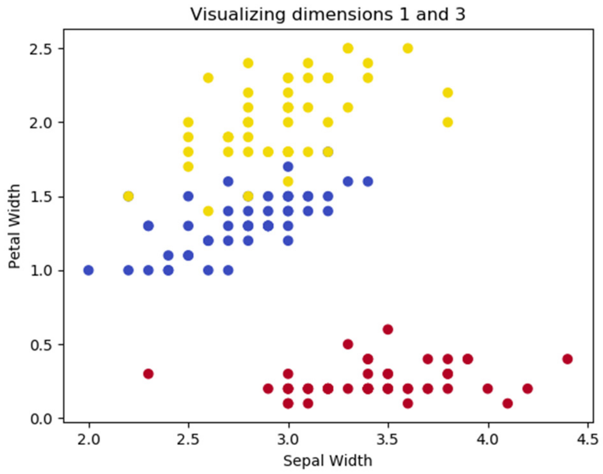


Fig. 8. Visualization of measurements 1 and 3

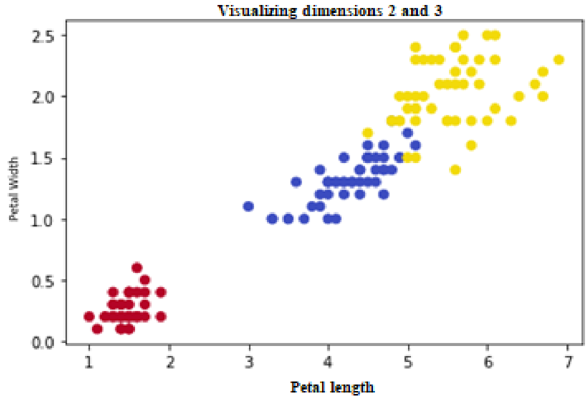


Fig. 9. Visualization of measurements 2 and 3

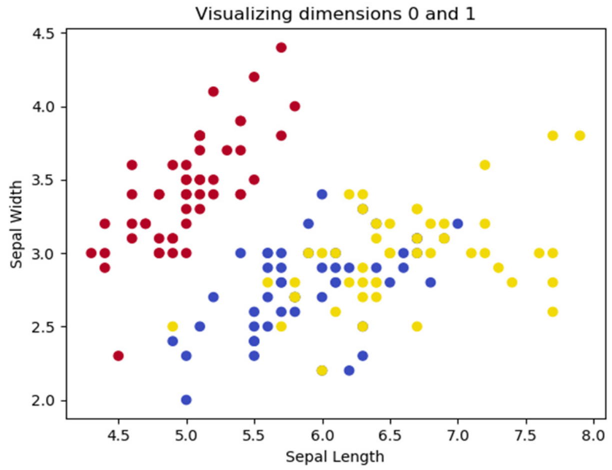


Fig. 10. Visualization of measurements 0 and 1

The figures demonstrate that two iris flowers cannot be linearly separated, however it is possible to separate this group from another iris flower. We have only 150 samples, but our feature space is four-dimensional. This is a case where feature extraction can be useful for training our model [10].

We equate the standard deviation (σ) to one, and the average value of the sequence of numerical data to zero ($\mu = 0.0$), by doing $x = \frac{x-\mu}{\sigma}$ for each dimension. As a result, we have changed the way values are displayed in the model space [14].

The following stage is to estimate the covariance matrix for the dataset. The covariance matrix in probability theory is a matrix made up of the pairwise covariances of the elements of one or two random vectors [8].

A variable. X is a mathematical representation of one measuremen measurement t of the dataset. Supposing that X presents petal width, numbers that present the petal width for one flower, can be also explained by the variable X .

Variable mean. Since the width of the lobe is dim (dimensionality of space points) = 3 in the visualization above, with an average value of 1.1993, it's possible to catch sight of how the numbers upward fit.

Variance (dispersion). Shows the «distribution» of data nearly the variable: $(x-\mu)^2$, where μ is expected value.

Covariance. Describes the direction of the linear dependence between variables:

$Cov(x, y) = (x-\mu_x)(y-\mu_y)$, where μ is expected value.

Covariance matrix for n variables. A covariance matrix for two dimensions X и Y looks as follows:

$$\begin{bmatrix} Cov(X,X) & Cov(X,Y) \\ Cov(Y,X) & Cov(Y,Y) \end{bmatrix}$$

Covariance matrix properties:

$Cov(X,X) = Var(X), Cov(X,Y) = Cov(Y,X)$.

So, our covariance matrix can be written as follows:

$$\begin{bmatrix} Var(X) & Cov(X,Y) \\ Cov(Y,X) & Var(Y) \end{bmatrix}.$$

In the matrix above, we see that the size of the covariance matrix is $n \times n$. It is essentially a symmetric matrix, i.e. a quadrature matrix that is equal to its transposition. The terms that construct a covariance matrix are called variations of a given variable that forms the diagonal of the matrix, or covariances of the two variables that fill the rest of the space. The j variable covariance with the k variable is equal to the covariance of the k variable with the j variable, i.e. « sjk »= « skj » [9].

With EIG-PCA we have the opportunity to expand the covariance matrix into eigenvectors and eigenvalues: $C = VLVT$, where V is the matrix of eigenvectors, is a diagonal matrix with eigenvalues and VT is the transpose of V .

Eigenvector matrix:

$$\begin{bmatrix} 0.52103086 & -0.37921152 & -0.71988993 & 0.25784482 \\ -0.27132907 & -0.92251432 & 0.24581197 & -0.12216523 \\ 0.57953987 & -0.02547068 & 0.14583347 & -0.80138466 \\ 0.56483707 & -0.06721014 & 0.63250894 & 0.52571316 \end{bmatrix}$$

Eigenvalues: 2.91912926, 0.91184362, 0.144265, 0.02476212.

In explaining variance, each principal dimension contributes, as follows: $\left[0.72978232; 0.2279609; 0.03606625; 0.00619053\right]$.

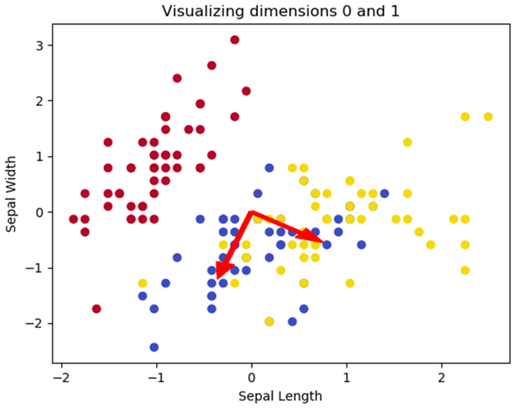


Figure 11 – Visualization of measurements 0 and 1

There are 73% is the contribution of the first principal dimension, 23% is the contribution of the second principal dimension. By changing the dimensionality to two, we get the dispersion description equal to 96% (73% + 23%).

Eigenpairs are sorted by eigenvalues: eigenvalues and corresponding eigenvectors must be sorted in descending order.

Thus, we have sorted eigenpairs, as we can see from the eigenvalues: 2.919129264835876, 0.9118436180017795, 0.14426499504958146, 0.024762122112763244.

The projection matrix projects a vector of observed values onto a vector of fitted values. It describes the effect of each observed value on each fitted value. Graph of the projected data is shown in Figure 12.

In this way, we have decreased the dimension to two without a big loss of data in the dataset.

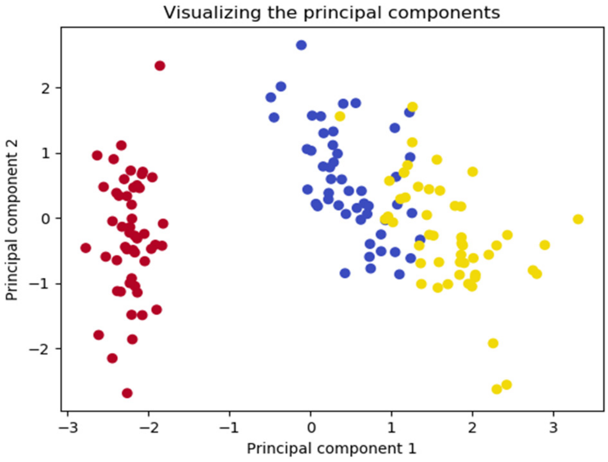


Fig. 12. Visualization of the principal components

B. Performing SVD on a data matrix

A convenient tool for decomposing a covariance matrix into eigenvectors and eigenvalues is the singular expansion. A singular decomposition is a representation of a matrix as a product of three matrices of a special form. Let M be a matrix of size $m \times n$, then it is represented by SVD as follows: $M = ULV^*$, where L is a matrix of size $m \times n$ with singular numbers on the main diagonal and the rest of its elements are zero, $V(n \times n)$ and $U(m \times m)$ are right and left singular matrices, $V^*$is a conjugate-transformed matrix to V (in case of real matrices equivalent to transpose). Omitting mathematical calculations, we note that the matrices U and V^T consist of eigenvectors of matrices $M * M^T$ and $M^T * M$, respectively.

If we zero some singular number in L, its corresponding eigenvector will be excluded from the product ULV^T. Hence, in order to apply the principal component method using SVD, the matrix $X = ULV^T$ must be decomposed. Depending on the problem, the components to get rid of are determined (usually the directions with the smallest variance). The singular numbers in L corresponding to these directions. These directions should be zeroed and a new value $ULV^T = X'$, where X' is the new matrix, which contains the data projected onto several principal components.

Let's look translating SVD outputs to usable vectors and values. The eigenvectors of the covariance matrix are as follows:

$$\begin{bmatrix} 0.52103086 & -0.37921152 & -0.71988993 & 0.25784482 \\ -0.27132907 & -0.92251432 & 0.24581197 & -0.12216523 \\ 0.57953987 & -0.02547068 & 0.14583347 & -0.80138466 \\ 0.56483707 & -0.06721014 & 0.63250894 & 0.52571316 \end{bmatrix}$$

We can collate them to the output of vh:

$$\begin{bmatrix} 0.52106591 & -0.26934744 & 0.5804131 & 0.56485654 \\ -0.37741762 & -0.92329566 & -0.02449161 & -0.06694199 \\ 0.71956635 & -0.24438178 & -0.14212637 & -0.63427274 \\ 0.26128628 & -0.12350962 & -0.80144925 & 0.52359713 \end{bmatrix}$$

As a result, with the exception of the sign, the columns of vh are equivalent to rows of the EIG-based eigenvectors.

Here we can also simply choose n components. As in the PCA-EIG script, we take $n = 2$ and consequently decrease the dimensionality from 4 to 2.

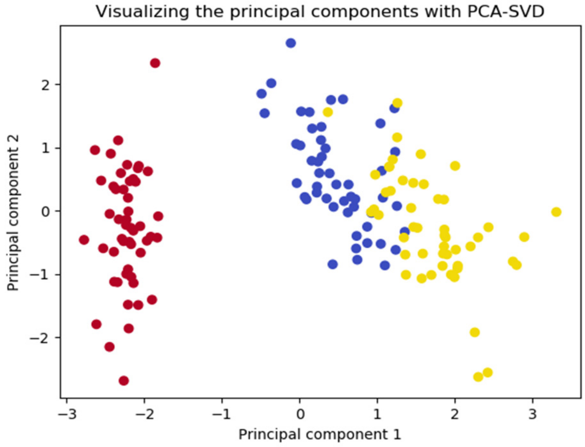


Fig. 13. Visualization of principal components using PCA-SVD

Now we can easily build a projection matrix like the one we did with PCA-EIG, project our data onto the main components, and plot the projection graph.

The final result is shown in Figure 13. It is identical to the result we obtained with the PCA-EIG approach.

Conclusion

To summarize, we reiterate that the use of PCA gives the opportunity to reconstruct the feature space with a lower dimensionality with a minimum of data loss. In this article, we have calculated the principal components and projected the dataset onto these components.

Eigenvector Decomposition (EIG) and Singular Value Decomposition (SVD) are two methods for obtaining eigenvectors, which were also discussed in our article. To find principal directions in a dataset, you need to compute eigenvalues, eigenvectors and then sort them to find principal directions in the dataset.

We have considered two approaches to perform the principal component method: PCA-EIG and PCA-SVD. PCA-EIG has been shown to work well with symmetric and quadratic matrices. However, it should not be forgotten that this method tends be numerically volatile. Because of this, PCA-SVD is often used in modern machine learning libraries. Singular value decomposition may be used on a matrix of standardized data to obtain eigenvectors. We ended up with the same final result as using PCA-EIG.

References

[1] V. A. Dokuchaev, “Digital Transformation: New Drivers and New Risks,” *2020 International Conference on Engineering Management of Communication and Technology (EMCTECH)*, 2020, pp. 1-7, doi: 10.1109/EMCTECH49634.2020.9261544.

[2] V.A. Dokuchaev, A.A. Kalfa. V.V. Maklachkova (2020). Architecture of Data Centers. Moscow: Hot Line-Telecom. 240 p. ISBN 978-5-9912-0849-9.

[3] V. A. Dokuchaev, V. V. Maklachkova, V. Yu. Statev, “Data subject as augmented reality,” *SYNCHROINFO JOURNAL*, vol.6, no.1, 2020, pp.11-15, doi: 10.36724/2664-066X-2020-6-1-11-15.

[4] S.V. Pavlov, V.A. Dokuchaev, V.V. Maklachkova, and S.S. Mytenkov, “Features of supporting decision making in modern enterprise infocommunication systems,” *T-Comm*, vol. 13, no. 3, pp. 71-74, 2019, doi: 10.24411/2072-8735-2018-10252.

[5] V.Yu. Statev, V.A. Dokuchaev, V.V Maklachkova, “Information security in the big data space”. *T-Comm*, vol. 16, no.4, 2022, pp. 21-28. (in Russian).

[6] V. A. Dokuchaev, E. V. Gorban and V. V. Maklachkova, "The System of Indicators for Risk Assessment in High-Loaded Infocommunication Systems," *2019 Systems of Signals Generating and Processing in the Field of on Board Communications*, Moscow, Russia, 2019, pp. 1-4, doi: 10.1109/SOSG.2019.8706726.

[7] S. V. Pavlov, V. A. Dokuchaev, V. V. Maklachkova, S. S. Mytenkov, “Features of supporting decision making in modern enterprise infocommunication systems”, *T-Comm*, vol. 13, no.3, 2019, pp. 71-74.

[8] X. Zhu, H. Dong, P. S. Rossi, M. Landro, «Feature Selection based on Principal Component Analysis for Underwater Source Localization by Deep Learning», Department of Electronic Systems, Norwegian University of Science and Technology, 2020, pp. 1-15, doi: 10.3390/rs13081486.

[9] Relationship between SVD and PCA. How to use SVD to perform PCA, StackExchange, Available: https://stats.stackexchange.com/questions/134282/relationship-between-svd-and-pca-how-to-use-svd-to-perform-pca.

[10] V.M. Efimov, K.V. Efimov, V.Y. Kovaleva, «Principal component method and its generalizations for sequences of any type», *Vavilovsky Journal of Genetics and Selection*, 2019, pp. 1032-1036, doi: 10.18699/VJ19.584.

[11] S. Raschka, «Principal Component Analysis», Available: https://sebastianraschka.com/Articles/2015_pca_in_3_steps.html

[12] J. V. Lambers, «PCA Mathematical Fundamentals, Available: https://www.math.usm.edu/lambers/cos702/cos702_files/docs/PCA.pdf.

[13] N. Cristianini, J. Shawe-Taylor, «An Introduction to Support Vector Machines and other kernel-based learning methods», *Cambridge University Press*, 2000, pp. 687-689, doi: 10.1017/CBO9780511801389.

[14] The Complete Guide to Principal Component Analysis - PCA in Machine Learning, machinelearningmastery, Available: https://en.wikipedia.org/wiki/Covariance_matrix.

[15] Understanding the output of SVD when used for PCA, StackExchange, Available: https://stats.stackexchange.com/questions/96482/understanding-the-output-of-svd-when-used-for-pca.

[16] Why does Andrew Ng prefer to use SVD and not EIG of covariance matrix to do PCA, StackExchange, Available: https://stats.stackexchange.com/questions/314046/why-does-andrew-ng-prefer-to-use-svd-and-not-eig-of-covariance-matrix-to-do-pca.

DIGITAL MANUFACTURING TRANSFORMATION
(INNOVATIVE BUSINESS OPPORTUNITIES FOR MANUFACTURERS)

Graham Immerman, David Westrom,
MachineMetrics, Northampton, Massachusetts, United States

DOI: 10.36724/2664-066X-2022-8-6-25-32

ABSTRACT

The digital thread represents the origination, flow, and consumption of data through machines, people, and systems across the enterprise of a manufacturer. In this paper, we explore unique manufacturing use cases and business opportunities derived from, and driven by, the digital thread and enabling technologies, with machine asset data as the cornerstone.The digital thread connects the various processes, entities, actions, and decisions across a company's supply chain through data and data relationships, helping to ensure an organization runs efficiently and in unison. By stitching together critical product information with digital assets across the full product lifecycle, a digital thread empowers manufacturing companies to drive continuous improvement and innovative new business processes and models. The benefits of data-driven manufacturing are far too significant to ignore and will enable many to deliver competitive advantages in an ever-competitive landscape. Digital transformation is about changing business models and about companies not just taking advantage of the huge opportunities created by these latest technologies but also preparing for their constant evolution. Amid the hype surrounding Industry 4.0, IIOT, and digital transformation, the introduction of Industry 4.0 has caused a bit of a culture shock for manufacturers. To build a roadmap to digital transformation, most companies are looking into the future, attempting to visualize where they want or need to be in twenty years, and planning backwards. However, a great deal of hesitancy exists for many manufacturers to embrace the technology and modernization that solves these new challenges.

KEYWORDS: *Industry 4.0, IIOT, Manufacturing's industrial iot platform.*

Information about authors:

***Graham Immerman**, responsible for global marketing at MachineMetrics. Graham has spent the majority of his career working at global marketing firms to craft successful digital transformation strategies for brands like Adidas, H&RBlock, and Starbucks. He has quickly become a thought-leader on IIoT technology for the manufacturing industry.*

***David Westrom,** responsible for global business development at MachineMetrics. Dave has spent much of his career in executive team roles at innovative Industrial Internet of Things (IIoT) companies. He has led business development organizations and driven strategy at three IIoT start-ups that experienced successful exits, including most recently ThingWorx (acquired by PTC) and Lighthammer (acquired by SAP).*

Introduction

Amid the hype surrounding Industry 4.0, IIOT, and digital transformation, the introduction of Industry 4.0 has caused a bit of a culture shock for manufacturers. The benefits of data-driven manufacturing are far too significant to ignore and will enable many to deliver competitive advantages in an ever-competitive landscape. Digital transformation is about changing business models and about companies not just taking advantage of the huge opportunities created by these latest technologies but also preparing for their constant evolution [1].

These new models for technology-enabled manufacturing have already moved into the implementation phase by many of the world's top manufacturers. However, a great deal of hesitancy exists for many manufacturers to embrace the technology and modernization that solves these new challenges. This hesitancy is a product of a few specific factors:

- **Lack of clear vision and strategy**

Roughly 50% of US companies admit to not having a systematic roadmap or toolbox for easy rollout of digital manufacturing solutions. Because no standard roadmap for digital manufacturing exists, companies are often uncertain around where to start and what foundational capabilities are required to succeed.

- **Lack of competent tech partners**

15% of all US companies identify lack of knowledge about suitable providers as their biggest obstacle. Business leaders need to understand which technology solutions address their core business problems as well as the right criteria for evaluating solution providers.

- **Difficulty managing and attracting digital talent**

21% of all US companies are facing a talent war as their biggest obstacle in transformation – companies need to build capabilities in-house in order to implement new strategies and tactics; experiential learning is the most effective way to build capabilities quickly.

Innovative business opportunities for manufacturers

Despite producing the most data, manufacturing is ranked last in digital transformation efforts. Compared to all other global industries, manufacturing is still caught in a state of reactivity. While some analytics companies have attempted to develop various solutions to address this problem, it has yet to be truly solved. There is no silver bullet. There is, however, a starting point.

Industry 4.0 help companies overcome the challenges along the digital transformation journey and to advance them forward from reactivity, to proactivity, and to predictivity. We are dedicated to empowering our customers to not just employ the latest technology but to achieve success along their individual journeys. How do we do this where so many other companies have failed? It's all about knowing where you stand and planning for the road ahead.

Most manufacturers are caught in a state of reactivity. Despite producing the most data, manufacturing is ranked last in digital transformation efforts. Compared to all other global industries, manufacturing is still caught in a state of reactivity. While some analytics companies have attempted to develop various solutions to address this problem, it has yet to be truly solved. There is, however, a starting point.

To build a roadmap to digital transformation, most companies are looking into the future, attempting to visualize where they want or need to be in twenty years, and planning back- wards. We often talk to companies who have predictive and preventative aspirations but who still don't have machines networked, the necessary IT infrastructure to capture and aggregate machine data, or the internal organizational resources required to decipher the data and implement continuous process changes.

For many however, a more proactive approach to planning would be to accept that "You can't know where you're going without knowing where you are now."

It's all about preparation

Here are a few areas of focus for any company to consider when building out their roadmap:

- *Organization*

To understand what you are solving for, it's essential firstly to be aware of what the problems are, and then to become capable of not just solving those problems but to ready ourselves for the greater problems in the journey ahead. Transformation requires buy-in at all levels, from the front office and on the shop floor, but it also requires internal leadership. It's critical for manufacturers to recognize the important role organizational attributes play in long-term project success and begin discussions about how the odds of project success can be increased by evaluating organizational gaps. Ask yourself: Where do we stand now? Does your team have the right people in place to implement new technology? Are there project leaders capable of owning this project?

- *Communication*

The information we need is available, but the hard part is actually applying it. Avoiding an "us vs. them" mentality is critical in this transition stage. It is vital to build trust between everyone involved in the manufacturing process so problems can be quickly identified, and new solutions can be effectively implemented as a team. Don't let a lack of communication stand in the way of change. Ask yourself: Does your team have an environment capable of communication and applying process changes not just from the top down but from the bottom up?

- *Waste reduction*

Before you embark on our digital transformation journey, it's important to get as lean as possible with your current capabilities. The goal of lean manufacturing is continuous improvement of production processes, while eliminating waste and cutting costs. However, setting the stage for a lean process is just the first step; implementing a system that allows you to maximize your manufacturing productivity results will take your lean model to the next level. Ask yourself: Are we lean? Do we know what our top areas of waste are? Have we embraced lean manufacturing principles?

- *Key performance indicators*

Depending on the systems and processes you have in place on your factory floor, you may face one of two problems; either you don’t know which key performance indicators (KPIs) you should track to enable you to improve your factory performance, or you are unable to collect sufficient data to accurately measure the KPIs you want to track. Having specific KPI’s will allow you to assess, analyze and track our manufacturing processes, as well as to evaluate success in relation to goals and objectives. Ask yourself: What are our key performance indicators that we want to measure as a benchmark for our improvement? Do we have any information now that we can use for this bench- marking? Some of our top suggestions to get started? OEE, Machine Utilization, Set Up Time, Cycle Time, and Scrap Rate.

- *Tools*

Digital manufacturing will transform every link in the manufacturing value chain, from research and development, supply chain, and factory operations to marketing, sales, and service. Having tools to measure your efforts, for designers, managers, workers, consumers, and physical industrial assets will unlock enormous value and change the manufacturing landscape forever.

Of course, every company will need tools to help them optimize their capabilities, but for this job some tools will make more sense than others. Your KPI’s will help you assess which tool will allow you to capture the information you are looking for that best fit your company’s needs. Ask yourself: What tools do we want to use to measure our efforts? What tools are we already using that we can leverage now?

- *Digital connectivity*

The cloud can be your best friend, and with security being better than most on-site solutions systems, the benefits are tremendous. Increasingly more companies are developing or moving their workloads to the cloud by the day, aiming to migrate everything onto the cloud over the next few years. This digitization of data will enable you to deliver competitive advantages in an ever-competitive landscape. Networking your machines and ensuring that all production data can be captured is one of the most essential capabilities for real- time analytics. Ask yourself: Are you ready to digitize our assets? Do you have the technical assets in place to capture and store the data?

Once you’ve completed your capabilities reality check, it's time to begin building your roadmap. Using our areas of focus, your roadmap should actually be quite logical at its core.

STEP 1: get capable

Let’s become as capable as we can and have all our ducks in a row to ready ourselves for the greater journey ahead.

STEP 2: digitize

Once we’ve optimized capability, it’s time to digi- tize our assets, visualize our manufacturing data in real-time, and measure the success of our KPI’s using our tools

STEP 3: analyze

We can then advance our use of this data to begin applying predictive and preventative models to our processes with the hopes of furthering our optimi- zation efforts.

STEP 4: virtualize

We can then virtualize these efforts into an inte- grated manufacturing system framework to sup- port the interoperability between our digital fac- tory tools to solve any real time problems as they arise

STEP 5: automate

Link design, engineering, manufacturing, supply chain, distribution and services into one intelligent (smart) automated system that can be used to self-improve both products and processes within the system.

It's time to shift the focus from the future to the now. Factory Floor Monitoring & Analytics. The problem:

– Poor production visibility

– La k of communication

– Shop-floor data isolated in silos

– Underutilized equipment

– Process nefficiency

The challenge:

– Roughly 50% of US companies admit they lack a systematic roadmap to digital manufacturing solutions and automation.

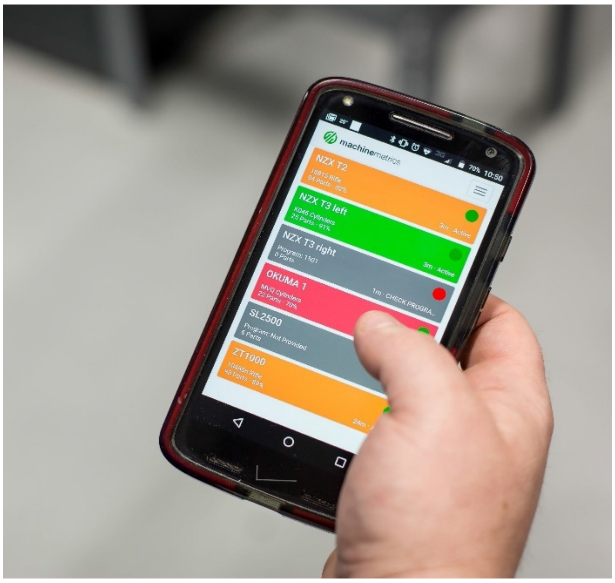
– Over 90% of companies have yet to attempt to integrate solutions.

We gathered feedback from 100+ manufacturers on why current solutions did not fit their needs – and how they could be designed better.

Despite producing the greatest amount of data, manufacturing is the furthest behind any other global industry in their digital transformation efforts. With simple "self-install" IIoT connectivity, it is predictive analytics and machine learning platform allows manufacturers to harness, structure, and take action on this data, driving manufacturing efficiency by more than 20% on average for customers.

Machine learning platform

Fully automated machine monitoring solution provides visualizations of real-time manufacturing production data, notifications, as well as historical analytics, allowing factory workers to make faster, smarter, more confident decisions based on real-time data.



Touchscreen interface allows for operators to add human-context to machine data with touch screen tablet interfaces mounted right at the machine tool. Having a touchscreen at each machine with an intuitive interface that asks the operator to categorize downtime as it's happening allows this information to be made available in real-time to managers in downtime pareto charts. Operators can also reject a part using the Operator View to manage and record quality data. Quality managers are able to view the quality pareto in real-time and when there are new rejects, head to the machine in question, re-inspect the parts and re-allocate as necessary.

From the performance dashboard to the operator workcenter to our reporting features, the interface is designed to be user-friendly for operators, managers, and upper-management. Our customers agree is that our aesthetic is both visually appealing and intuitive. Because of the platform's simplicity, minimal training is required to get your team using the system and confident interfacing with the product.

MachineMetrics is incredibly easy to integrate and requires far less time for setup than most other machine analytics and monitoring platforms. We allow for the option of self-integration or to work with our on-site integration team. Manufacturers can start collecting data in minutes from networked machines. As a cloud application, there are no servers to manage, and no applications to update. All that's required is available internet and that your machines are on your network and accessible from our gateway.

We provides robust and superior reporting features including better OEE reports, job reports, downtime and quality pareto reports. Realtime OEE is available in various reports including the real-time dashboard, historical reports that can be compared by shift, and when viewing individual job/ part reports. Utilization, TEEP and OOE are also measured and visible in historical reports. Information such as cycle times, performance, number of parts produced, rejects, downtime reasons, and reject reasons can be reported on for each part operation.

This information is presented in eminently understandable form, allowing managers to quickly identify issues that are related to a specific machining operation and help measure the effect of process improvements.

MachineMetrics has the unique capability to connect to other types of software, including the ability to tie into a manufacturer’s production goals / ERP to give real-time feedback on a job's performance, and a comparison with previous job runs. To date we have released deep integrations with Epicor and Infor Visual and our open API will allow to directly integrate with any ERP. By partnering with other best in class manufacturing software, we provide seamless connectivity of information silos via our digital threading.

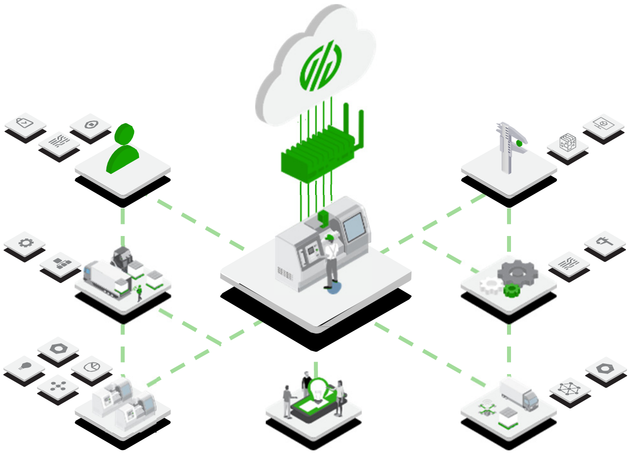
We have a team of developers that are adding new features weekly including ideas brought to us by our current customers. We understand that every manufacturer is unique and we don't want to shoe- horn them into a profile to sell them our software. Rather, we offer a set of tools to our customers and support them to use the tools for whatever fits them best.

There's minimal IT infrastructure required. Our mobile friendly software allows you to access your data from anywhere with a secure connection from your mobile phone or home PC without IT support or complicated firewalls and VPNs.

We can also roll out new features instantly upon approval, provides Full-Time unlimited support forever, with absolutely no hidden costs and provides all customers with a customer success manager (CSM) that, through routine meetings, helps train their team on our software and meeting their performance goals. Customers often make feature suggestions, and CSMs are dedicated to making those requests a reality.

Digital manufacturing

The digital thread represents the origination, flow, and consumption of data through machines, people, and systems across the enterprise of a manufacturer. In this paper, we explore unique manufacturing use cases and business opportunities derived from, and driven by, the digital thread and enabling technologies, with machine asset data as the cornerstone. The digital thread connects the various processes, entities, actions, and decisions across a company's supply chain through data and data relationships, helping to ensure an organization runs efficiently and in unison. By stitching together critical product information with digital assets across the full product lifecycle, a digital thread empowers manufacturing companies to drive continuous improvement and innovative new business processes and models [2].



Data as a foundation

At the heart of every manufacturing operation is the machine assets and the people that make the products. Machine assets in manufacturing plants produce thousands of data points every second and represent the preeminent component of the digital thread.

The insights and actions driven from the data provides the foundation for manufacturers to grow their business and differentiate themselves competitively. Today, manufacturers find themselves at different stages of a journey to leverage data and the digital thread to optimize and automate their business processes. We engage with many manufacturers who discover early in their journey that data from their plant operations is insufficient and unreliable. For key metrics such as equipment utilization, manufacturers, who do not have automated systems for capturing transformed, or contextualized, data from their machine assets, are often under the mistaken belief they are performing at a satisfactory level.

Our average manufacturing customer starts with a utilization of 28%, significantly lower than what is perceived. Similar unsatisfactory statistics are also found for OEE, downtime, and other key metrics.

The underlying cause of the subpar, yet inflated, perception of performance originates with data that is captured manually. Manual data capture often results in data that is inaccurate, manipulated, or missing. It is a shaky foundation that cracks and buckles when attempts are made to drive continuous improvement and innovation. On the other hand, accurate real-time data automatically captured and transformed from machine assets creates a solid base for capturing insights and driving value.

Performance can be baselined across similar machines, lines, and plants with confidence knowing the underlying data is accurate. Machine asset performance can even be measured and compared to similar assets across a particular industry outside of the company. In the end, there is only one version of the truth and the machine does not lie. With accurate data capture and transformation, combined with visibility and actionability through notification and workflow triggers, a 15 to 20% improvement in utilization performance can be realized in a matter of months.

The operational thread

The most valuable assets of a manufacturing operation are the people who make the products. The machine operators, supervisors, and plant engineers are also significant contributors to the digital thread. Weaving the digital thread from machine assets through the people assets stitches together machine data with operational data, creating unique insights and improvement opportunities.

One such area of opportunity is downtime reduction. By enabling operators to categorize downtime events, manufacturers are provided with a thread that drives downtime reduction. Real-time machine data correlated with annotated downtime events from operators provides the foundation for optimizing processes that reduce downtime, such as machine and tooling setup processes. Additionally, the same data sets can be leveraged to generate algorithms and applications that predict tool wear and failure on critical machine assets.

GETTING OFF THE ISLAND

Data on an island, while not useless, does not deliver on the ultimate potential of the digital thread. Value achievement is accelerated by integrating data found in machine assets and operations with data residing on enterprise systems. Value realized from threading data through an ERP system uncovers multiple opportunities and use cases. For discrete manufacturers, a glaring example of lost opportunity associated with 'island data' lies in the inaccuracy of job standards and cycle times.

High mix discrete product manufacturers base their pricing, and by extension, their profitability, on the time it takes to machine a specific product. Inaccurate or sub-optimal job standards result in lost profits.

In many cases, job standards are entered and stored on an ERP system. How accurate are they? Are they updated? How are they calculated? Who entered the data? For many manufacturers, the subject of job standard accuracy leads to an ongoing debate without any clear answers. By capturing and tracking real-time machine and operational data, actual cycle times can be compared to those found in an ERP system, continuously updated, automated, and optimized.

By stitching together these key pieces of the digital thread, companies can optimize cycle times, reliably report on production performance indicators to their customers, and increase profitability.

Maintaining assets

The digital thread weaves its way through many facets of manufacturing operations and is key to maintaining critical manufacturing assets- both machine and people. The maintenance of critical machine assets in manufacturing plants is often accomplished by leveraging a Computerized Maintenance Management System (CMMS). One goal of a CMMS is to optimize the maintenance schedules of machine assets, minimizing the amount of scheduled machine downtime.

Achieving this, however, is difficult without threading, into the CMMS, real-time machine asset data that accounts for the actual operation of the asset itself. How do you optimize a maintenance schedule if you have no idea how long the machine has been running? Would knowing the load over time on the machine also be helpful in determining when a machine asset should be maintained? Threading machine asset data and insights into the CMMS can transform a maintenance strategy from reactive and calendar based to usage or condition based. Without a digital thread originating at the machine asset, the value a manufacturer can expect to achieve from a CMMS investment is significantly diminished.

And what about maintaining a manufacturer’s most valuable asset- its people? Forward thinking companies are creatively weaving the digital thread into their Human Resource systems to enable innovative new processes. One example is generating performance reviews for plant personnel based on measured performance of the machine assets they are responsible for operating and maintaining. The same data can also be leveraged to capture best practices and improve operator and supervisor training programs.

Beyond a manufacturing plant

Extending the digital thread beyond the walls of a single manufacturing facility further increases opportunities to innovate and drive value. For manufacturers with multiple manufacturing facilities, for example, the digital thread enables operation of many facilities virtually as if they were all one. Business Intelligence (BI) systems enable manufacturers to create dashboards and reports providing real-time visibility of key performance indicators across multiple plant sites. This enables comparative measure of performance across manufacturing plant machines, lines, cells, operators, and locations.

By having visibility into available capacity, manufacturers can make better decisions on where to manufacture their products, how to price them, and the timing of delivery. By weaving the digital thread through their planning, forecasting, and financial systems, manufacturers can also make more informed decisions on when plant capacity may need to be increased or decreased. In the past, decisions on capital purchases, specifically the purchase of new machine assets, was often made in a vacuum without the data required to justify the purchase.

Today, the digital thread ties together all required data and insights from multiple systems to ensure these critical financial decisions are data driven and justified.

The digital thread drives a similar set of business cases for original equipment manufacturers (OEMs) and their component part suppliers. Traditionally, suppliers manually provided historical reports to document their ability to meet OEM requirements for a specific part or product. These reports were also part of a process to compete for, and obtain, the business from the OEM initially.

Today, static reports with dated information, are a thing of the past. OEMs can treat their suppliers as virtual extensions of their own business through the digital thread. By insisting on automated machine data capture and performance insights from operations, along with BI and other system technologies, OEMs can gain a real-time view of the performance of their suppliers and the status of their component products. This also results in a data driven approach for evaluating and selecting suppliers. For the suppliers, the enabling technology and systems that power the digital thread, will be a requirement to stay in business.

The product thread

Products are designed, developed, manufactured, sold, shipped, and serviced. The digital thread ties together the various life cycle stages, creating opportunities to innovate and optimize the processes and systems in each functional silo. When manufactured products are developed, or new features are added to existing products, it is often difficult to determine how effectively the products or features are being used once the product reaches the customer. The digital thread has created new opportunities to ‘close the loop’ between design and development, and the actual use of the product by the customer.

Advanced sensors built into manufactured products collect product usage data and transmit that data back to the manufacturer, allowing the manufacturer to obtain real-time feedback on the performance of the product. Stitching customer usage data into CAD/CAM and Product Lifecycle Management (PLM) systems provides the insights required to optimize the design and development of new products and incremental product features and enhancements.

For a manufacturer, the digital thread also closes the loop between product design and the manufacturing process. Does the product, when manufactured, meet design requirements? If not, what adjustments need to be made to the manufacturing process? For discrete manufacturers, this often requires capturing specific data from machine control systems, machine tools, quality systems, and other systems, and weaving the data into design and 3D simulation systems.

The ability to leverage the digital thread to create 'digital twins' represents an advancement where objects, processes, and systems can be modeled and represented virtually. The virtual representation of a system, driven by data from the digital thread, can then be continuously optimized to meet the goals and objectives of the manufacturer.

Machine as a product

At MachineMetrics, we work closely with the manufacturers who manufacture the machinesmachine builders. The machine builders are focused on ensuring their customer, the manufacturer, receives outstanding service and support. The digital thread facilitates this goal by enabling diagnostic data, related to the health and performance of the machine, to be accessed remotely by the machine builder.

By extending the digital thread to the machine service provider, the machine builder or distributor can remotely diagnose health related machine issues, potentially resolve the issue with the customer, or determine if a technician needs to be dispatched to the plant. If a service call at the plant is required, the technician can diagnose and identify issues ahead of time, ensure the right tools and spare parts are available. The result is a higher quality, more efficient service.

Some machine builders, either directly or through partners, provide turnkey systems that include performance guarantees around the operation of a machine, cell, or complete production line. Having remote access to key performance data allows the machine builder to monitor performance and ensure performance guarantees are being met. The machine building business is highly competitive. Machine builders and their distributors are often required to prove their capabilities in competitive run offs with other builders.

They are benchmarked and measured around a variety of factors ranging from speed, quality, and versatility. The ability to prove out any claims through data provides a competitive advantage. The ability to leverage the digital thread to document benefits and competitive advantages while the machine asset is in production on site creates an even more distinct advantage. This advantage is magnified when conveyed in the context of a manufacturer's specific needs and requirements-capacity constraints, machine asset bottlenecks, performance and benefits relative to older equipment, etc. Tying this machine data thread to a Customer Relationship Management (CRM) system and targeted customer data can provide a potent weapon for a machine builder or distributor working to drive sales growth.

Digital thread as a service

The digital thread is accelerating the creation of new business models that promise to disrupt just about every industry. For manufacturers, whether it is selling complementary products or services around the manufactured product or providing the product as a service, the digital thread is creating opportunity for some while eliminating the need for others.

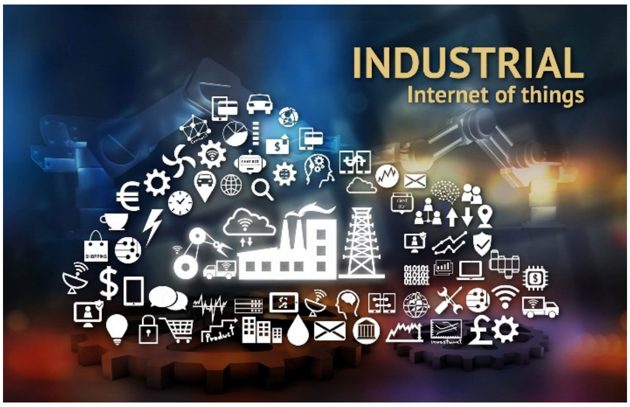
Participants in the manufacturing ecosystem need to understand the potential ramifications, both positive and negative, on their business going forward. One example of business model innovation brings us back to the machine builder. How does the manufacturing industry evolve if machine builders stop selling their machines and begin providing them as a service? In other words, the manufacturer would not buy or own the machine asset, they would purchase the right to run the machine for a specified time or to make a specified number of parts or products.

They would essentially be purchasing a machining service. That service would include training, support, maintenance, and servicing of the machine all included in the service fee. No hidden fees provided the machine is operated within specifications. But how do you determine if the machine is being operated properly? What if the machine is operated outside of the guidelines and parameters for performance and safety? With the digital thread, a machine builder would have remote monitoring capabilities, access to all critical operational data, and the ability to automatically trigger alerts, notifications, and work processes if specific operating parameters were exceeded.

What other services could the machine builder potentially offer under this arrangement? Perhaps a service to manage the tooling supply and processes for the machine equipment. Augmenting this with the ability to predict tool failure and guarantee uptimes would drive tremendous value for a manufacturer. What about managing spare parts inventory? And for those companies who provide specialized machinery, or turnkey machine cells and lines, this could extend to the complete manufacture of certain products as a service.

Conclusion

In this disruptive model, who takes the lead? The machine builder seems like the obvious choice with built in advantages and much to gain. But perhaps it is a different group – maybe insurance companies or the machine distributors? How about the control system suppliers? Who are the partners and who are the competitors? And who gets left out? Does the machine builder still need distributors?



What about many of the ancillary services that in the past were provided by others? And how might this impact the business of the manufacturer? This could lead to many functions of the manufacturer being outsourced to other entities, resulting in smaller, leaner, manufacturing companies.

While this entire discussion may sound hypothetical, I can assure you it is not. There are companies who are moving to some variant of this model as we speak. The digital thread and the technologies that enable and support both the thread and the business model exist today. First mover advantage is already underway.

The Industrial Internet of Things is fundamentally changing the entire economic model of supplier-consumer interaction. This allows:

– automate the process of monitoring and managing the life cycle of equipment;

– organize effective self-optimizing chains from enterprises-suppliers to companies-end consumers;

– ove to "sharing economy" models and much more.

In the most advanced cases, the Industrial Internet of Things makes it possible not only to improve the quality of technical support for equipment using advanced telemetry tools, but also to ensure the transition to a new business model for its operation, when the equipment is paid for by the customer upon the fact of using its functions.

The introduction of network interaction between machines, equipment, buildings and information systems, the ability to monitor and analyze the environment, the production process and its own state in real time, the transfer of control and decision-making functions to intelligent systems lead to a change in the "paradigm" of technological development, also called "fourth industrial revolution".

References

[1] Graham Immerman. Digital Manufacturing Transformation. Roadmap. machinemetrics.com.

[2] David Westrom, Graham Immerman. Weaving the manufacturing digital thread. machinemetrics.com.

[3] Abdel-Basset M., Moustafa N., Hawash H. Deep Learning Approaches for Security Threats in IoT Environments. Piscataway: Wiley-IEEE Press, 2022. 387 p.

[4] Bell Charles. Beginning IoT Projects: Breadboard-less Electronic Projects. Apress, 2021. 875 p. ISBN 978-1-4842-7233-6.

[5] Chaudhuri Abhik. Internet of Things, for Things, and by Things MOBI CRC Press, 2019. 303 p. ISBN: 978-1-315-20064-4

[6] Dagnino A. Data Analytics in the Era of the Industrial Internet of Things. Springer, 2021. 150 p. ISBN 13 9783030631390

[7] Tomar Pradeep. Integration and Implementation of the Internet of Things Through Cloud Computing. IGI Global, 2021. 388 p. ISBN 978-1799869818

ADVISORY BOARD ANNOUNCED FOR GEO WEEK 2023

Geo Week Advisory Board set to help craft programming, recommend speakers, and deliver critical insights to geospatial and built world professionals

Organizers of Geo Week, the premier event that champions the coming together of geospatial technologies and the built world, have announced an impressive list of influential leaders within the geospatial and built world industries who will be participating on the 2023 event's Advisory Board.

The 2023 event will take place February 13-15, 2023 in Denver, Colorado.

www.geo-week.com

BUILD BACK BETTER WITH BROADBAND

Ki-Hong Park, Mohamed-Slim Alouini,
King Abdullah University of Science and Technology, Thuwal, Saudi Arabia

Yunfei Chen,
University of Warwick, Coventry, England, UK

Edward Asiedu,
University of Ghana, Legon, Ghana

David Botchie, Weifeng Chen,
Brunel University, London, United Kingdom

Shang Gao,
Orebro University School of Business, Orebro, Sweden

Michael Canares,
Step Up Consulting, Bohol, Philippines

Francois van Schalkwyk,
Stellenbosch University, Stellenbosch, South Africa

Wondwossen Mulualem Beyene,
freelance researcher

Abraham Tulu Mekonnen, Samson Alemayehu Mamo,
Hawassa University, Hawassa, Ethiopia

DOI: 10.36724/2664-066X-2022-8-6-33-42

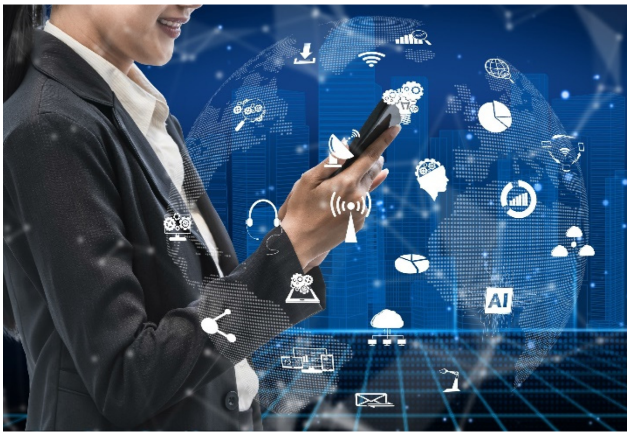
ABSTRACT

The United Nations Secretary-General's Roadmap for Digital Cooperation states that "meaningful participation in today's digital age requires a high-speed broadband connection to the Internet", and that every person should have "safe and affordable access to the Internet by 2030, including meaningful use of digitally enabled services". As part of efforts to achieve these goals, ITU launched the Connect2Recover initiative in September 2020, to help countries transition from responding to the coronavirus disease (COVID-19) pandemic and natural hazards to building back better with broadband. The initiative has the strong support of Australia, Japan, Lithuania and Saudi Arabia. As part of the Connect2Recover initiative, a research competition was launched in July 2021 to identify promising research proposals from across the world to accelerate digital inclusion during recovery from the COVID-19 pandemic. This resulted in the selection of 15 winning research proposals in December 2021. The 15 research teams, which represent 43 universities and institutions from 22 countries, focused on the themes of digital inclusion (in the areas of education, health care, enterprises and job creation, and vulnerable groups), and digital connectivity and resilience. The wealth of knowledge and insights compiled within are based on diverse methodologies, including desktop research, surveys, interviews and focus groups, which covered 17 countries in Africa, the Americas, Arab States and Asia-Pacific. The research showed that, while the use of broadband and digital technologies has been critical for coping with the pandemic, many people faced challenges and barriers in their adoption and use. In schools and universities, teachers and students struggled to get access to online education. During lockdowns, many in rural communities were isolated from health- care providers in cities. In the business sector, the financial needs of micro, small and medium- sized enterprises (MSMEs) were not adequately addressed by financial institutions. The digital needs of vulnerable groups – such as women and girls, ageing populations and persons with disabilities – were also not adequately addressed. Ubiquitous and reliable network infrastructure, as well as affordable and accessible services, are essential to deliver digital solutions such as telemedicine, e-education and e-business services. Policy and regulatory enablers are also critical. Outdated policies or regulations that are not inclusive or do not meet post-pandemic recovery requirements need to be revamped. Digital skills gaps need to be addressed through sustained efforts for institutional and human capacity building. For instance, teachers, healthcare providers and enterprises require digital skills and competencies to thrive and be successful; digital literacy is important for everyone, including vulnerable groups, so that they can fully participate in digital societies and economies. An estimated 2.7 billion people – or one-third of the world's population – remain unconnected to the Internet in 2022. The goal of universal and meaningful connectivity cannot be addressed through improving coverage alone. By leveraging the lessons learned from these 15 published research reports – and working to ensure access, adoption, affordability and resiliency of broadband services – together we can build back better with broadband.

KEYWORDS: *Connect2Recover initiative, broadband and digital technologies, digital skills, COVID-19.*

INTRODUCTION

As part of the Connect2Recover initiative, a research competition was launched in July 2021 to identify promising research proposals from across the world to accelerate digital inclusion during recovery from the COVID-19 pandemic. This resulted in the selection of 15 winning research proposals in December 2021.



The 15 research teams, which represent 43 universities and institutions from 22 countries, focused on the themes of digital inclusion (in the areas of education, health care, enterprises and job creation, and vulnerable groups), and digital connectivity and resilience. The wealth of knowledge and insights compiled within are based on diverse methodologies, including desktop research, surveys, interviews and focus groups, which covered 17 countries in Africa, the Americas, Arab States and Asia-Pacific.

The research showed that, while the use of broadband and digital technologies has been critical for coping with the pandemic, many people faced challenges and barriers in their adoption and use. In schools and universities, teachers and students struggled to get access to online education. During lockdowns, many in rural communities were isolated from health- care providers in cities. In the business sector, the financial needs of micro, small and medium- sized enterprises (MSMEs) were not adequately addressed by financial institutions. The digital needs of vulnerable groups – such as women and girls, ageing populations and persons with disabilities – were also not adequately addressed. Ubiquitous and reliable network infrastructure, as well as affordable and accessible services, are essential to deliver digital solutions such as telemedicine, e-education and e-business services.

Policy and regulatory enablers are also critical. Outdated policies or regulations that are not inclusive or do not meet post-pandemic recovery requirements need to be revamped. Digital skills gaps need to be addressed through sustained efforts for institutional and human capacity building. For instance, teachers, health-care providers and enterprises require digital skills and competencies to thrive and be successful; digital literacy is important for everyone, including vulnerable groups, so that they can fully participate in digital societies and economies.

An estimated 2.7 billion people – or one-third of the world's population – remain unconnected to the Internet in 2022. The goal of universal and meaningful connectivity cannot be addressed through improving coverage alone. By leveraging the lessons learned from these 15 published research reports – and working to ensure access, adoption, affordability and resiliency of broadband services – together we can build back better with broadband.

Research competition: The journey, research stories, lessons learned and the opportunity

The story of the Research Competition began in early 2021 with discussions between ITU and Huawei to design the concept, modalities and desired outcomes within the scope of the Connect2Recover initiative. Connect2Recover aims to build back better with broadband by reinforcing digital infrastructure and digital ecosystems of beneficiary countries, so that they can better leverage information and communication technologies (ICTs) to support COVID-19 pandemic recovery efforts and preparedness for a post-COVID-19 normal, and to remain resilient in hazardous times. In that light, the Connect2Recover Research Competition was designed with the objective to identify promising research proposals that could provide empirically sound and targeted insights, as well as recommendations for fostering digital inclusion during the global COVID-19 recovery. There were 307 research proposals received from 80 countries, demonstrating overwhelming support and interest in this area.

RESEARCH STORIES

The 15 research reports are a compilation of stories and lessons learned from the front line during the COVID-19 pandemic. These include case studies, focus groups and interviews collected from 17 countries in Africa, Asia-Pacific, the Americas and Arab States. The research reports are organized into sections, shown in Figure 2, as follows:

- Digital inclusion in health;
- Digital inclusion in education;
- Digital inclusion for enterprises and jobs;
- Digital inclusion for vulnerable persons; and
- Digital connectivity and resilience.

Section 1 focuses on digital inclusion in health. The first two research reports explore the technologies and solutions required to overcome the challenges in providing health services in rural and remote areas.

The first research report shares a “network-in-a-box” technical solution – that is portable, low-priced and easily deployed. The solution supports broadband and Internet of Things (IoT) health data, and has the potential to revolutionize health services in rural and remote areas.

The second research report describes the successful deployment of communication satellites to deliver telemedicine services in Nigeria.

The third and fourth research reports analyse the deployment of telemedicine. While the third report focuses on the needs of Dominica, the fourth report considers the state of telemedicine and the needs of vulnerable persons in Ghana. They collectively assess the challenges, and propose solutions to address health needs, particularly to those in rural and remote areas.

Section 2 focuses on digital inclusion in education. The first two research reports spotlight digital inclusion in higher education. The first compares three higher education systems in Australia, the Philippines and South Africa, while the second focuses on higher education in Ethiopia and the needs of vulnerable students. The third research report addresses the digital divide in education by considering the cities of Benguerir, Morocco, and Nairobi, Kenya. They collectively highlight the challenges in online learning, and propose measures and policies to address them and enhance digital inclusion in education.

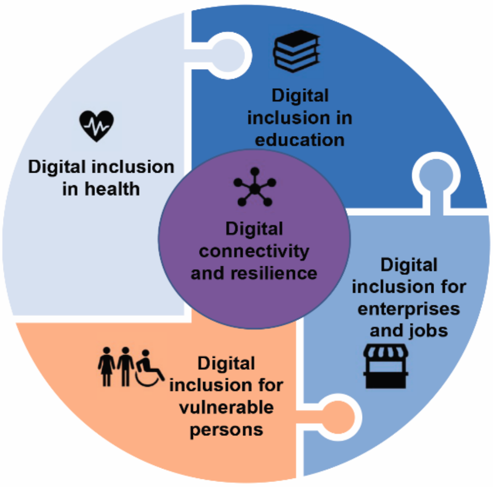
Section 3 focuses on digital inclusion for enterprises and jobs. The first research report focuses on the opportunities for micro-enterprises to leverage digital technology in Ghana; the second report investigates e-business usage and digital financial inclusion of micro, small and medium-sized enterprises (MSMEs) in the Common Market for Eastern and Southern Africa (COMESA) region; and the third proposes a roadmap, with the goal of transforming the smallholder agriculture sector into a digital agriculture ecosystem in Botswana. They collectively demonstrate that digital technologies are necessary to transform smaller enterprises and advance the agricultural sector, and effectively improve jobs and livelihoods.

Section 4 focuses on digital inclusion for vulnerable persons. Both research reports focus on challenges faced in accessing digital services. The first report focuses on the vulnerable persons in Uganda and South Africa, and the second focuses specifically on the challenges faced by older persons in Malaysia. They collectively establish the opportunity available to society by empowering vulnerable persons with digital services.

Section 5 focuses on digital connectivity and resilience. The first two research reports focus on Kenya. The first provides a macro view on infrastructure and policies and their impact on the economy; and the second explores the challenges faced by rural counties in the area of education and health, and the opportunity provided by community networks to address these needs. The third report explores the opportunities at the grass-roots level to leverage community networks in South Africa and India. They collectively demonstrate the need to ensure ubiquitous, reliable and affordable services to support the digital inclusion highlighted in the four earlier sections.

These research stories will be detailed in greater detail in the following sections (Fig. 1).

The challenges of digital inclusion and digital connectivity and resilience are not new; nevertheless, the extent of the challenges have been exacerbated during the COVID-19 pandemic.

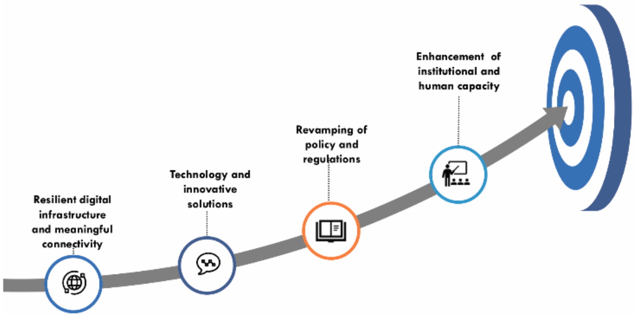


Digital inclusion in education
Digital inclusion in health
Digital connectivity and resilience
Digital inclusion for enterprises and jobs
Digital inclusion for vulnerable persons

Fig. 1. Five thematic areas of focus

Lessons learned

In the current stage of the pandemic, universal and meaningful connectivity and digital inclusion have taken centre stage. As such, the lessons learned from the 15 research reports are not only applicable to “building back better” in the recovery from COVID-19, but can be applicable to the broader effort to close the digital divide. The insights and recommendations of the 15 research reports are distilled into four enablers (Fig. 2).



Resilient digital infrastructure and meaningful connectivity
Technology and innovative solutions
Revamping of policy and regulations
Enhancement of institutional and human capacity

Fig. 2. Four enablers for digital inclusion and digital connectivity and resilience

- The first enabler is resilient digital infrastructure and meaningful connectivity. Digital solutions require that the network and infrastructure are ubiquitous and reliable, and that the services are affordable and accessible by all.
- The second enabler is technology and innovative solutions. Digital solutions to enable e-learning, telemedicine, e-business and digital financing are available and should be embraced. The solutions should also be inclusive and accessible by vulnerable persons.
- The third enabler is revamping of policy and regulations. It is recommended that outdated and restrictive policies and regulations be reviewed to ensure that they facilitate technological developments, and are fit-for-purpose and inclusive in nature.

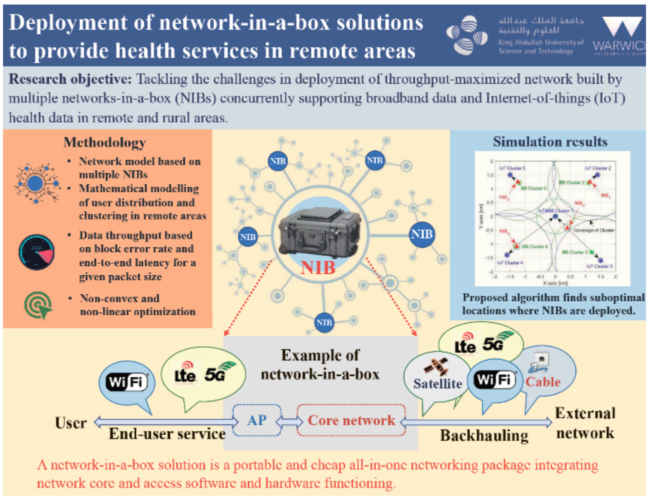
- The fourth enabler is enhancement of institutional and human capacity. Institutions need to prioritize capacity building, particularly in digital skills. This applies to all sectors and institutions – such as health, education and enterprises – and covers health-care providers, teachers, students and employees. Schools should also emphasize equipping their students with digital skills. In addition, vulnerable persons should be empowered with digital skills.

The opportunity

This Research Competition journey that started in July 2021 does not end with the launch of the 15 research reports. In fact, the focus shifts to implementation. We encourage policy-makers to consider reviewing the recommendations and implementing them with impact, as appropriate. Researchers are encouraged to consider the research conducted and to further the work. Finally, we also encourage pilot projects to be conducted or multi-stakeholder partnerships to be formed, and resources to be mobilized, to address these challenges and ensure a more inclusive and connected world.

SECTION 1: DIGITAL INCLUSION IN HEALTH

Network-in-a-box to provide health services in remote areas



Deployment of network-in-a-box solutions to provide health services in remote areas

Research objective: Tackling the challenges in deployment of throughput-maximized network built by multiple networks-in-a-box (NIBs) concurrently supporting broadband data and Internet-of-things (IoT) health data in remote and rural areas.

Methodology

- Network model based on multiple NIBs
- Mathematical modelling of user distribution and clustering in remote areas
- Data throughput based on block error rate and end-to-end latency for a given packet size
- Non-convex and non-linear optimization

Simulation results

Proposed algorithm finds suboptimal locations where NIBs are deployed.

A network-in-a-box solution is a portable and cheap all-in-one networking package integrating network core and access software and hardware functioning.

Fig. 3. Deployment of network-in-a-box solutions to provide health services in remote areas

Background

This winning research project involves two research teams: Mr Mohamed-Slim Alouini and Mr Ki-Hong Park of the King Abdullah University of Science and Technology, Saudi Arabia; and Mr Yunfei Chen of the University of Warwick, United Kingdom. Mr Alouini is an active research participant in this project, and is responsible for overseeing all its aspects. Mr Yunfei and Mr Park have capitalized on their expertise to formulate and model the challenge, build a new methodology and discover new findings.

In this project, the team tackled the challenges of network deployment, while concurrently supporting broadband data and IoT health data in remote and rural areas. It introduced a network-in-a-box (NIB) solution, a portable and cheap all-in-one networking package integrating network core and access functioning. It proposed and verified the deployment algorithm to maximize the total throughput in a network supported by multiple NIBs.

Report findings and outcomes

In order to find the optimal deployment of multiple NIBs, the team will face several new challenges:

- First, the end users in rural areas are sparsely located over a wide area, and it is not appropriate to apply a traditional network model. The network topology in this area is likely to be disjointed and irregular, and therefore the user distribution should be modelled by geometrically reflecting and clustering the real user density with reference to local census. The Gaussian density function is used to characterize the distribution of users in a cluster. Its mean and variance reflect the location of the cluster centre and the population concentration in a cluster.
- Second, the team carefully estimated the performance metric that each NIB could provide at one position. The IoT information in health services might have different characteristics, distinguishing it from conventional network data. It requires periodic sensing over multiple wearables, implants and IoT devices on the human body or using wireless sensing through off-body monitoring units. It usually consists of short data packets that are delay-limited, highly reliable and energy-efficient. Therefore, the throughput of health IoT information in rural areas should be characterized by block error rate and latency for a given packet size. The end-to-end latency is characterized by decoding processing delay and transmission delay. Average aerial throughput of IoT health information can be computed under the statistical models of user clustering and distribution. When the IoT health data is limited to a certain latency requirement, the optimal aerial throughput exists with the unique coding rate. On the other hand, the throughput of broadband data can be featured by a fundamental Shannon capacity formula, and averaged over the modelled user distribution and clustering.
- Third, the conditions in multi-NIB-based IoT health networks have to be carefully considered. NIB is capable of supporting multiple networks, but users in different networks will not overlap. This means that the NIB can support low-rate massive IoT health networks, while servicing broadband. Thus, the resource allocation and position of multiple NIBs need to be jointly considered for all networks. Minimum data rate requirements for users and maximum delay constraints for NIBs are constraining the problem of maximizing the total throughput in a network where multiple NIBs are

receiving data from multiple IoT health devices and from mobile broadband users in cellular systems.

The optimal deployment of NIBs can be obtained by solving non-convex and non-linear optimization problems with block coordinate descent methods and the relaxation techniques used in its inner loop algorithm, such as successive convex approximation, fi approximation and linear relaxation. The optimal coding rate for IoT health data and user association are jointly determined in the proposed algorithm. The simulation results have validated that multiple NIBs are optimally deployed near the centre of user clusters for mobile broadband services to provide high data throughput, while maintaining a bias towards serving IoT health clusters to support low-data-rate IoT service.

Insights from the ground

The network environment varies dynamically, as networks are spontaneously and sporadically demanding over time and space. Temporary and intermittent medical sites require network connectivity only during the concerned time or locations. The network has to be resilient to the circumstances confronting the defect and malfunction due to natural disaster or temporary blackout. Accordingly, 6G initiative groups are pursuing the goal of sustaining ubiquitous network connectivity. In this regard, pop-out networking is more advantageous than fixed network infrastructure in dealing with unusual, temporary and intermittent data on demand. NIB is an excellent pop-out solution with limited but user-friendly hardware and software capability, which saves capital and operation expenditure in a mobile network. NIB perfectly fits in varying environments, owing to ease of deployment, mobility and network flexibility.

Ultra-low latency and ultra-high reliability will be the key indicators to evaluate the quality of physical experience in the private network sectors, based on the extended reality and nearly zero delay interaction in remote rescue and medical operations. Smart ambulances can be ubiquitously connected to the network and provide super-accurate medical assistance. Such potential networks must operate under 5G standards or 6G recommendations. Bearing in mind the advantages of NIB, one consideration will be the manner in which cost-effective NIBs can be used to support these 6G use cases in rural areas.

Key recommendations

- The proposed mathematical modelling of two different communication systems for combining broadband and IoT service enables researchers to further analyse and optimize mobile and heterogeneous networks with latency constraints. If the user distribution and clustering for the target remote areas are modelled precisely, the network operator can deploy and operate a cost-effective, NIB-based network rather than a conventional fibre- optic-based network.
- The proposed algorithm for NIB deployment is a suboptimal solution for finding the positions of NIB to be deployed in the specific network model under user distribution and clustering in sparsely populated areas.
- We should carefully consider the interoperability and coexistence of NIB-based networks along with conventional cellular networks operated by mobile network operators. The mobile broadband users in the town centre and nearby in the rural areas are most likely supported by traditional cellular networks. Spectrum management techniques – such as spectrum sharing, spectrum allocation, or cognitive radio – are required.
- Finally, the user's distribution and clustering in rural areas can vary drastically over time, since the users are sparsely populated, and commercial and residential areas are separated irregularly. The deployment scheduling of NIBs can be dynamically designed over time thanks to the mobility of NIBs.

Improving resilience in developing countries: Digital health provision through telemedicine ecosystem against the pandemic, epidemics and natural disasters in sub-Saharan Africa

Background

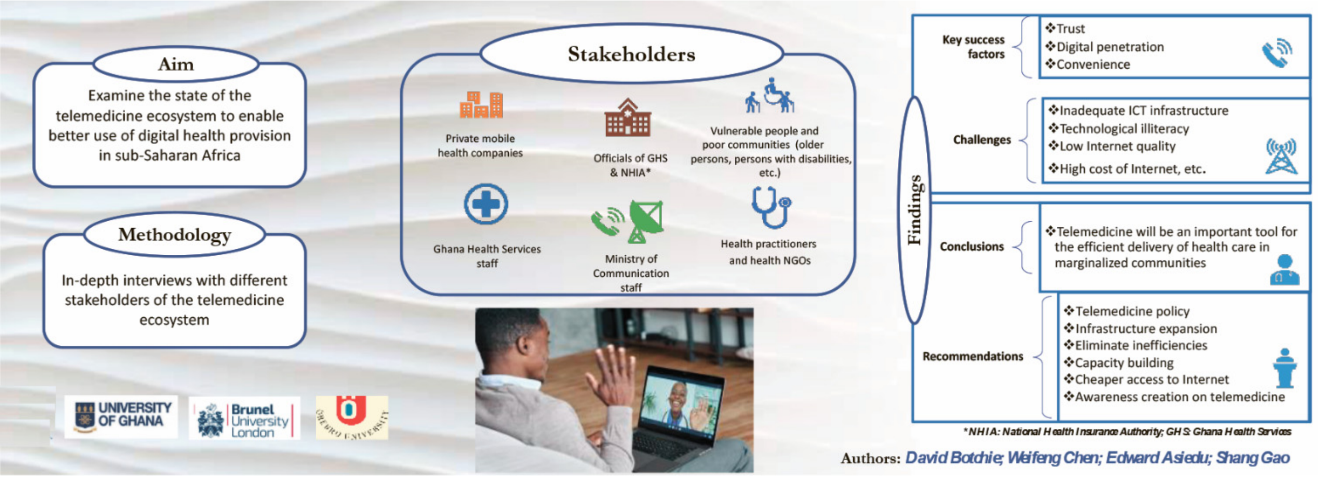
Primary objective: To create an understanding of the dynamics of the telemedicine ecosystem and proffer recommendations to facilitate the sustainable adoption of telemedicine in sub-Saharan Africa (Fig. 4).

Specific objectives

- To determine the state of the telecommunication and telemedicine ecosystems in sub-Saharan Africa.
- To examine how telemedicine has been diffused to vulnerable groups (older persons, disabled and poor) in sub-Saharan Africa.
- To examine how the telemedicine ecosystem has been leveraged to improve resilience to pandemics, epidemics and natural disasters.
- To determine the challenges and successes of expanding access to telemedicine in sub-Saharan Africa.

Research methodology

The study examines secondary data on the topic across various sub-Saharan African countries, and then concentrates on the telemedicine ecosystem of Ghana using in-depth qualitative tools, including community and focus group discussions, to address its objectives. Primary data were obtained through one-on-one interviews with 63 relevant stakeholders in the telemedicine ecosystem using a semi-structured questionnaire. Stakeholders in the health and telecommunication sectors, and community members constitute the target population.



Aim

Examine the state of the telemedicine ecosystem to enable better use of digital health provision in sub-Saharan Africa

Methodology

In-depth interviews with different stakeholders of the telemedicine ecosystem

Stakeholders

Private mobile health companies

Officials of GHS & NHIA*

Vulnerable people and poor communities (older persons, persons with disabilities, etc.)

Ghana Health Services staff

Ministry of Communication staff

Health practitioners and health NGOs

Findings

Key success factors
- ❖Trust
- ❖Digital penetration
- ❖Convenience

Challenges
- ❖Inadequate ICT infrastructure
- ❖Technological illiteracy
- ❖Low Internet quality
- ❖High cost of Internet, etc.

Conclusions
- ❖Telemedicine will be an important tool for the efficient delivery of health care in marginalized communities

Recommendations
- ❖Telemedicine policy
- ❖Infrastructure expansion
- ❖Eliminate inefficiencies
- ❖Capacity building
- ❖Cheaper access to Internet
- ❖Awareness creation on telemedicine

*NHIA: National Health Insurance Authority; GHS: Ghana Health Services

Authors: David Botchie; Weifeng Chen; Edward Asiedu; Shang Gao

Fig. 4. Improving resilience in developing countries: Digital health provision through the telemedicine ecosystem against the pandemic, epidemics and natural disasters in sub-Saharan Africa

Research findings

- Even though telemedicine was not formally in place, people used various digital means to address their health challenges and seek clarifications regarding some symptoms they were feeling during earlier phases of the COVID-19 pandemic.
- Digital penetration, trust and convenience are among the key success factors for a telemedicine ecosystem.
- A telemedicine ecosystem is faced with challenges that hinder the deployment of telemedicine. These include:
  - poor telecommunication and road networks, low ICT capacity of health-care professionals, ICT illiteracy of community members, and financial constraints;
  - the most prominent among these challenges, poor communication network connectivity.
- Government, non-governmental organizations (NGOs) and philanthropists are the main sources of funding for telemedicine interventions. This finding contradicts those studies that conclude that some telemedicine pilots in Ghana were not successful due to lack of funds and little government support.

Results and outcomes

- The telemedicine ecosystem of Ghana is in its early stage and, even though there is improvement in telecommunication services, there are some communities in Ghana with poor network connectivity, which hampers the agenda to expand telemedicine to rural communities.
- Generally, at the policy level, the results show that the telecommunication infrastructure deficit, which has created a digital divide, and policies that do not allow health practitioners in rural communities to attend to some health situations and prescribe certain medication, are the hindering factors to the deployment of telemedicine in Ghana.
- Even though most people are not aware of the existence of telemedicine, they have resorted to various digital means, such as calling friends, health workers they know etc., to seek health advice during the COVID-19 pandemic.
- The success of any telemedicine infrastructure in sub-Saharan Africa will depend to a large extent on the creation of awareness and trust that the digital health services received are of the same quality as those that would have been provided at a health facility.

Insights from the ground

Relevant quotes from participants/interviewees:

- One community member stated: “In my community, the main challenge is network connectivity and the high illiteracy rate. If the people in the community get education about telemedicine and a good network, there will be no issues with it being initiated… the availability of the hardware, not just availability, but its maintenance to be consistent.”
- A District Health Director said: “Telemedicine can be used for emergencies and things that are beyond the level of whoever the client is seeing… If somebody collapses, if a pregnant woman is bleeding, if a pregnant woman is convulsing, you can use it. You are attending to someone; you realize the condition has changed and you need to make a call for information; you can use telemedicine.”

Key recommendations

Based on the research findings, the study proffers the following recommendations for policy action:

- Health-care policy-makers should constantly collaborate with academia to undertake evidence-based studies to support health-care policy-making in Ghana.
- The Government, in consultation with health-care policy-makers and other policy-makers, should develop a

national telemedicine policy to aid in the implementation of telemedicine.

- Health-care professionals and community members should be sensitized to the need to mainstream telemedicine into the health-care delivery system.
- Government should work with the telecommunication companies to address outstanding inefficiencies, such as poor communication networks.
- Ghana Health Service (GHS) is encouraged to consult, design and implement appropriate training modules on telemedicine to build the capacity of health-care professionals.
- The Government should work with the telecommunication companies to implement a toll-free system for telemedicine-related services.
- Ghana Investment Funds for Electronic Communication should expedite action to ensure that telecommunication services are extended to rural communities to support the adoption of telemedicine.
- GHS, through district health directorates, should work with the National Communications Authority to organize clinics to enhance ICT literacy in rural communities.

SECTION 2: DIGITAL INCLUSION IN EDUCATION

2.1. Making higher education truly inclusive

Background

ICT infrastructure has proven vital in helping countries and citizens adapt and respond to the COVID-19 pandemic. Whether this reliance has resulted in greater immediate and longer-term inclusion of marginalized communities is an important question in terms of equitable access to higher education.

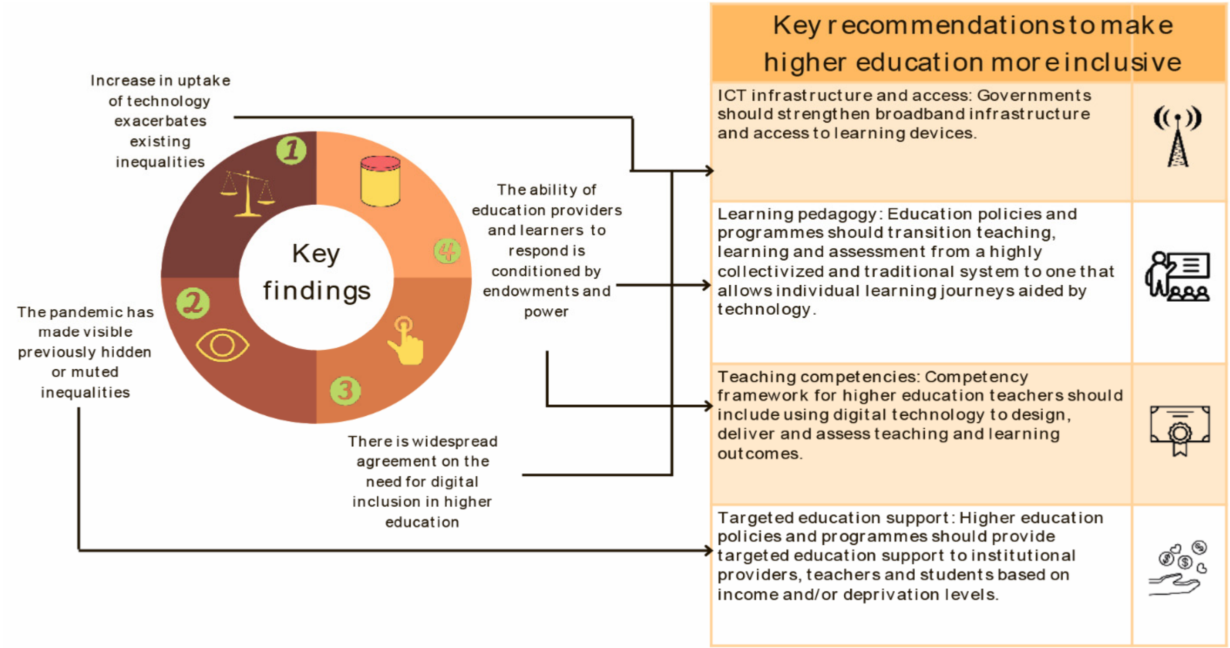
This report investigates the response to the COVID-19 pandemic by three higher education systems. It also describes the outcomes of these interventions in terms of the inclusion (or exclusion) of marginalized students. Finally, it dissects further the situation as to whether the disruption to higher education – particularly the uptake of new modes of instruction, learning and assessment – results in greater inclusion in the future provision of education (Fig. 5).

Report findings and outcomes

Three countries – Australia, the Philippines and South Africa – were selected to study the effects and future outcomes attributable to the COVID-19 pandemic in relation to ICT infrastructure, access and inclusion in higher education. A case study approach was used to study and compare the countries.

Three observations emerge from the South African experience:

- The first is the realization that, as predicted by scholars, neither technology nor open resources necessarily lead to the anticipated democratization effects. Instead, in highly unequal societies (and university systems), an increase in the uptake of technology and open resources is more likely to exacerbate existing inequalities.
- The second observation is that the pandemic has made the invisible visible, especially the historical, economic and geospatial inequalities within and between countries studied. Regardless of the availability of data and devices, the quality of online and other forms of digital educational provision during the pandemic remains open to question.


Key findings
1
Increase in uptake of technology exacerbates existing inequalities
2
The pandemic has made visible previously hidden or muted inequalities
3
There is widespread agreement on the need for digital inclusion in higher education
4
The ability of education providers and learners to respond is conditioned by endowments and power
Key recommendations to make higher education more inclusive

ICT infrastructure and access: Governments should strengthen broadband infrastructure and access to learning devices.

Learning pedagogy: Education policies and programmes should transition teaching, learning and assessment from a highly collectivized and traditional system to one that allows individual learning journeys aided by technology.

Teaching competencies: Competency framework for higher education teachers should include using digital technology to design, deliver and assess teaching and learning outcomes.

Targeted education support: Higher education policies and programmes should provide targeted education support to institutional providers, teachers and students based on income and/or deprivation levels.

Fig. 5. Making higher education truly inclusive

- One more positive observation is how the pandemic has brought together government, higher education institutions and private mobile operators in acknowledging the greater need for digital inclusion. This atypical cooperation resulted in the provision of devices to students in need as well as "zero-rated" (free) access to university learning platforms and other educational websites and resources.

The case of the Philippines, a system more diverse in its mix of both public and private institutions compared with South Africa, suggests that the capacity of institutions and individuals to adapt to, cope and mitigate the impact of COVID-19 on teaching and learning is significantly differentiated; well-resourced, well-connected, and strategically located actors were more able to transition to new modes of education delivery. Several studies have pointed out that there are differences in processes and outcomes, not only due to resources, but also due to actors such as political leadership, social security provisions, degrees of autonomy and centralization, as well as underlying capacity. Education policies in the country have not considered these differences in capacity, leaving actors to respond and cope on their own.

In the case of Australia, the biggest impact on the higher education sector was the result of the international travel ban and the resultant decline in student fee income from international students. The foreign student market is the third biggest export industry in Australia and the country's largest service-sector export. In terms of the impact of the pandemic on social exclusion, the perilous position of international students stranded in Australia attracted the most attention. Despite some measures taken by the Government of Australia to support these students, the Government's response was seen as largely unsympathetic, vividly expressed by the "go home" response. In general, the effects of COVID-19 on the higher education system in terms of the exclusion of segments of the student population received only cursory attention in the media and the academic literature compared with South Africa and the Philippines.

Insights from the ground

By not succumbing to the hype and allure of new digital educational technologies, and by providing a context-sensitive synthesis of the literature and the surveys conducted during the COVID-19 pandemic, the team has taken a step towards making more explicit the actual conditions and their effects on specific segments of the higher education student populations in South Africa, the Philippines and Australia. By providing an account of how the responses of governments, institutions and the private sector impacted on students with limited resources or abilities, this report has shown the limitations of an overreliance on ICTs for education purposes. Report findings and outcomes

Five of the oldest universities – including Hawassa University, Addis Ababa University, Arba Minch University, Jimma University and Bahir Dar University – were selected, presumably for their relatively better experience of using ICT. The exploratory sequential mixed method was used in the research. Thus, the research began with in-depth interviews with 15 staff members, including teachers, college deans and ICT directors chosen purposefully from the five universities. The interviews included questions about the respondents' background, ICT access, digital literacy, ICT use and motivation, students' ICT usage, and opinions on barriers of ICT use in education.

The COVID-19 pandemic has shown that the rapid deployment of technology by various stakeholders is possible. Theoretically, the availability of technologies to a broader segment of the population should result in greater inclusion (that is, participation in communication networks for educational purposes). However, the evidence provided in this report shows that, without the capabilities – many of which are non-material and do not relate to technical skills or access alone – and without an acknowledgement of the social dynamics of systems and networks, parts of the population will always remain excluded.

Key recommendations

Any future integration of online learning as complementary to contact modes of instruction will require substantial investment in the following areas:

- ICT infrastructure and access: In a context where access to technology is challenging, governments should strengthen broadband infrastructure on the one hand, and access to learning devices on the other.
- Learning pedagogy: From a pedagogical perspective, there is a need to formulate policies and programmes that transition teaching, learning and assessment from a highly collectivized and traditional system to one that allows individual learning journeys aided by technology. More research is required to understand better the outcomes and impacts of these new modes of teaching, learning and assessment about COVID-19.
- Teaching competencies: The competency framework for higher education teachers should include using digital technology to design, deliver and assess teaching and learning outcomes.
- Targeted education support: Higher education policies and programmes should provide targeted education support to institutional providers, teachers and students, based on income and/or deprivation levels to transition towards better use of technology in education.

Determinants of digital inclusion in higher education: Exploring the Ethiopian context

Background

ICTs support the United Nations Sustainable Development Goal 4 (Ensure inclusive and equitable quality education and promote lifelong learning opportunities for all) by providing an alternative route to education, as shown during the COVID-19 pandemic, when face-to-face communication becomes inconvenient. However, that requires identification and removal of the digital divide that would create inequalities in ICT access and use. This study aims to explore the digital divide in the Ethiopian higher education context, and recommend solutions that could be used by higher education institutions and policy-makers.

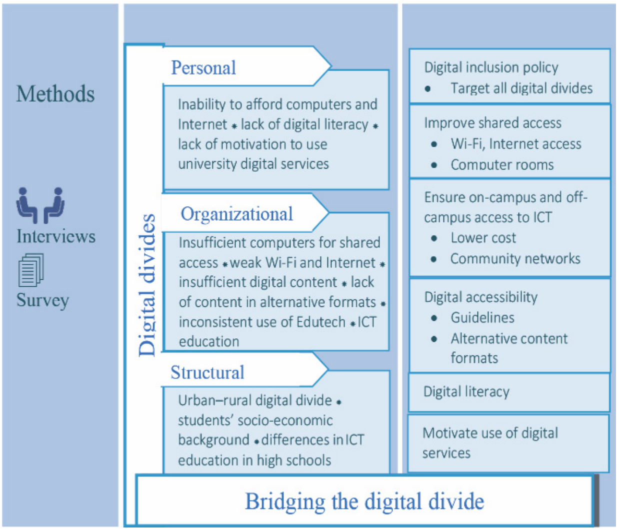

Methods
Interviews
Survey
Digital divides
Personal
Inability to afford computers and Internet • lack of digital literacy • lack of motivation to use university digital services
Organizational
Insufficient computers for shared access • weak Wi-Fi and Internet • insufficient digital content • lack of content in alternative formats • inconsistent use of Edutech • ICT education
Structural
Urban–rural digital divide • students' socio-economic background • differences in ICT education in high schools
Digital inclusion policy
• Target all digital divides
Improve shared access
• Wi-Fi, Internet access
• Computer rooms
Ensure on-campus and off-campus access to ICT
• Lower cost
• Community networks
Digital accessibility
• Guidelines
• Alternative content formats
Digital literacy
Motivate use of digital services
Bridging the digital divide

Fig. 6. Determinants of digital inclusion in higher education: Exploring the Ethiopian context

The interview data was used to design a questionnaire, which was completed by 418 undergraduate students selected from the universities using a stratified proportionate sampling technique. A total of 43 per cent of the students said that they owned PCs, and 90 per cent said they owned smartphones, while 7 per cent said they had tablets. Those who didn't own computers made use of shared access facilities, such as library computer sections (22 per cent) and computer labs (22 per cent), while 10 per cent said they borrowed laptops from friends.

The study identified problems related to ICT access, digital content, accessibility, digital literacy, ICT use and ICT policy. The access-related problems included students' inability to afford computers and Internet connection, an insufficient number of computers in computer rooms, weak Wi-Fi and weak Internet connections. The universities had digital libraries that were not well developed and not accessible outside of university compounds. Learning management systems (LMSs) were underutilized, though they saw improved utilization during the COVID-19 shutdown. The e-learning attempted during the shutdown remained largely inaccessible for undergraduate students, since most of them were from rural areas where there was no access to the Internet. Efforts to address the needs of students with disabilities were limited to production of content in Braille form, and provision of audio recorders to students with visual impairments. There was a lack of knowledge on accessibility and accessibility guidelines.

There are differences in digital literacy among students as well as teachers. Teachers' digital literacy levels have affected the production and delivery of digital content as well as the utilization of LMSs. Those with better ICT literacy would include multimedia content and links to other sources, whereas others were limited to PDF documents and PowerPoint slides.

A related problem was inconsistency among teachers in the use of educational technology – for instance, some used LMSs, while others did not. That reflects a lack of institutional norms that govern consistent use of educational technology.

Underutilization of digital services is the other problem. When asked about tools for content sharing, 52 per cent of the respondents said they preferred Telegram, 36 per cent said they would like the materials sent via e-mail, and 11 per cent said they would use university portals.

The little motivation to use university portals was attributed to platform complexity, lack of digital literacy, and teachers' low expectations on the use of LMSs by their students.

Insights from the ground

As one student put it: "I mostly am dependent and feel comfortable with high-end smartphones and I use them instead of computers." The prevalence of smartphones and students' preference of least interactive technologies such as Telegram imply the importance of incorporating user needs and preferences to develop usable and accessible educational applications.

Removing barriers of access is an important step to digital inclusion. Nevertheless, barriers are revealed through usage. In this research, there were students who did not own computers and had no Internet connection, but said they had not faced barriers. On the other hand, there were others with computers and access to the Internet who listed a number of barriers.

Higher education institutions, therefore, would have to promote usage of their digital services and actively work to identify and remove barriers. Maximizing ICT use would require dealing with motivational issues. The use of ICT policies and guidelines would enforce a consistent use of ICT in higher education institutions. Digital literacy programmes would intrinsically motivate students as well as teachers to use ICT to their best advantage.

Key recommendations

The first important step could be the development of a digital inclusion policy that recognizes the technical, socio-demographic and socio-economic barriers that are explored in the study. This would help to create a shared understanding of digital inclusion, and institute consistent practices that maximize the use of available educational resources.

Thereafter, implementing digital literacy programmes (including computer literacy, ICT literacy, information literacy and media literacy) that target different groups would be important. It would also be important if the existing continuous professional development for teachers incorporated courses and trainings on digital literacy. Establishment of inclusive ICT infrastructures that incorporate the needs of persons with disabilities, working with different governmental and non-governmental partners to ensure on-campus and off-campus access to ICT resources, could be important steps to expand access to ICT and digital services.

Digital inclusion in education will not be complete without accessible digital content. Thus, utilization of accessibility guidelines to produce content that can be accessible to all, including students with disabilities, would be important. Moreover, production of content in alternative formats (such as PDF, HTML and audio) would help to address the needs of students with different needs and preferences.

Abbreviations

AFRALTI	African Advanced Level Telecommunications Institute
BDT	ITU Telecommunication Development Bureau
COMESA	Common Market for Eastern and Southern Africa
CWN	community wireless network
DM	diabetes
DSA	dynamic spectrum access
GHS	Ghana Health Service
HTN	hypertension
ICT	information and communication technology
IDP	internally displaced person
IoT	Internet of Things
ISP	Internet service provider
ITU	International Telecommunication Union
ITU-D	ITU Development Sector
LMS	learning management system
MSMEs	micro, small and medium-sized enterprises
NGO	non-governmental organization
NIB	network-in-a-box
SIDS	Small Island Developing States
UN-Habitat	United Nations Human Settlements Programme
VSAT	very small aperture terminal

Acknowledgements

"Build back better with broadband: Research stories from the front line" was prepared by the Connect2Recover initiative of ITU under the overall supervision of Mr Alex Wong, Chief of Special Initiatives. The core team was comprised of Mr Sameer Sharma, Ms Karen Woo and Mr Chuluunbat Tsendsuren. We would also like to especially acknowledge the support by Huawei, a Sector Member of the ITU Development Sector (ITU-D) – in particular Mr Rene Arnold and Mr Andy Hu – for their contributions to the Connect2Recover Research Competition.

We are grateful to the global five-member jury that selected the 15 winners of the Connect2Recover Research Competition. The jury was led by Mr Ahmad Reza Sharafat, Professor of Electrical and Computer Engineering at Tarbiat Modares University, Islamic Republic of Iran. Other members of the jury include Ms Ellen Helsper, Professor of Digital Inequalities at the Department of Media and Communications at the London School of Economics and Political Science, United Kingdom; Ms Ida Nganga, Regional Head of Anglophone Countries of the United Nations Educational, Scientific and Cultural Organization (UNESCO) Emerging Technologies for Development Steering Committee, based in Bordeaux, France; and from ITU, Mr Cosmas Zavazava, Chief of Partnerships for Digital Development Department, and Ms Nur Sulyna Abdullah, Chief of Digital Knowledge Hub.

Finally, we would also like to gratefully acknowledge the efforts undertaken by the 15 research teams. The research stories, insights and recommendations from their 15 research reports have been compiled in this publication.

References

[1] ITU, Connect2Recover Research Competition – Winning Projects Booklet, available at www.itu.int/en/ITU-D/Documents/connect2recover/research-competition/Connect2Recover-winning-projects-booklet-final.pdf.

[2] ITU, "Information Sessions on Connect2Recover: Research Competition Papers focusing on Africa", available at www.itu.int/en/ITU-D/Pages/events/connect2recover/infosessions-research-competition-papers-focusing-on-Africa/default.aspx.

[3] ITU, "ITU's 'Best practices and recommendations for digital inclusion through resilient infrastructure'", available at www.itu.int/en/ITU-D/Regional-Presence/Africa/Pages/EVENTS/2022/P2C_Addis.aspx.

STRATEGY PAPER FOR CIRCULAR ECONOMY: MOBILE DEVICES

Materials prepared by GSMA Association,
London, United Kingdom
www.gsma.com

DOI: 10.36724/2664-066X-2022-8-6-43-51

ABSTRACT

Sustainability challenges can only be addressed at a systemic level, and this is why the GSMA is proud to play a role in helping the mobile industry become more sustainable. In 2016, the mobile industry was the first industry to commit fully to the 17 United Nations Sustainable Development Goals and, in 2019, the GSMA Board set a climate ambition on behalf of the industry to reach net zero carbon emissions by 2050 at the latest. Earlier this year, the GSMA published its first Strategy Paper on the Circular Economy, which focussed on how network equipment can evolve towards more circular business models. Continuing the exploration of circularity, this paper looks at the largest environmental impact of the mobile industry – mobile devices. The report has been developed with Ethos, a Swedish management consultan-cy specialising in sustainability, in collaboration with Tele2, as the Project Group lead, and Project Group members from the GSMA.

KEYWORDS: *GSMA, Tele2, mobile devices, circular economy.*

Introduction

Consumption of natural resources is already at an unsustainable rate and is increasing. Scientific evidence indicates this will lead to a collapse in the natural systems upon which humans depend. However, existential challenges such as climate change, waste, pollution, resource scarcity and biodiversity loss can be solved by moving to a more circular economy, and this idea is gaining recognition globally.

For the telecommunications industry, one of its biggest environmental impacts is from customers accessing connectivity through connected devices. This strategy paper therefore focusses on the opportunities to transition both mobile devices and customer premises equipment such as routers to more circular business models. In developing a circular approach for the industry, the research has referenced widely agreed principles of a circular economy as well as existing frameworks and metrics that are already being used, both within the industry and in other sectors. There has been consideration of current and proposed circular economy policies that governments around the world are implementing.

The circular model includes a vision for 2050 to help drive the industry towards a sustainable future. This is defined as a future where devices have as long a lifetime as possible, where they are made with 100% recyclable and recycled content using 100% renewable energy and where no device ends up as waste. The strategy paper explores how telecommunication operators can understand their current position within the circular economy, how they can accelerate the circular transition by engaging with key stakeholders in the value chain and how to measure progress by using circular metrics and actions covering both ‘entry-level’ participation and ‘leadership-level’.

The benefits of this approach are broad, being environmental, social and economic:

- Extending the lifetime of all smartphones in the world by just one year has the potential to save up to 21.4 million tonnes of CO2 emissions annually by 2030, equal to taking more than 4.7 million cars off the road.
- A reduction in the 30m adults and children currently experiencing adverse health impacts from informal e-waste recycling.
- A refurbished mobile device market predicted to be worth more than $140bn by 2030, compared to $50bn in 2020.

This strategy paper further explores the barriers to achieving a circular economy for devices, along with circular incentives and existing examples of best practice. The paper outlines four immediate opportunities to improve circularity:

1. Understand product flows, increase the number of devices collected from consumers and create a foundation to measure reclaimed devices and treatment method by share of recycled, repaired, reused and reclaimed devices.
2. Increase consumer awareness, based on understanding consumption habits in terms of end-oflife treatment and incentives to increase longevity of devices.
3. Engage with suppliers to improve eco-design and sustainable production leading to greater repairability and durability of devices, which will increase their lifespan.
4. Engage with repairers and recyclers to increase the number of devices that are reclaimed, repaired and recycled to maximise value retention within the economy.

Positioned between consumers, device suppliers and repairers/recyclers, telecommunication operators have a fantastic opportunity to contribute to a circular transition for devices, both from a direct control perspective as well as through influence and partnerships.

By moving to a circular business model for the industry, negative environmental and social impacts will be reduced. This means the industry can meet its demand for materials without depleting the global supply of finite resources. It will also create new market and employment opportunities and will support a just transition given supportive government policies and incentives.

Two product groups are included in the scope of this strategy paper: mobile devices and customer premises equipment. The term ‘devices’ will be used hereafter to describe both product groups.

Mobile devices

The product category ‘mobile devices’ includes smartphones, tablets and feature phones, which may be similar in material content but can vary in size.

Customer premises equipment

The customer premises equipment (CPE) product category includes in-home devices such as set-top boxes, internet routers, Wi-Fi hubs and access points.

The strategy paper was developed through:

- Desktop analysis of new and existing research on the circular model and the circular economy within the telecommunications industry.
- Interviews with industry experts and circular economy practitioners and dialogues with project members.

Currently, the global population is using natural resources corresponding to 1.75 Earths. This means that the global economy uses resources at a rate faster than nature can regenerate, causing resource depletion. The consumption of resources is also accelerating – by 2060, global GDP is projected to triple in size and the world’s resource consumption is estimated to double.

If these trends continue, the environmental and socioeconomic consequences will be severe. The effects of resource depletion will be seen not only through a reduced ability to mitigate and adapt to climate change, but also through its impact on biodiversity and ecosystems. Disruption of planetary systems is already seen through global warming and more extreme weather such as heatwaves, storms and flooding.

The environmental impact of the telecommunications sector is derived from activities throughout the value chain, from raw material extraction and processing, production and assembly of electronic devices to packaging and transportation, as well as by the energy consumed through use of devices and in waste management.

The use of connected devices is expected to grow, and this digitalisation can enable the future low carbon economy and a more resilient society. Mobile technology is already harnessing the Internet of Things (IoT) and artificial intelligence (AI) to create solutions that allow societies not only to mitigate emissions, but also to adapt and become more resilient to the impacts of climate change.

At the same time, demand for these solutions will further accelerate the consumption of raw materials6 required to manufacture devices such as mobile phones and routers. Currently, around two billion phones are sold annually and more than 90% of the global population owns a mobile phone. In 2021, there were an estimated 7.78 billion active smartphones and feature phones around the globe. [1-3]

This number is projected to increase; by 2030, the total number of smartphones and feature phones is predicted to reach nine billion. A similar trend is seen in the global router market, which is predicted to almost double from 2020 to 2030.

The value chain of devices

The value chain of devices is long and complex, with hundreds of businesses involved.

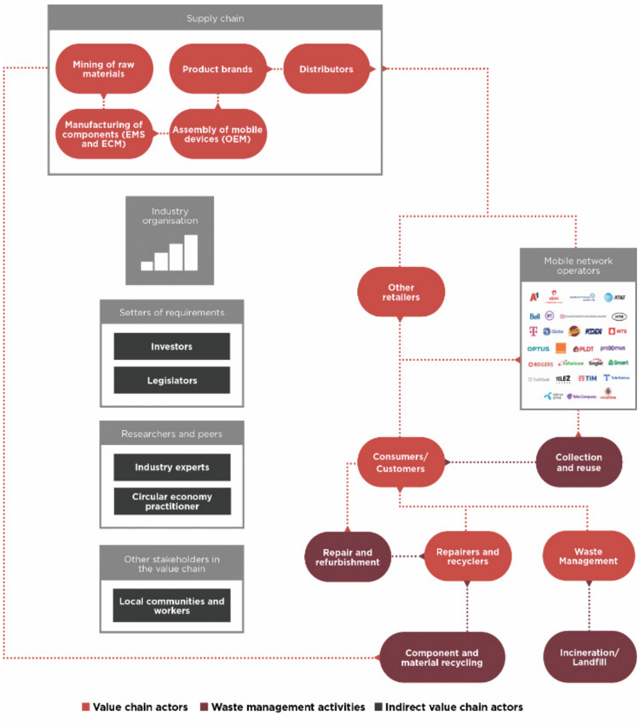


Fig. 1. A simplified version of the value chain of devices. Depending on countries and business model, the value chain can vary to some extent

As an example, an iPhone contains components from more than 200 suppliers. Each step of the value chain entails circular economy-related challenges and opportunities. Simplified, the manufacturing of devices consists of three main areas: raw material extraction, component manufacturing and assembly.

More than 50 different materials could be found in an average smartphone, such as: 29% plastic, 16% ceramics, 15% copper and compounds, 10% silicon plastics, 10% other metals, 9% epoxy, 8% other plastics and 3% iron13. The material in all 7.78 billion smartphones and feature phones around the globe could contain an estimated 124,000 tonnes of copper, 2,721 tonnes of silver, 264 tonnes of gold and 117 tonnes of palladium.

Raw material extraction, primarily mining practices, has negative environmental and social impacts due to contamination of air, soil and water by chemicals, heavy metals or acidic minerals when these are emitted or mixed with wastewater. Mining activities can cause soil erosion and loss of biodiversity as the practices include modification or destruction of habitats.

Production and assembly of components is not only material-intensive, but is also energy-intensive. It often uses fossil fuel energy sources, which is why approximately 80% of the climate impact from a smartphone comes from the production stage of the device and its components. According to the United Nations Environmental Programme (UNEP), resource extraction and processing of fossil fuels, metals and minerals make up 36% of global greenhouse gas emissions and 7% of global biodiversity loss.

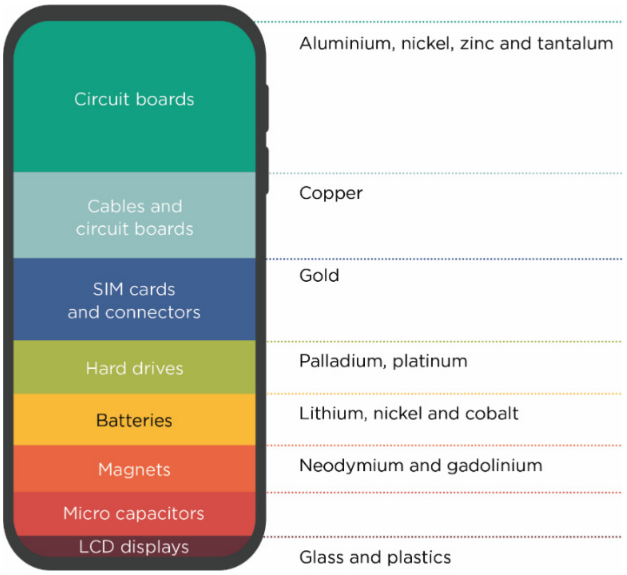


Fig. 2. An example of materials in a mobile phone

The average use time of a phone is around three years. However, the technical lifespan is between four and seven years and the optimal lifetime for a mobile phone in terms of minimising its climate impact could be at least 25 years. However, extending the lifetime of all smartphones in the world by one year has the potential to save up to 21.4 million tonnes of CO2 emissions annually by 2030, equal to taking more than 4.7 million cars off the road [4-7].

The current rate of consumption of devices contributes to the growing generation of e-waste (electronic waste such as discarded electrical or electronic devices), with a considerable amount of it being outside of formal waste management. It is estimated that as much as 86% of global e-waste is estimated to be treated outside of formal waste management, with small IT and electronics such as devices constituting around 9% of the total e-waste generated.

The final destination of many of these e-waste streams is unknown, but can end up in regular waste collection, dumped in landfills or burned in both formal and unregulated settings. There are data and knowledge gaps across all electronic waste streams, including devices such as routers and mobile phones.

As an example, where specific regional data is available, official take-back rates of mobile phones rarely exceed 15%, meaning that 85% of mobile phones are not formally recycled. However, this data does not include mobile phones stored in people's homes or those which are either passed on or sold to other consumers.

As significant amounts of e-waste are handled outside formal systems, fully functional devices could be discarded or recycled instead of collected and reused or repaired, losing the full potential of their useful lifecycles as well as the embedded value in components, materials and energy.

Social impacts

Apart from environmental impacts, the life cycle of devices is also associated with social issues that impact individuals and communities. Evidence has been uncovered of extraction of much-needed minerals present in mobile devices being associated with human rights risks. For example, in raw material mining of conflict minerals and cobalt, there are recorded instances of child labour, such as in mining of gold in Ghana and mining of cobalt in the Democratic Republic of the Congo.

Poor labour conditions including the lack of safety equipment or deficient security practices have also been documented in several countries such as Bolivia and Bulgaria, where materials such as tin, silver and copper are sourced for devices. Mining of raw materials may also lead to the release of toxic metals and contamination of soil or freshwater, not only affecting workers' health, but also local communities in proximity to mining areas or dumpsites. Moreover, communities that can be affected by proposed mining activities or which are in proximity to planned projects within the value chain are at risk of being neglected from the planning process. In Armenia, the planning of a gold mine failed to include local residents in the Environmental Impact Assessment, resulting in a number of protests.

Freedom of speech can be negatively impacted in the value chain of devices. In gold and copper mining in Bulgaria, cases of intimidation and silencing workers from speaking up regarding poor working conditions have been reported. Similar issues are also visible in manufacturing, for example in China and Vietnam, where workers can be exposed to poor working conditions related to (for example) long working hours and excessive overtime, lack of worker representation and trade union rights as well as insecure contracts.

The increasing generation of e-waste, with significant amounts processed in the informal sector, also pose a risk to people's health and safety. For example, as many as 18 million children and adolescents and 12.9 million women could be at risk from adverse health outcomes associated to e-waste recycling. Poor e-waste management may also lead to contamination of nearby areas. Much of the informally treated e-waste is estimated to be illegally traded or dumped, predominantly in Ghana and Nigeria.

The way forward

With these negative environmental and social impacts in mind, there is an urgent need to accelerate the circular transition of the economy to be able to meet the demand for materials without depleting the global supply of finite resources. This would also reduce both the environmental and social impact of devices. The devices in scope should be used for longer and resource efficiency, reuse, repair and recycling rates should increase.

While demand for new devices remains high, there is already evidence of a budding circular economy – 11% of smartphones sold worldwide today are refurbished and the market is increasing. Consumers are also becoming more interested in second-hand products as well as sustainability at large.

Devices are being used for longer. In the past seven years, the mobile phone replacement cycle has increased by 10 months, from 24 months in 2014 to 34 months in 2021 worldwide. This trend is expected to continue, with the refurbished mobile device market predicted to be worth more than $140bn by 2030 compared to $49.9bn in 2020. To put this into context, the global telecommunications market was valued at $1,708bn in 2021.

The environmental, social and economic benefits of moving towards a more circular economy for devices are clear. The purpose of this strategy paper is therefore to explain how the industry stakeholders can take leadership in working towards a sustainable telecommunications industry in general, and to improve the circularity and long-term sustainability of telecommunication devices in particular.

A circular economy

A circular economy is defined as an economy that retains the value of materials and products for as long as possible, moving from a linear economy (take-make-dispose) to a system where resources are used more efficiently and waste is reduced.

Moving from a linear to a circular economy requires transformation to a system that uses less material, extends the longevity of products, increases product use and recirculates products, components and resources back into the material flows of the economy.

In 2020, the global economy was estimated to be 8.6% circular, meaning that more than 90% of the world is still stuck in a linear economy where material and products do not get recycled or reused, but end up being wasted.

The circular economy is beginning to be embedded in policy across several continents:

- The EU’s Green Deal51 and the Circular Economy Action Plan
- The US’s National Recycling Strategy
- hina’s Development Plan for the Circular Economy
- Africa’s African Circular Economy Alliance
- Latin America’s Circular Economy Coalition Latin America and the Caribbean.

However, the focus on how to deal with circular devices varies, from increasing recycling rates to prolonging the lifetime of devices or empowering consumers with the right to repair. There is a need to have an agreed approach for the circular economy of devices to achieve a system shift in the global economy. By exploring best practice across these regions, along with what individual businesses are doing, this strategy paper proposes a globally relevant approach. In response to national strategies and action plans, as well as to boost circular economy overall, numerous organisations and companies have published frameworks on the topic [8-10].

Circularity frameworks and metrics

Circular economy frameworks and indicators are widely discussed topics by internationally acknowledged organisations, such as the ‘Circular Transition Indicators’ by the World Business Council for Sustainable Development (WBCSD), ‘Circulytics’ by the Ellen MacArthur Foundation, ‘CIRCelligence’ by Boston Consulting Group and ‘the Circular Gap Metric’ by Circle Economy. The GSMA also recently published the ESG Metrics for Mobile, including metrics connected to waste, repair, reuse and recycling.

In addition to metrics developed by industry initiatives, emerging regulation and internationally developed standards are also incorporating the circular economy – for example, European regulations such as the Corporate Sustainability Reporting Directive (CSRD), the Sustainable Finance Disclosure Regulation (SFDR) and the EU Taxonomy.

Internationally adopted standards like Global Reporting Standards (GRI) and Sustainability Accounting Standards Board (SASB) also have metrics regarding circular economy.

Circularity model for devices

To materialise the potential of circular devices, the GSMA has defined a shared industry vision to 2050. All actors in the telecommunication ecosystem will need to work together to achieve this vision:

“Devices with as long a lifetime as possible, made with 100% recyclable and recycled content, 100% renewable energy and where no device ends up as waste”

To achieve the vision, a circularity model has been developed. The model includes two overarching concepts of ‘maximised longevity’ and ‘zero waste’, which permeate the four solutions to commonly observed barriers related to a circular economy for telecommunication operators. As circularity is not achieved in isolation, the solutions entail collaboration with stakeholders across the value chain.

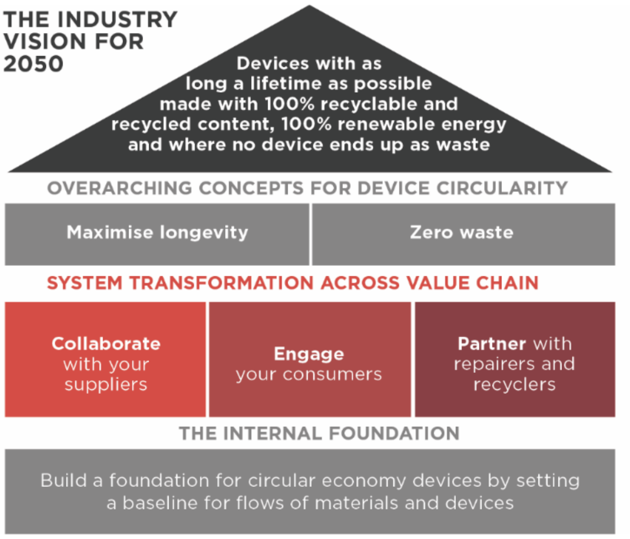

THE INDUSTRY VISION FOR 2050
Devices with as long a lifetime as possible made with 100% recyclable and recycled content, 100% renewable energy and where no device ends up as waste
OVERARCHING CONCEPTS FOR DEVICE CIRCULARITY
Maximise longevity
Zero waste
SYSTEM TRANSFORMATION ACROSS VALUE CHAIN
Collaborate with your suppliers
Engage your consumers
Partner with repairers and recyclers
THE INTERNAL FOUNDATION
Build a foundation for circular economy devices by setting a baseline for flows of materials and devices

Fig. 5. The circular model displaying the circular transition for devices.

Source: GSMA

Boosting circularity: barriers and opportunities

The roadmap to a circular economy for devices introduces four steps to incentivise the transition towards maximized longevity and zero waste. Actions are both directly for telecommunication operators as well as in cooperation with others. Given that every telecommunication operator is embarking on the circular economy journey from a different starting place, each step of the roadmap consists of both ‘entry-level’ and more advanced ‘leadership-level’ actions.

Lack of data on end-of-life of devices

The data of global take-back rates of devices varies in availability and, where data is available, the take-back rates rarely exceed 15%. Available data only concerns devices that have been handed into a formal recycling facility and do not consider the secondary market that exists among customers. There is an evident data gap on what happens to the device when the first consumer no longer needs the device but does not return it to a telecommunication operator.

In many parts of the world, data is lacking and this could have several explanations. For example, there are limited organised or formal systems for the take-back of e-waste and devices in many countries in Africa. However, the data are estimates and, as an example, MobileMuster – accredited under the Australian Government's Recycling and Waste Reduction Act 2020 – reported a 98% recycling rate of the collected mobile phones.

Choosing suitable metrics to measure device flows

By gathering data on the flow of devices, a foundation for circular economy for devices can be built and actions can be targeted to improve circularity. There are two sets of actions and metrics recommended: 'entry-level' and 'leadership'. Measuring device flows will help operators understand their current circular status for devices, and also improve the understanding across the industry.

Tele2 'Mapping of material flows'

As a result of Tele2's ambitious efforts to reduce emissions by 90% in their own operations in two years, the majority of Tele2's emissions lie within its value chain today. Transitioning to a more circular economy has the potential to decrease the environmental impact of Tele2's value chain and is one of four key focus areas of Tele2's sustainability strategy.

In the autumn of 2021, Tele2 mapped out the most important material flows of its operations and identified key questions for moving forward in the circular transition. The material flow analysis included both network infrastructure, offices and stores as well as customer products in B2B and B2C offers. Material inputs of around 3,000 tonnes were identified. Excluding the network equipment, the largest material flows in terms of weight concerned plastics and different metals such as aluminium and coppe.

The analysis showed that around 8-15% of all procured mobile devices are either reused or recycled. To close the loop for mobile devices, the number of reclaimed devices must increase. The insights from the material mapping will be used to further develop Tele2's understanding and im-plementation of circular economy for devices.

Choosing suitable metrics to measure device flows

Entry-level actions and metrics will help to understand what proportion of devices are wasted, recycled and refurbished. Leadershiplevel actions and metrics go a step further and measure overall device reclamation rates, and what devices are composed of.

Entry-level action:

• Set up a structure (e.g. system tool or Excel) to collect data regarding devices.

• Quantify number of collected devices and number of devices being incinerated or landfilled, recycled, repaired and reused.

Entry-level metrics:

• Mobile device and CPE waste generated in tonnes per fiscal year.

• % of mobile devices and CPEs recycled by unit sold per fiscal year.

• % of mobile devices and CPEs recycled by purchase price per fiscal year.

• % mobile devices and CPEs repaired and reused by unit sold per fiscal year.

• % mobile devices and CPEs repaired and reused by purchase price per fiscal year.

Leadership-level action:

• Based on the entry-level data, conduct an assessment on data gaps, how to improve data quality and leverage points.

• Estimate material content (plastic, metals, critical minerals, etc.) of procured and reclaimed products to understand the material value of input and output flows.

Leadership-level metrics:

• Percentage of mobile devices and CPE collected from consumers of total units sold in fiscal year.

• Weight and share of renewable, reused and recycled material in procured devices and CPEs.

Consumers want to do more, but need information

Consumers are critical to achieving a circular transition for the devices because, once the devices are sold, the consumer is in control of the device. Understanding consumer behaviour is a difficult and complicated task, and aspects such as affordability, information availability, social norms and preferences can all affect the final behaviour of the consumer [11-12].

However, consumer awareness of sustainability and circularity is on the rise. For example, 51% of consumers globally think that the consumer electronics sector is not doing enough to reduce, reuse and recycle waste.

In addition, 72% of consumers would like to buy products that are more durable and 47% want to buy second-hand instead of brand-new items. Also, 53% of consumers are comfortable with using second-hand phones. When it comes to purchasing a device, consumers feel less engaged and aware. A majority (80%) of EU citizens think it is hard to find information on durability and repairability, and 64% think that it is hard to tell how long a device will last65.

India	74%	UK	44%
China	62%	US	44%
Spain	52%	Sweden	37%
Italy	50%	Germany	43%
France	47%	Netherlands	37%
Australia	47%	Norway	29%
		Japan	24%
OVERALL	45%		

Fig. 8. Consumer interest in buying exclusively from brands that practice circularity. Percentage of respondents who say they are interested in buying exclusively from brands that concentrate on circular and sustainable practices

Source: Capgemini Research Institute, ciruclar economy surbey, August-September 2021, N=7,819 consumers

Moreover, consumers are concerned with how data security and integrity are ensured when a device is collected, creating a barrier to returning the device68. Around the world, consumers are still holding on to old devices; an estimated 700 million second-hand mobile phones are left unused in the EU alone when they could be recycled, refurbished or even reused by someone else69. A majority save the devices as a spare but, for others, either concerned regarding data security or not knowing where to hand in old devices, they just end up in peoples' drawers.

Consumers must have the knowledge and opportunity to care for devices while in use to maximise longevity and also return devices they no longer need so the device can be reused or recycled.

Telecommunications operators can provide consumers with the information needed to promote them into a change of behaviour. To measure consumer engagement, telecommunication operators can use not only descriptive and qualitative metrics, but also estimations of the destination of the fate of devices based on findings from consumer surveys.

Once the company has gathered information on consumer habits and mapped the potential destination, the next step involves communicating with the consumer to potentially affect their behaviour – for example, through communication campaigns to raise awareness on how to turn in devices, caring for devices or promoting the use of ecolabels.

The supply chain for devices is complex

To achieve a circular economy for devices, there is a need for collaboration and communication within the value chain. A recent study shows that 49% of telecommunication operators see the complexity of supply chains as a barrier for circular economy. At the same time, 84% believe that a circular economy can help solve challenges within the supply chain.

Stakeholders in the supply chain, such as manufacturers and product brands, play a key role in terms of the circular performance of the devices, with the possibility to design devices to prolong their life, increase recyclability and keep them free from hazardous materials. It is also important to consider possible vulnerable stakeholders within the supply chain, such as workers and local communities.

In order to ensure a fair transition of the industry towards a circular economy, manufacturers can play a key role to ensure that violations against human and labour rights are avoided within their supply chains.

Telefónica: Accelerating Circular Economy through innovating supply chain processes. Regarding CPE, Telefónica has developed and deployed VICKY in Brazil, an initiative that has been recognised for its innovation by the industry (i.e: Gartner, Forbes).

This platform, based on Blockchain, is now tracing millions of CPE yearly, from components manufacturing to distribution, installation and reverse logistics across the E2E value chain. Telefónica uses this initiative to lead the transition to circular economy, while creating business value through a more efficient, faster, simpler and sustainable supply chain. With more than 100 different companies involved in the process, placing more than 15 million pieces of equipment on the market each year, it has drastically improved product recovery rates (up to +25p.p.), refurbishment processes, product lifespan, product design and recycling and scrapping rates, intending to collect 100% of the uninstalled or inactive equipment, both in customer premises and in collecting points.

In addition, for mobile devices, Telefonica uses MARA, a fully omnichannel process that allows consumers to automatically assess their devices and access the Telefónica trade-in programmes anywhere, providing instant and real value added to customers without risks (0% discrepancies rate) and, at the same time, defining the best device destination (reuse, resell, repair or recycle) before collecting them.

Targeted and specific supplier engagement is needed

Preconditions to prolong product lifecycles such as repairability and durability lie outside telecommunication operators' direct control as such factors are defined in the design phase of a product. Nevertheless, operators jointly hold significant purchasing power as retailers of devices around the world.

There are also technical limitations on the longevity of devices (for example, software upgrades, lifetime of components such as the battery or the availability of replaceable components). The technically useful lifetime is limited because these obstacles prevent use of the device after a certain number of years. These limitations should be considered when telecommunications operators have ownership over the design process, such as white-labelled CPE.

All forms of industry initiatives and partnerships are key to creating a systemic change towards a circular economy for devices. Telecommunications operators may engage with suppliers and product brands to target the barriers created by complex supply chains. By requesting brands to design and produce devices with longer lifespans, higher repairability and more recycled components and materials, the telecommunications operators have the potential to positively influence the market.

Three metrics are suggested in the engagement with suppliers. The first and second relate to criteria set towards suppliers to increase circularity of devices. The second metric enables the organisation to further comprehend the circularity of their flow of devices in terms of aspects that could result in a longer lifetime for devices. The third metric concerns procedures for following up suppliers.

Untapped potential in secondary markets

Although the market demand for second-hand and refurbished devices is starting to increase, only 11% of smartphones sold worldwide are refurbished. A global circular transition will move employment from primary production, such as resource extraction, to secondary and tertiary production, such as the recycling and refurbish sectors. Estimations predict a loss of eight million jobs globally within industries such as mining and manufacturing, but these can be relocated within new growing sustainable businesses, if given the right incentives.

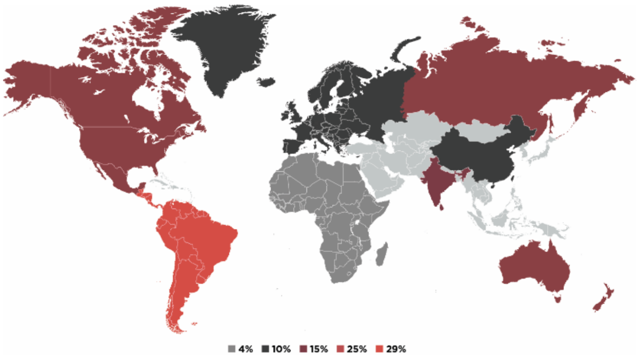


Fig. 10. The map displays the market growth rate of refurbished smartphones in 2020-202176. The market growth rate for refurbished smartphones rose by 15% in 2021 from 2020, with Latin America and India as frontiers

**Figures for Oceania and Russia are based on the global average as data is unavailable. Source: GSMA*

Forecasts show that the world market for second-hand smartphones has the potential to grow by 11% each year from 2019 to 202477. However, economic feasibility and cross-border legislative challenges hinder the secondary markets to tap into more of the value from circular devices. Currently, the economic feasibility of repairing a broken smartphone ends before the device reaches three years of use.

Affordability is another aspect that needs to be included to incentivise purchases of second-hand or circular devices. Currently, the cost of a second-hand phone is, on average, about half the cost of a new one, but the cost depreciation varies depending on model and where in the world the phone is given a new life. It is important to ensure that second-hand or circular devices are not too expensive because the most common barrier to owning a mobile phone, according to people in both low- and middle-income markets, is the cost.

The global average cost of a smartphone is around 26% of an average monthly income. However, there are large differences and the cost can be more than double in some regions. In addition, the industry must also understand and overcome logistical and legislative challenges related to exporting functional second-hand devices to enable cross-border export – for example, within the EU, where the WEEE regulation prevents e-waste from being transported across EU countries.

This also concerns second-hand devices when it comes to shipping refurbished products from one country to another. All actors that either recycle, refurbish, repair or enable the reuse of devices are retaining resources and value in the industry.

Many opportunities to enable the cascading use of devices lie in the secondary market. With a variety of offers, devices can be used several times with new consumers, ideally within the same country, but also in others. Repairers, recyclers, and other waste management operators are vital partners in the transition towards a circular economy because they enable the transitions of the material flows from being linear to become circular. Waste management operators are also important to be able to safely discard hazard-ous waste that needs to be disposed of. telecommunication operators can develop partnerships with repairers to overcome and capitalise on barriers to secondhand markets and to ensure the quality of refurbished and repaired devices – for example, by incorporating quality warranties of refurbished devices [13].

Conclusions and Recommendations

This strategy paper outlines a model telecommunications operators can use to increase circularity for devices and contribute to the vision for 2050. The model includes actions and metrics both for ‘entry’ and ‘leadership’ levels, as well as how to engage with key stakeholders in the value chain.

To implement the actions presented in this strategy paper, both company-wide and industry-wide collaboration is vital. Implementation will be strengthened and accelerated by support from senior management within and between companies. As telecommunications operators learn more about the circular economy for devices, implementation of actions must be continuously evaluated and adjusted.

This includes consideration of operators’ unique business models and markets. To contribute to the industry reaching the common vision by 2050 and to go further in the circular transition for devices, telecommunications operators will need to develop more interventions than those presented in this strategy paper. In addition, a circular transition requires a unified industry with consistent methodologies and communication [14, 15].

Even though individual companies incorporate the circular economy and work towards more circular devices, all actors within the industry and the value chain are needed to create a circular economy for devices. Aspects include, for example but not exclusively:

• B operators:

– Develop common regional take-back schemes within the industry with a focus on transparency.

– Evaluate circular services and refurb/recycling methods.

– Lead by example – apply circularity principles for own-branded CPE.

• International collaboration with actors upstream in the value chain

– For example, by deciding on a harmonised method and criteria to classify circular products both to use for ecolabelling as well as procurement criteria.

• ngaging with downstream actors such as consumers and waste operators:

– Further understand consumer incentives at national levels by conducting cross-operator consumer surveys about successful consumer incentives.

– Promote the development of industry accreditation standards for recycling stakeholders to ensure the maximum economic and sustainable benefits are consistently derived from devices.

- Within the GSMA community:

– Promote the findings of this strategy paper by launching training to put the entry and leadership actions into practice.

– Promote and facilitate a global approach to collaboration between the telecommunications industry and other actors within the industry – for example, to initiate research projects.

Acknowledgments

The GSMA would like to pay special thanks to Tele2 and Ethos for leading this Project Group, and would also like to thank the following organisations for their contribution to this strategy paper and the development of the circular model for devices.

References

[1] www.footprintnetwork.org/our-work/ecological-footprint

[2] www.ipcc.ch/report/ar6/wg2/downloads/report/IPCC_AR6_WGII_SummaryForPolicymakers.pdf

[3] pubdocs.worldbank.org/en/961711588875536384/Minerals-for-Climate-Action-The-Mineral-Intensity-of-the-Clean-Energy-Transition.pdf

[4] www.gartner.com/en/newsroom/press-releases/2022-03-01-4q21-smartphone-market-share and www.counterpoint-research.com/morethan-a-billion-feature-phones-to-be-sold-over-next-three-years/

[5] www.bankmycell.com/blog/how-many-phones-are-in-the-world

[6] www.apple.com/supplier-responsibility/pdf/Apple-Supplier-List.pdf

[7] World Economic Forum, 2019 - A New Circular Vision for Electronics, Time for a Global Reboot.

[8] The estimate has been made based on the US Environmental Protection Agency that estimates that for every million cell phones that are recycled, around 15.9 tonnes of

[9] www.bankmycell.com/blog/how-many-phones-are-in-the-world

[10] www.statista.com/statistics/786876/replacement-cycle-length-of-smartphones-worldwide/

[11] Miliute-Plepiene, J. and Youhanan, L. (2019) - E-waste and raw materials: from environmental issues to business models. IVL Swedish Environmental Research Institute.

[12] EEB. 2019. Coolproducts don't cost the earth – full report. Brussels: EEB. and Miliute-Plepiene, J. and Youhanan, L. (2019). E-waste and raw materials: from environmental issues to business models. IVL Swedish Environmental Research Institute.

[13] The calculations have been made using the estimation from the European Environmental Bureau, 2019.

[14] Assuming a typical passenger vehicle emits about 4.6 metric tons of carbon dioxide per year (EPA, 2022www.responsiblemineralsinitiative.org/about/faq/general-questions/what-are-conflict-minerals/

[15] www.oecd.org/daf/inv/mne/OECD-Due-Diligence-Guidance-Minerals-Edition3.pdf