

KNOWLEDGE GENERATION FOR STRATEGIC INVESTMENT IN STI WITH OPPORTUNITIES FOR MACHINE LEARNING AND CYBERSECURITY IN ZIMBABWE

Professor Gabriel Kabanda 

Secretary General


Zimbabwe Academy of Sciences, TREP Building, University of Zimbabwe
Harare, ZIMBABWE

Email: gabrielkabanda@gmail.com/ profgkabanda@hotmail.com

Abstract

Research creates both knowledge and technology which are put into practical use through the process of innovation. The success in achieving applied scientific technologies can be measured in the form of technological solutions, patents, inventions, published research papers, etc. The purpose of the research was to formulate an economic framework and develop technological solutions for Zimbabwe with respect to knowledge generation, innovation and enterprise development. This was compounded by an exploration for opportunities in cybersecurity and machine learning for use in the knowledge generation and dissemination business. Cybersecurity is an amalgamation of technologies, processes and operations purposed to preserve and protect computer information systems from cyber attacks or unauthorized access. Machine Learning (ML) entails the automatic data analysis of large data sets and production of models for the general relationships found among data. The Pragmatism paradigm was used as the research philosophy in this research as it epitomizes the congruity between knowledge and action. The qualitative aspect was primarily used in the knowledge generation component which was based on an integral research architecture which combines descriptive, narrative, theoretical, and experimental survey methods, through focused group discussions as the major research design. The quantitative dimension used an experiment as a research design to explore prototype models for cybersecurity and machine learning. Priority projects for strategic investment were identified for commercialization and these were on post harvest technologies; small scale mining/mineral value addition/bio mining; clean water alternatives; tiles technologies from mining waste; ICT innovations in Machine Learning and Cybersecurity; and defence technologies. A Bayesian Network model for Cybersecurity was developed to guide implementation of future cybersecurity systems in Africa. The research used the KDDCup 1999 intrusion detection benchmark dataset in order to build an efficient network intrusion detection system. The sample comprised primary data with 42 variables in a set of 494,020 instances that was analysed using mainly the SNORT open source software and other Bayesian Network supportive platforms. A Bayesian Network model was developed which took into consideration the most efficient ML algorithms.

Key Words: Knowledge generation, innovation, sustainable development, economic framework, Cybersecurity, Artificial Intelligence, Machine Learning.

 <http://orcid.org/0000-0001-6699-080X>

1. INTRODUCTION

1.1 Background

Research creates both knowledge and technology which are put into practical use through the process of innovation. Knowledge co-creation is a synergetic process of combining selected value-adding content and process from disciplinary traditions to synthesize new ways of knowing. As a management initiative, knowledge co-creation brings forth a blend of ideas and harmonisation of different parties together to jointly produce a mutually valued outcome. Innovation is the successful exploitation of an idea, renewal and enlargement of the range of products and services, the establishment of new methods of production, supply and distribution, the introduction of changes of management, work organisation. Innovation is the process of creating and putting into use combinations of knowledge from different/multiple sources to create development impact, and obtaining commercial value from inventions. Technology comprises objects, knowledge, activities, processes and a socio-technical system that comes with it (Kabanda G., 2013). The success in achieving applied scientific technologies can be measured in the form of technological solutions, patents, inventions, published research papers, etc.

The Zimbabwe National Vision 2030 is ***“Towards a Prosperous and Empowered Upper Middle Income Society by 2030, with Job Opportunities and a High Quality of Life for its Citizens”***. The ultimate goal for Vision 2030 is transform Zimbabwe to an upper middle income economy with respect to its per capita Gross National Income from the current US\$1 440 to over US\$5000 in real terms by 2030. The National Vision for Zimbabwe, enunciated as Vision 2030, provided the basis upon which the National Development Strategy was formulated and is being implemented as the two successive Five-Year National Development Strategies: NDS1 (2021-2025) and NDS 2 (2026-2030).

The era of the Internet of Things (IoT) generates huge volumes of data collected from various heterogeneous sources which may include mobile devices, sensors and social media. A hybrid cybersecurity model which uses Artificial Intelligence (AI) and Machine Learning (ML) techniques may mitigate against IoT cyber threats on cloud computing environments. Cybersecurity consolidates the confidentiality, integrity, and availability of computing resources, networks, software programs, and data into a coherent collection of policies, technologies, processes, and techniques to prevent the occurrence of an attack (Berman, D.S., et al, 2019). Cybersecurity is an amalgamation of technologies, processes and operations purposed to preserve and protect computer information systems from cyber attacks or unauthorized access (Sarker, I.H., et al, 2020). The major cybersecurity applications are intrusion detection and malware detection, which have necessitated a radical shift in the technology and operations of cybersecurity to detect and eliminate cyber threats so that cybersecurity remains relevant and effective in mitigating costs arising from computers, networks and data breaches (Sarker, I.H., et al, 2020).

Artificial Intelligence (AI) is the simulating of human intelligence in machines, through programming computers to think and act like human beings (Nielsen, R., 2015). Machine Learning (ML) is a special category of AI where computers are instructed to learn. ML is essentially the automatic data analysis of large datasets and development of models for ascertaining the relationships found among data. ML algorithms require empirical data as input and then learn from this input. The three classes of ML according to Umamaheswari, K., and Sujatha, S., (2017) are:

1. ***Supervised learning***: where the methods are given inputs labeled with corresponding outputs as training examples;
2. ***Unsupervised learning***: where the methods are given unlabeled inputs;
3. ***Reinforcement learning***: where data is in the form of sequences of observations, actions, and rewards.

Machine Learning essentially includes programming analytical model construction and is a technique of big data analytics (Napanda, K., et al, 2015). The emergence of Big data analytics as a discipline of ways of data analysis and data mining most appropriate for large datasets beyond the capability of traditional data-processing methodologies (Nielsen, R., 2015). Big Data came into existence when the traditional relational database systems were not able to handle the unstructured data generated by organizations, social media, or from any other data generating source (Mazumdar, S., and Wnga, J., 2018). In an age of transformation and expansion in the Internet of Things (IoT), cloud computing services and big data, cyber-attacks have become enhanced and complicated (Wilson, B.M.R., et al, 2015), and therefore cybersecurity events become difficult or impossible to detect using traditional detection systems (Hashem, I.A.T., et al, 2015; Siti,N.M., et al, 2017). Big Data Analytics (BDA) is rich in functionality with respect to provision of security dimensions in network traffic management, web transactions access patterns, network servers' configuration, data sources for the network, and user identity and authentication information. These activities have brought a huge revolution in the domains of security management, identity and access management, fraud prevention and governance, risk and compliance.

The supervised machine learning algorithm which can be used for both classification or regression challenges is called the Support Vector Machine (SVM). The easiest and simplest supervised machine learning algorithm which can solve both classification and regression problems is the k-nearest neighbors (KNN) algorithm. Both the SVM and KNN are applicable to the determination of optimal handover solutions in heterogeneous networks derived from diverse cells. Given a set of contextual input cues, machine learning algorithms have the capability to exploit the user context learned. The list of supervised learning algorithms includes Regression models, K-nearest neighbors, Support Vector Machines, and Bayesian Learning (Thomas, E.M., et al, 2013).

Intrusion detection involves monitoring events on networked computer systems and conducting analysis of possible intrusions or violation of various computer security policies. Network Intrusion Detection Systems (NIDS) have precipitated from the monotonic increase in the use of the internet and its associated threats. The NIDS is a type of computer software that can distinguish between the legitimate network users from malicious ones, and monitors system usage to identify behaviour breaking the security policy (Bringas, P.B., and Santos, I., 2010, p.229). NIDS exist in two categories, misuse network detectors and anomaly detectors. Misuse detection systems comprehensively invigilate all incoming network traffic and detect any sequence that appears in the knowledge base. Conversely, anomaly detection systems' focus is on detection of new unknown threats (Bringas, P.B., and Santos, I., 2010, p.229). Anomaly detection heavily relies on the use AI paradigms, and in particular on ML. Bayesian networks represent the most appropriate tool that can help us to achieve this integration of both misuse network detectors and anomaly detectors.

Bayesian Networks (BNs) are directed acyclic graphs that have an associated probability distribution function which are used as graphical probabilistic models for multivariate analysis (Bringas, P.B., and Santos, I., 2010, p.231). Boudali, H., and Dugan, J.B., (2006, p.86) define a Bayesian network simply

as a directed acyclic graph comprising nodes and arcs, where the nodes represent random variables (RV), and directed arcs between pairs of nodes represent dependencies between the RV. Furthermore, the probability function illustrates the strength of these relationships in the graph. Formally, let a Bayesian Network B be defined as a pair, $B = (D, P)$, where D is a directed acyclic graph; $P = \{p(x_1|\Psi_1), \dots, p(x_n|\Psi_n)\}$ is the set composed of n conditional probability functions (one for each variable); and Ψ_i is the set of parent nodes of the node X_i in D . The set P is defined as the joint probability density function (Bringas, P.B., and Santos, I., 2010, p.232).

$$P(x) = \prod_{i=1}^n p(x_i | \Psi_i)$$

A Bayesian network G is a probabilistic graphical model that encodes a joint probability distribution over a set of variables $X = \{X_1, X_2, \dots, X_n\}$ based on conditional independencies. This is also a directed acyclic graph (DAG) where each node represents a random variable and an edge denotes a direct probabilistic dependency between the two connected nodes (Xiao, L., 2016, p.10).

The Bayesian network accurately represents the joint probability distribution as

$$P(X_1, X_2, \dots, X_n) = \prod_{i=1}^n p(X_i | \text{PaG}(X_i))$$

where $\text{PaG}(X_i)$ denotes the set of parent nodes of X_i in G , and $p(X_i | \text{PaG}(X_i))$ specifies the conditional probability distribution (CPD) of X_i given $\text{PaG}(X_i)$ (Xiao, L., 2016, p.10).

The greatest strength of Bayesian Networks is in their ability to determine from a given a historical dataset the probability that a certain hypothesis is true. According to Margaritis, D. (2003, p.2), the reasons for choosing Bayesian networks for this type of research are that they:

1. Are graphical models, capable of displaying relationships clearly and intuitively.
2. Are directional, thus being capable of representing cause-effect relationships.
3. Can handle uncertainty.

1.2 Statement of the Problem

Zimbabwe faces a sustainable development problem with a limited capacity for industrialisation and modernisation. There is a need to create wealth and establish an innovation-led knowledge economy through breaking silos, synergising and creating smart partnerships in the National Science, Technology and Innovation System (NSTIS). The Government of Zimbabwe is faced with a limited fiscal space where its ability to mobilize resources to finance scio-economic programmes is very thin.

Firewall protection on computer systems and networks in Information Communication Technologies (ICTs) has proved to be inadequate because of gross limitations against external threats. The fact is

that the most network-centric cyberattacks are carried out by intelligent agents and combating them with intelligent semi-autonomous agents that can detect, evaluate, and respond to cyberattacks has become a requirement (Kabanda, G., 2021). The rapid development of computing and digital technologies has necessitated the need to revamp cyberdefense strategies for most organisations (Kabanda, G., 2021). Consequently, there is an imperative for security network administrators to be more flexible, adaptable, and provide robust cyber defense systems in real-time detection of cyber threats. It is of paramount importance to explore the opportunities of Machine Learning (ML) and Big Data Analytics (BDA) paradigms for use in Cybersecurity.

1.3 Purpose or Aim

The purpose of the research was to formulate an economic framework and develop technological solutions for Zimbabwe with respect to knowledge generation, innovation and enterprise development. The ultimate aim was to generate, exploit and commercialise at least 2 priority applied scientific technologies within 100 days, and then explore the opportunities in Machine Learning and Cybersecurity.

1.4 Main Research Question

How do we formulate an economic framework and develop technological solutions for the sustainable development of Zimbabwe in the context of advances in technologies such as Machine Learning and Cybersecurity?

1.5. Research Questions

- a) How is knowledge generated in the National Science, Technology and Innovation System (NSTIS) of Zimbabwe through?
- b) How do you exploit and commercialise innovative technological solutions?
- c) What are the indicators for successful knowledge generation and enterprise development in the NSTIS of Zimbabwe?
- d) How are the Machine Learning and Big Data Analytics paradigms effectively used in Cybersecurity to ensure secure ICTs?
- e) How do you develop a Bayesian Network Model that can handle the complexity in Cybersecurity?

1.6. Rationale and Justification for the Research

The vision of the African Union is “An integrated, prosperous and peaceful Africa, an Africa driven and managed by its own citizens and representing a dynamic force in the international arena” which was clearly enunciated in its Agenda 2063. The mission of the Science, Technology and Innovation Strategy for Africa (STISA-2024) is to "accelerate Africa 's transition to an innovation-led, knowledge-based economy". The STISA-2024 research and innovation priority areas for Africa are shown on Table 1 below and will focus on addressing six distinct socio-economic priorities, of which particular interest is placed on priority areas 3 (communication - physical and intellectual mobility) and 6 (wealth creation).

Table 1: STISA 2024 Priority Areas

PRIORITIES	RESEARCH AND/OR INNOVATION AREAS
1. Eradicate hunger and ensure food and nutrition security	<ul style="list-style-type: none"> ❖ Agriculture/ Agronomy in terms of cultivation technique, seeds, soil and climate ❖ Industrial chain in terms of conservation and/or transformation and distribution infrastructure techniques
2. Prevent and Control Diseases and ensure Well-Being	<ul style="list-style-type: none"> ❖ Better understanding of endemic diseases - HIV/AIDS, Malaria Hemoglobinopathy ❖ Maternal and Child Health Traditional Medicine
3. Communication (Physical and Intellectual Mobility)	<ul style="list-style-type: none"> ❖ Physical communication in terms of land, air, river and maritime routes equipment and infrastructure and energy ❖ Promoting local materials Intellectual communications in terms of ICT
4. Protect our Space	<ul style="list-style-type: none"> ❖ Environmental protection including climate change studies Biodiversity and Atmospheric Physics ❖ Space technologies, maritime and sub-maritime exploration Knowledge of the water cycle and river systems as well as river basin management
5. Live Together - Build the Society	<ul style="list-style-type: none"> ❖ Citizenship, History and Shared Values ❖ Pan Africanism and Regional Integration ❖ Governance and Democracy, City Management and Mobility ❖ Urban Hydrology and Hydraulics ❖ Urban Waste Management
6. Create Wealth	<ul style="list-style-type: none"> ❖ Education and Human Resource Development ❖ Exploitation and management of mineral resources, forests, aquatics, marines, etc. ❖ Management of water resources

The National Innovation System (NIS) comprises a set of institutions which interact and drive the innovative performance of a nation. The innovation system of any country often includes the institutions, policies, legal framework, and practices and procedures on the creation, dissemination, preservation and application of knowledge. In all these initiatives, excellence, innovation and leadership are the critical success factors. The National Science, Technology and Innovation System of Zimbabwe is illustrated by the diagram on Figure 2. However, a careful balance is required on the three facets of innovation, which are creativity, entrepreneurship and commercialisation, and diffusion and adaptation. The Government of Zimbabwe has continued to prioritise the eradication of poverty, as was the key agenda for the development blue print for the period from October 2013 to December 2018, publicly called the ‘*Zimbabwe Agenda for Sustainable Socio-Economic Transformation*’ (Zim-Asset). The Government of Zimbabwe, through the Ministry of Finance and Economic Planning of Zimbabwe is now implementing the National Development Strategy which whose implementation is being done through the two successive Five-Year National Development Strategies: NDS1 (2021-2025) and NDS 2 (2026-2030).

2. REVIEW OF LITERATURE

2.1. The Sustainable Development Goals (SDGs) Context

The Research Council of Zimbabwe (RCZ) appointed a team of prominent scientists led by the researcher from the Zimbabwe Academy of Sciences to embark on a ground-breaking research programme purposed "to generate, exploit and commercialise at least 2 priority applied scientific technologies within 100 days". The Knowledge Generation research priority areas for Zimbabwe are informed by the Sustainable Development Goals (SDGs), the African Science, Technology and Innovation Strategy for Africa 2024 (STISA 2024) priority areas, Vision 2030 and the national research priority areas. The United Nations published the SDGs where SDG Goal 9 is about the need to “*build resilient infrastructure, promote inclusive and sustainable industrialisation and foster innovation*” (<http://www.un.org/sustainabledevelopment/sustainable-development-goals/>). Zimbabwe as a country prioritised goals 2,3,4,5,6,7,8,9,13 and 17. However, Zimbabwe as a nation must not be viewed as just gravitating from one developmental guideline to another without making any meaningful progress during each dispensation.

The vision for the Zimbabwe Academy of Sciences (ZAS) is “*The Zimbabwe Academy of Sciences seeks to be the leading catalyst for knowledge-sharing, innovative solutions, evidence-based policy formulation and advisory services in Zimbabwe, Africa and beyond*”. The Mission states that ZAS exists to “*Monitor the environment, identify problems and opportunities, provide and communicate the ultimate evidence-based solutions that benefit society for sustainable development by mobilising the science community and other resources through smart partnerships with government, academia, private sector, development partners and civil society*”. However, there is need for self-renewal, task-oriented, relevancy, agility, flexibility and consistency in the renewed Mission to develop innovative solutions to address Zimbabwean challenges and strategically advance Zimbabwe to be a global power. The ZAS guiding philosophy is about mutual respect and quality, stated clearly as “*Mutual respect and equality is important because my humanity is bound up with yours*”. This is buttressed by the core values of Innovativeness, Integrity, Professionalism, Reliability, Institutional Independence, Respect and Ethics. At a national level, ZAS is desirous to provide national leadership on scientific initiatives and innovations in key areas that include heritage studies, water and sanitation, climate change, sustainable environmental management, national security, etc., as guided by the national research priorities and key projects of national significance. The national research priorities are as follows:

1. Social Sciences and Humanities

- ❖ Fiscal Reform Measures
- ❖ Public Administration, Governance and Performance Management
- ❖ Strengthening social and economic fabric
- ❖ Strengthening policy making processes
- ❖ Social Services and Poverty Eradication
- ❖ Culture and Heritage
- ❖ Creative and Cultural Industries
- ❖ Regional and world cultures

2. Sustainable Environmental and Resource Management

- ❖ Food Security and Nutrition
- ❖ Land
- ❖ Water
- ❖ Minerals
- ❖ Aerospace and other sensing technologies
- ❖ Energy
- ❖ Sustainable livelihoods
- ❖ Transforming Agriculture
- ❖ Value addition and Beneficiation to Natural Resources
- ❖ Impact of new technologies to sustainability and / or resource management.

3. Promoting and Maintaining Good Health

- ❖ Revitalising Health Delivery Systems
- ❖ Increasing access to health facilities
- ❖ Preventive Health Care
- ❖ Food Security and Nutrition
- ❖ Social Services and Poverty Eradication

4. National Security

- ❖ Geo-Information Sciences
- ❖ Terrorism and crime
- ❖ Transformational defence technologies
- ❖ Invasive species, diseases and pests
- ❖ Infrastructure and Utilities

The new priority areas that require special attention include the following:

1. Natural and Cultural Heritage
2. Indigenous Knowledge Systems
3. Post-harvest technologies
4. Rural transformation
5. Small scale mining/mineral value addition/bio mining
6. Clean water alternatives
7. Tiles technologies from mining waste
8. Cyber security systems
9. Defence technologies (double use technologies (drones))

The 17 Sustainable Development Goals (SDGs) describe the major development challenges for humanity in order to secure a sustainable, peaceful, prosperous and life of equitability, and these are depicted by the diagram below on Figure 1. Zimbabwe prioritised Goals 2,3,4,5,6,7,8,9,13 and 17. At the core of the 2030 Agenda are 17 Sustainable Development Goals (SDGs) which describe the major development challenges for humanity in order to secure a sustainable, peaceful, prosperous and life of equitability. Humanity needs to build peace and drive sustainable development. The world faces a sustainable development problem with a limited capacity for industrialisation and modernisation in developing nations.

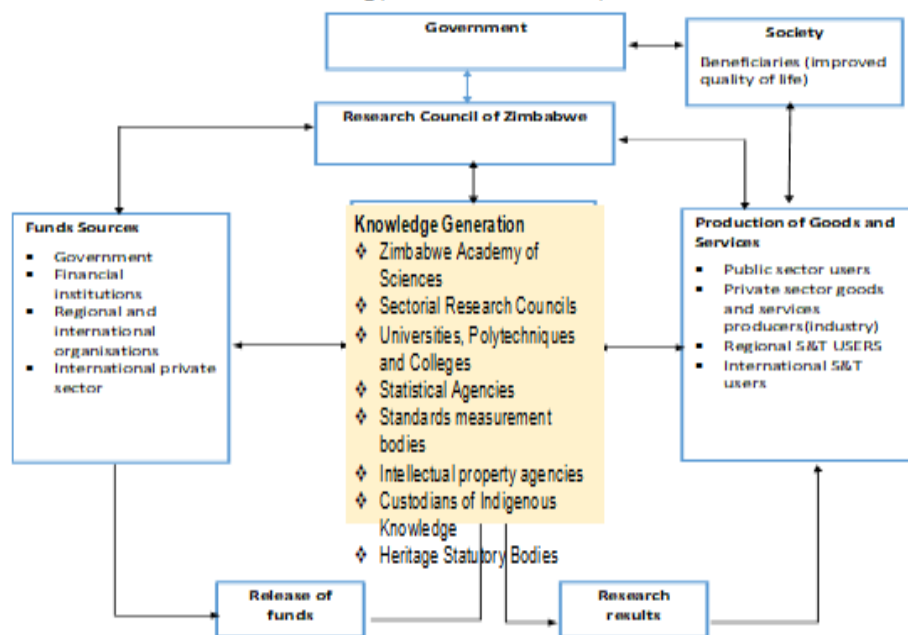
The African Union has a clearly defined Mission on Science, Technology and Innovation (STI), which was considered by Zimbabwe as an African country. The African Union, through its Agenda 2063, desires a prosperous and peaceful Africa. The Science, Technology and Innovation Strategy for Africa (*STISA-2024*) is to *"accelerate Africa 's transition to an innovation-led, knowledge-based*

economy''. At a national level, there is a need to create wealth and establish an innovation-led knowledge economy through breaking silos, synergising and creating smart partnerships in the National Science, Technology and Innovation System (NSTIS).



Figure 1: The United Nations Sustainable Development Goals

The Science, Technology and Innovation System of Zimbabwe



The Science Technology and Innovation System of Zimbabwe – A Schematic View

Adapted from Musize, S (2009)

Figure 2: Zimbabwe's National Science, Innovation and Technology System (Source: Muzite, S., 2009)

The National Science, Technology and Innovation System of Zimbabwe is summarised by the diagram on Figure 2. Scientific Knowledge Generation in Zimbabwe is led by ZAS and involves various stakeholders that include Zimbabwe Academy of Sciences, Sectorial Research Councils, Universities and Colleges, Statistical Agencies, Standards Measurement Bodies, Public Laboratories, Research Centres, Private Laboratories, Intellectual Property Agencies, Custodians of Indigenous Knowledge, Heritage Statutory Bodies, etc. Zimbabwe Academy of Sciences works in partnership with a number of national institutions to meet its objectives and goals. These institutions include the Government institutions, research institutions, Universities, parliamentarians, etc.

The basic driving force behind economic growth is technological change, where the main catalyst is investment on research and development. A comparison is made between two countries, Israel and Zimbabwe, where Israel has made extensive investment in R&D, as shown on the Table 2 below (Kabanda G., 2013).

Table 2: Effects of R&D to economic performance

	Zimbabwe	Israel
Population	14,627,000	7,700,000
Size	200,000 SQ km	20,000 SQ km
GDP - Per Capita	<i>\$1,530</i>	<i>\$31,004</i>
Infant Mortality Rate	79 dead per 1000 live births	<i>4 dead</i> per 1000 live births
Life Expectancy	<i>50</i>	<i>81</i>

It is noted that Israel, which has half the population of Zimbabwe and a geographical land ten times smaller than that of Zimbabwe, has invested extensively in Research and Development (R & D) over the years and that its economy has grown to be in the league of developed nations and now boasts of a GDP per capita of \$31,004 and life expectancy of 81 years whilst Zimbabwe remains with a GDP per capita of \$1,530 and life expectancy of 51 years, respectively. Israel, which has half of its land in a desert, now exports more oranges and agricultural products than Zimbabwe and South Africa put together because of the huge investments in R&D. How does Zimbabwe improve its average life expectancy of 50 years to the life expectancy of Israel of 81 years, or improve its GDP per capita from just \$1,530 to \$31,004 like Israel? In Southern Africa (SADC Region), high GDP per capita are evidenced in Seychelles which has a GDP of \$16,434, Mauritius with a GPD per capita of \$11,228 and Botswana has \$8,258. South Africa has a GDP per capita of 6,354 and Namibia \$6,013.

The GDP per capita analysis for Southern Africa as of August 2021 is shown on the table 3 below.

Table 3: SADC Regional GDP per capita - August 2021

Country	Population	Annual GDP (US\$)	GDP per capita (US\$)
Angola	30,809,762	105, 902M	3,437
Botswana	2,254,126	18,615M	8,258
DRC	84,068,091	47,099M	560
Lesotho	2,108,132	2,739M	1,299
Madagascar	26,262,368	13,853M	528
Malawi	18,143,315	7,065M	389
Mauritius	1,266,000	14,210M	11,228
Mozambique	29,495,962	14,396M	488
Namibia	2,414,000	14,513M	6,013
Seychelles	96,762	1,590M	16,434
South Africa	57,939,000	368,135M	6,354
Swaziland	1,136,191	4,711M	4,146
Tanzania	56,318,348	56,852M	1,009
Zambia	17,351,822	26,720M	1,540
Zimbabwe	14,439,018	20,401M	1,530

The Zimbabwe Academy of Sciences (ZAS) was established in October 2004 following research work done by the Research Council of Zimbabwe with the purpose to provide independent evidence-based advice to the government and the nation at large on addressing national challenges using scientific knowledge and innovative expertise; and to recognize, honour, and perpetuate the achievements of those Fellows of ZAS who have made immense contributions to the scientific development of Zimbabwe and the rest of the world and have helped to bring recognition, honour, distinction, and excellence to science, technology, engineering, and mathematics (STEM) related programmes, projects and research.

The Knowledge Generation work for the NSTIS of Zimbabwe is guided by Goal 9 of the globally defined 17 Sustainable Development Goals (SDGs), which largely inform some of the national developmental programmes, are listed in Table 4 below. Zimbabwe as a country prioritised goals 2,3,4,5,6,7,8,9,13 and 17.

Table 4: The 17 Sustainable Development Goals (SDGs)

(<http://www.un.org/sustainabledevelopment/sustainable-development-goals/>)

Goal 1	End poverty in all its forms everywhere
Goal 2	End hunger, achieve food security and improved nutrition and promote sustainable agriculture
Goal 3	Ensure healthy lives and promote well-being for all at all ages
Goal 4	Ensure inclusive and equitable quality education and promote lifelong learning opportunities for all
Goal 5	Achieve gender equality and empower all women and girls
Goal 6	Ensure availability and sustainable management of water and sanitation for all
Goal 7	Ensure access to affordable, reliable, sustainable and modern energy for all
Goal 8	Promote sustained, inclusive and sustainable economic growth, full and productive employment and decent work for all
Goal 9	<i>Build resilient infrastructure, promote inclusive and sustainable industrialization and foster innovation</i>
Goal 10	Reduce inequality within and among countries
Goal 11	Make cities and human settlements inclusive, safe, resilient and sustainable
Goal 12	Ensure sustainable consumption and production patterns
Goal 13	Take urgent action to combat climate change and its impacts*
Goal 14	Conserve and sustainably use the oceans, seas and marine resources for sustainable development
Goal 15	Protect, restore and promote sustainable use of terrestrial ecosystems, sustainably manage forests, combat desertification, and halt and reverse land degradation and halt biodiversity loss
Goal 16	Promote peaceful and inclusive societies for sustainable development, provide access to justice for all and build effective, accountable and inclusive institutions at all levels
Goal 17	Strengthen the means of implementation and revitalize the global partnership for sustainable development

Capital for the infrastructure and equipment and extensive investments into people (labour) are required in the ICT revolution (Kabanda G., 2008). The revolutionary technological change or productivity levels from an ICT perspective is related to labour and capital by the Cobb-Douglas production function in the form:

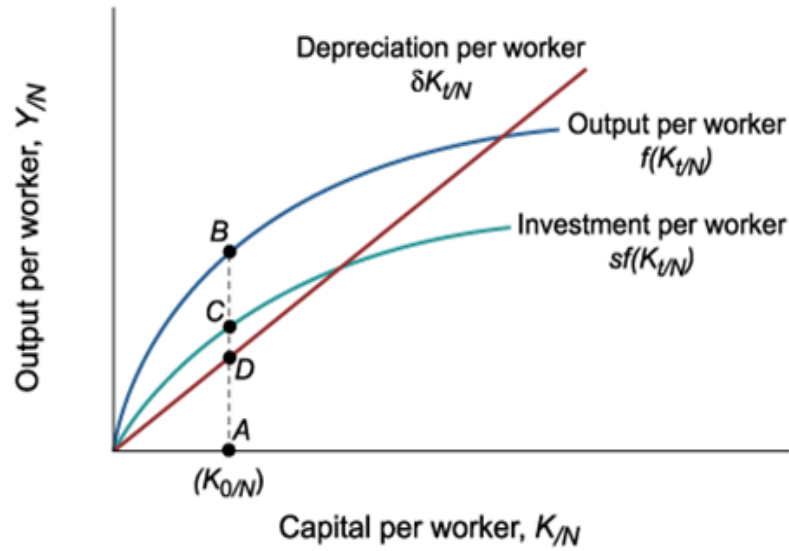
$$Q = A K^a L^b$$

is used for the analysis of technological progress and attended economic growth, where A , a and b are empirical parameters.

- K = capital input (very meaningful amounts)
- L = labour input (high technical competence)

Production capacity can be multiplied several times more through an investment in technology. Technological change is the basic driving force behind economic growth. Technological change is determined by deliberate activities of economic agents in response to the financial incentives, and so is endogenous. High-skilled workers enhance technological innovations and their diffusion, and so high-skilled labour is complementary with capital and low-skilled labour. The Brain Drain problem is a result of a continuous outflow of high – skilled labour from a country. An endogenous neoclassical economic growth model is illustrated by the diagram below which relates the output per worker to the capital per worker in how it relates to the investment per worker and the output per worker.

Neoclassical Endogenous Growth Model



Source: Jones, 1988, Chapter 2, Solow Neoclassical Growth Model

Figure 2: Neoclassical Endogenous Growth Model

2.2. Classical Machine Learning (CML)

Machine Learning (ML) is a field in artificial intelligence where computers learn like people. We present and briefly discuss the most commonly used classical machine learning algorithms, as shown on Table 5.

Table 5: The Classical Machine Learning Algorithms

Name of ML Algorithm	Description
1. Logistic Regression (LR)	As an idea obtained from statistics and created by Sit, N.M., et al (2017), logistic regression is like linear regression, yet it averts mis-classification that may occur in linear regression but its results are basically either '0' or '1'.
2. Naive Bayes (NB)	Naive Bayes (NB) classifier is premised on the Bayes theorem which assumes independence of features and overcomes the curse of dimensionality.
3. Decision Tree (DT)	A Decision tree has a structure like flow charts, where the root node is the top node and a feature of the information is denoted by each internal node.

4. K-Nearest Neighbor (KNN)	K-Nearest Neighbor (KNN) is a non-parametric approach which uses similarity measure in terms of distance function classifiers other than news cases, and this stores the entire training data, requires larger memory and so is computationally expensive.
5. Ada Boost (AB)	Ada Boost (AB) learning algorithm is a technique used to boost the performance of simple learning algorithms used for classification.
6. Random Forest (RF)	Random forest (RF), as an ensemble tool, is a decision tree derived from a subset of observations and variable, and gives better predictions than an individual decision tree.
7. Support Vector Machine (SVM)	Support Vector Machine (SVM) can be used to solve classification and regression problems, and belongs to the family of supervised machine learning techniques.

2.3 Modern Machine Learning

Deep learning has the capability of taking raw inputs and learning the optimal feature representation implicitly.

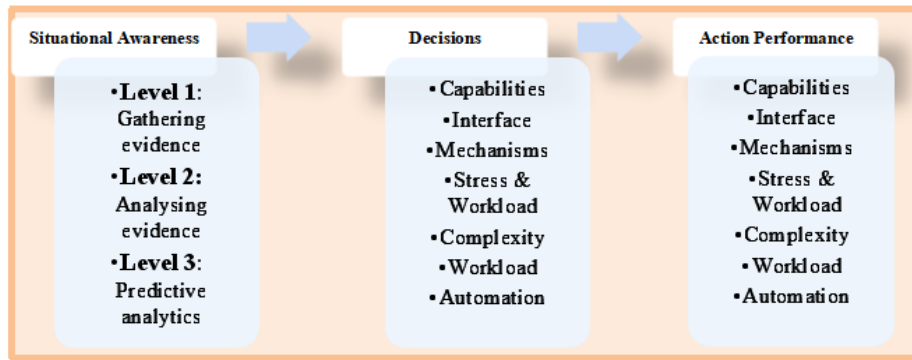
2.2.1 Deep Neural Network (DNN)

An artificial neural network (ANN) as is characteristics of biological neural networks. The family of ANN includes the Feed forward neural network (FFN), Convolutional neural network and Recurrent neural network (RNN). The traditional examples of machine learning algorithms include Linear regression, Logistic regression, Linear discriminant analysis, classification and regression trees, Naïve bayes, Support Vector Machines (SVM), K-Nearest Neighbour (K-NN), Kmeans clustering Learning Vector Quantization (LVQ), Monte Carlo, Random Forest, Neural networks and Q-learning.

2.2.2 The future of AI in the fight against cybercrimes

Big Data Analytics requires new data architectures, analytical methods, and tools. Threat intelligence is the process purposed to gather threats from big data, analyze and filter information about these threats and create an awareness of cybersecurity threats (Sarker, I.H, et al, 2020). The situation awareness model consists of situation awareness, decisions and action performance as shown in Figure 3. There is consensus in prior literature that cybersecurity has evolved to become a problem for big data analytics. Further, even the data mining models that have been used in the past are no longer sufficient for the challenges in cybersecurity (Hashem, I.A.T., 2015). A big data analytics model for cybersecurity can be evaluated on the basis of its agility and robustness (Hashem, I.A.T., 2015).

FIGURE 3: SIMPLIFIED THEORETICAL MODEL BASED ON SITUATION AWARENESS



2.3 Cybersecurity in Network Intrusion Detection and Prevention System

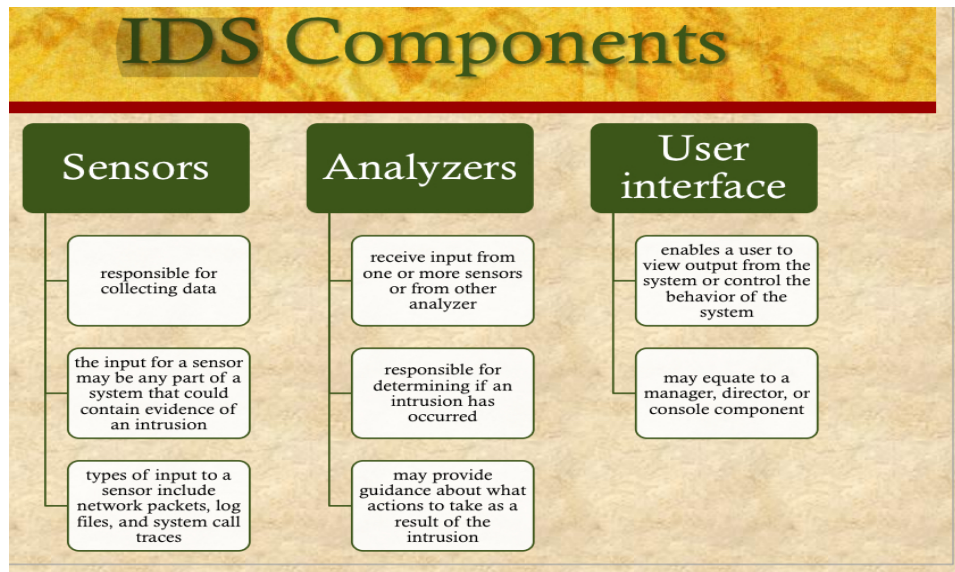
Infocommunication is a natural scientific discipline that studies the structure of objects and the process of interaction between these objects whose objective is the construction of formalized models of data structures and data transmission from one object to another (Kuznetsov, N.A., 2005, p.1). The digital convergence process initially affected the information technology and telecommunications sectors by amply manifesting the unification of the technologies, the integration of their markets and the harmonization of their regulation (Sallai, G., 2012, p.2). Accordingly, the various contents have been associated with separated networks, services and user terminals and their markets and regulation have been separately managed. The term information and communications technology (ICT) is generally used and usually refers to the integration of information and telecommunication technology sectors involving their convergence with the media technology sector based on common digital technology.

Information systems security comprises computer and communications security dimensions. The weakest link in Cybersecurity determines its overall strength (Nielsen, R., 2015, p.8). Access controls and security mechanisms must be clearly enunciated in the company objectives. Its important to give the employees internet access only for the purposes which are of great importance to the organization. However, the privileges given need to be constantly monitored, especially when accessing from outside the company premises. All network traffic in network security should be redirected through a single point and only open the ports on the firewall necessary for business traffic. The network configuration can be strengthened by the provision of VPN support (Nielsen, R., 2015, p.18). Both an Intrusion Detection Systems (IDS) and Intrusion Prevention Systems (IPS) are required in network security. However, organizational policies should spell out the procedures for handling information security, with some legal assistance. As shown on Figure 4 below, the Intrusion Detection System (IDS) usually operates with sensors, analyzers and a user interface. The policies should cover the following areas (Nielsen, R., 2015, p.14):

- ◆ Personal Electronic Devices (PED)
- ◆ Acceptable Use
- ◆ Records Retention
- ◆ Identity Protection
- ◆ Server, Service and Project Computing Security
- ◆ Data Encryption

The IDS can either be network-based or host-based.

Figure 4: The IDS Components [\(Source: Stallings, W., 2015, p.6 \)](#)



A firewall provides network security against external threats and is essentially a computer server that interfaces with external computer systems with a mechanism to protect sensitive files on computers within the network (Stallings, W., 2015, p.10). Operating Systems Hardening secures an operating system (Stallings, W., 2015, p.28), and involves installing and patching the operation system, and then hardening and/or configuring the operating system to protect the system by:

- ◆ removing unnecessary services, applications, and protocols
- ◆ configuring users, groups and permissions
- ◆ configuring resource controls
- ◆ additional security controls installation and configuration.

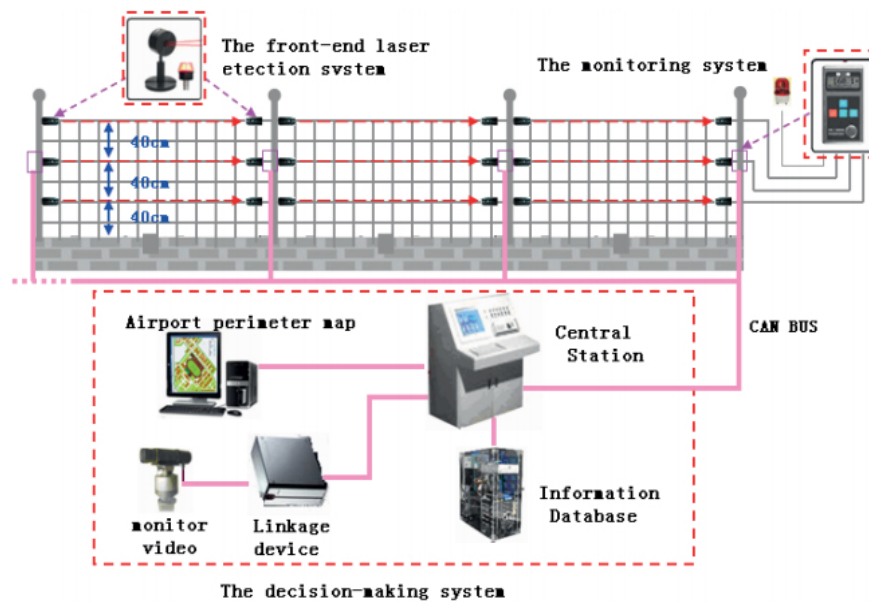
The Bayesian network creation and setup comprises the following phases:

- ❖ Traffic sample obtaining to establish the information source in order to gather the sample
- ❖ Structural Learning, which defines the operational model
- ❖ Parametric Learning of the quantitative model
- ❖ Bayesian Inference
- ❖ Adaptation.

Next generation access networks are expected to support an increased number of users, increased bandwidth demand and longer-range coverage. Optical fiber as the prevailing solution for next generation fixed access network has been adopted and deployed worldwide. Network intrusion and the probability of risk can be adequately handled by a probability model. Principal component analysis (PCA) method is applied to preprocess the network signal to avoid the problem of denoising methods involving the use of low (high) pass filter (Wei and Liu, 2016). The network signal is transformed into a new coordinate system by the orthogonalized linear transformation through making the first variance of the data projected at the first and second coordinates (referred to as the first principal component and the second principal component, respectively). PCA can eliminate noise from the background environment and reduce the dimension of the network signal collected on the receiving device (Wu, 2018, p.2).

The current detection technologies for airport perimeter security usually include infrared detection, vibration cables, underground cables, microwave detection, video surveillance, tension fencing, and other technologies, but it is rare that laser detection technology is used (Wu *et al*, 2016, p.1). In consideration of the airport surroundings, Wu *et al* (2016) divided the laser anti intrusion security system into three parts: the decision-making system, the monitoring system, and the front-end laser detection system, as shown on Figure 5.

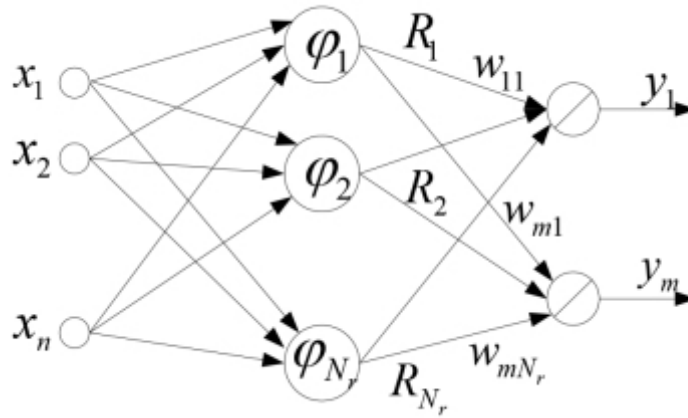
Figure 5: The antiintrusion laser alarm system [\(Source: Wu et al 2016, p.2\).](#)



Intrusion incidents at airport perimeters currently uses pattern recognition methods comprising use of

artificial intelligence, fuzzy theory, expert systems, neural networks, genetic algorithms, and other relevant methods. Artificial neural networks have emerged to possess the unique capability of processing nonlinear information and distributed storage ways for information (Wu *et al*, 2016, p.6). The radial basis function (RBF) neural network is constructed as a two-forward type neural network whose value for RBF is determined by the intermediate layer node's output, as shown on the network model on Figure 6.

Figure 6: The structure of the Radial Basis Function (RBF) neural network [\(Source: Wu et al, 2016, p.6\)](#)



$X = \{x_1, x_2, \dots, x_n\}$ are n -dimensional input vectors and the output of the hidden layer nodes are the RBF values. The input mapping to a new space is provided by the hidden layer unit which performs a nonlinear transformation. RBF is essentially a Gaussian function which is expressed as follows (Wu *et al*, 2016, p.6):

$$R_j = \varphi_j(X) = e^{-\|X - C_j\|^2 / (2\sigma_j^2)}, \quad j = 1, 2, \dots, N_r$$

Network intrusion detection systems were developed to detect any network attack. The attacks or malicious behavior can be determined from an analysis of packet contents of the network. However, the packet inspection is a complex resource-hungry process usually impossible to attain. Karimpour *et al* (2016, p.1) reveal the attacks by combining flow-based and graph-based procedures. Karimpour *et al* (2016, p.2) categorized the general overview of intrusion detection approaches in four approaches as follows:

- 1) *Feature-based approaches*: these approaches are based on the concept of similar graphs sharing common attributes inclusive of diameter, eigenvalues, and a distribution of degree. These methods can be used for checking the structure of a graph in order to find patterns and explore anomalies.
- 2) *Decomposition-based approaches*: use tensor decomposition and graph structure to interpret eigenvectors and convergence of graph attributes to find the patterns, respectively.

3) *Community-based approaches*: the main action in these approaches is graph clustering where the clustering algorithms are employed to create cluster parameters of data, and the anomalies are recognized based on their values.

4) *Window-based approaches*: in this category, the patterns of the evolutionary behavior are revealed by the time intervals, and thereby determine whether the behavior of the network is a normal or malicious case.

A general view of these intrusion detection methods according to the above 4 categories are shown on Table 4 below.

Table 4: Anomaly detection methods [\(Source: Karimpour et al \(2016, p.3\)\)](#)

Method	Data type	Attack	Proposed system	Accuracy
Graph in time series	Flow-based	DDoS	Graph-based	94.2%
Dispersion graph	Flow-based	DDoS	Graph-based	100%
Using flow concept	Flow-based	Dictionary	Flow-based	99%
Graph clustering and local deviation coefficient	Packet-based	DoS, Scan	Graph-based	95.3%
Graph clustering and local deviation factor				
Packet heard analyzing	Packet-based	DoS, Scan	Packet-based	95.4%

An attack in the network was detected by using the flow and graph-clustering concepts by Karimpour *et al* (2016, p.3) in a manner that reflected the nodes, the edges, and the weight of edges through the IPs, the flows, and the number of flows in the graph, respectively. The anomaly points could be detected from the average weight of clusters that are reached from the graph-clustering algorithm and comparing it in several time intervals and threshold points. The outcome of the research by Karimpour *et al* (2016, p.4) involved 7 weeks of network traffic and 5 types of attacks: DoS, scan, local access, user to root, and data, which are shown on Table 5 below, indicating the number and types of attacks in each categorized attack.

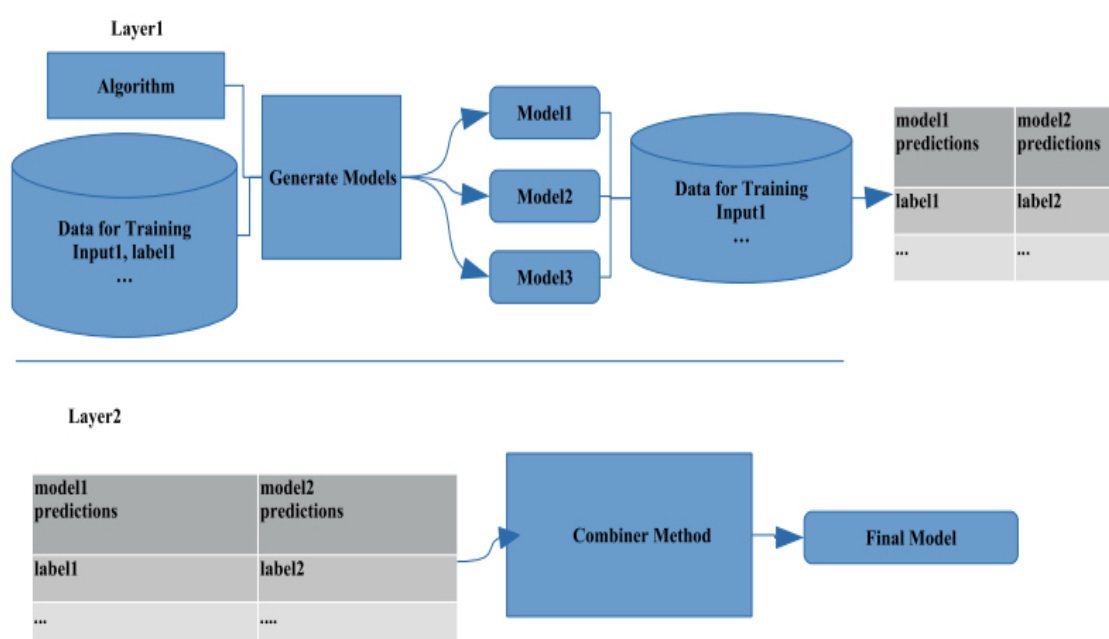
Table 5. Various attack types and their descriptions [\(Source: Karimpour et al, 2016, p.4\)](#)

Attack type	Description
DoS	Denial of service; an attempt to make a network resource unavailable to its intended users: temporarily interrupt services of a host connected to the Internet
Scan	A process that sends client requests to a range of server port addresses on a host to find an active port
Local access	The attacker has an account on the system in question and can use that account to attempt unauthorized tasks
User to root	Attackers access a user account on the system and are able to exploit some vulnerability to gain root access to the system
Data	Attackers involve someone performing an action that they may be able to do on a given computer system, but that they are not allowed to do according to policy

When the cluster-based data are given as input to the model, the final model of attack detection can be developed, and the proposed criterion is then calculated in the time series based on defined threshold points. Accordingly, the best threshold is identified within the time series and extracted from the detection rates of the suggested way (Karimpour *et al*, 2016, p.5). Li (2018) developed a collaborative intrusion detection method with data mining techniques for a marine distributed network. An efficient The marine distributed network intrusion model created was very efficient, used less memory space and had a detection rate above 92%. The marine distributed network provides guarantees for the normal sailing of a ship as used in a ship navigation system. ML can provide researchers with opportunities to detect network intrusion without using a signature database. Demir and Dalkilic (2017) improved the model generation and selection techniques by using different classifications algorithms as a combiner method. Model generation was performed using subsets of the dataset with randomly selected features and obtained better accuracy levels than pure machine learning techniques. In comparison with other studies, the study obtained the highest detection rate for user-to-root attacks.

In ML the same classification or regression problem is solved by ensemble learning whose methods develop a set of models which are then combined. Demir and Dalkilic (2017) established that weak learners could be assisted to become strong learners. Boosting, bootstrap aggregating, and stacking are the three most commonly used types of ensemble techniques (Demir, N., and Dalkilic, G., 2017, p.1). Bootstrap aggregating is when each model is trained by drawing random subsets of the training set. The random forest algorithm combines random decision trees and uses bagging. Boosting incrementally builds an ensemble model by using the misclassified training instances that previous models misclassified for training each new model. Stacking, also known as stacked generalization, is a method where an algorithm is used. Stacking is the generalization of other ensemble methods and involves using an algorithm to to combine the outputs of other models' predictions. Models are generated from the training algorithm using data. The “model generation” phase generates n models during the stacking implementation by using the algorithm with randomly drawn sub-datasets. A two-layered training phase consists of a training set with each algorithm and then the predicted labels of each model, as shown on Figure 7 below.

Figure 7: Training phase of tracking approach [\(Source: Demir, N., and Dalkilic, G., 2017, p.4\)](#)



Demir and Dalkilic (2017) developed a threat model that collects information on the packet level. The following four assumptions may possibly occur:

- ◆ The attacker can exploit various vulnerabilities of the applications running on the target host and get access to a user right,
- ◆ The attacker already can gain root access by exploiting various vulnerabilities of the applications running on the target host,
- ◆ The attacker can the vulnerabilities of the applications running on the target host to launch a denial of service (DoS) attack, and
- ◆ The target host is probed by an attacker with various techniques to gain information.

The IDPS components must first and foremost be secure since it is the primary target of attackers who try to prevent the IDPSs functioning of detecting attacks or to access the sensitive data on IDPSs like host configuration and known vulnerabilities.

3. RESEARCH METHODOLOGY

3.1. Overview

The research philosophy, methodology and research design were guided by the research onion shown on Figure 8 below. The Pragmatism paradigm was used in this research and this is intricately related to the Mixed Methods Research (MMR). Knowledge Generation for Strategic Investment in STI with opportunities for Machine Learning and Cybersecurity is a huge area for consideration and in order to address problems within it, there is need for contextualisation.

The Research methodology is a way of solving a research problem thoroughly and meticulously and includes steps followed in carrying out the research and the reasoning behind (Kotari, C.R., 2004). The Mixed Methods Research methodology was used and underpinned by the pragmatic paradigm. The researcher adopted mainly a qualitative approach in the form of focus group discussion for the knowledge generation stage and followed by the quantitative approach with the experimental research design that involved the development of a Bayesian Network Model for Cybersecurity using the Snort platform. The researcher adopted a descriptive research design because of the need to systematically describe the facts and characteristics of big data analytics models for cybersecurity. The purpose of the study was essentially an in-depth description of the models (Burt, D., et al, 2013). The researcher adopted a postmodern philosophy to guide the research. The researcher noted that the definition, scope and measurement of cybersecurity differs between countries and across nations (Wilson, B.M.R., et al, 2015). Prior research has tended to use case studies in relation to the study of cybersecurity (Wilson, B.M.R., et al, 2015).

The research onion

(Saunders et al., 2009:138)

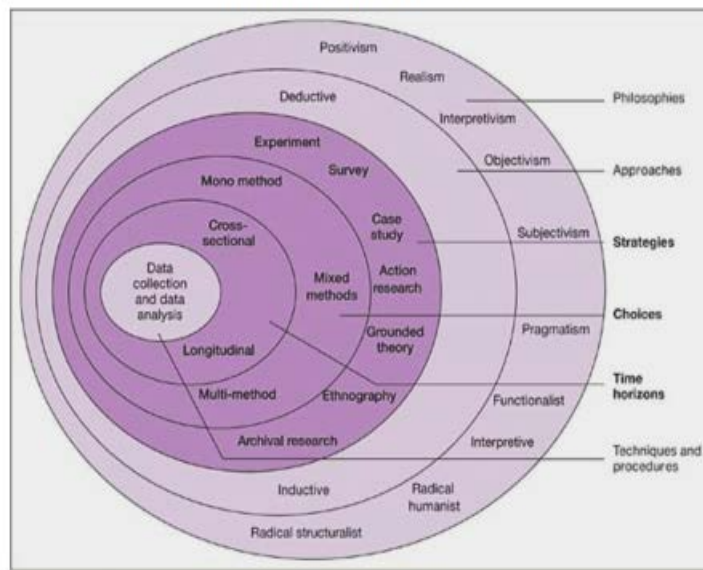


Figure 8. Research onion

The research methodology being used in the research Programme on Knowledge Generation is largely qualitative and based on an integral research architecture which combines descriptive, narrative, theoretical, and experimental survey methods, through focused group discussions as the major research design. The integral research architecture, which is illustrated in Figure 9 below, uses a combination of descriptive methods, experimental and survey methods, methods of theorising, and narrative methods. These methods are related to the four (4) human modes (being, doing, knowing, and becoming), respectively. The core methods used in integral research methods are (empirical phenomenology (descriptive methods), storytelling (narrative methods), grounded theory (methods of theorising) and case study (experimental and survey methods)).

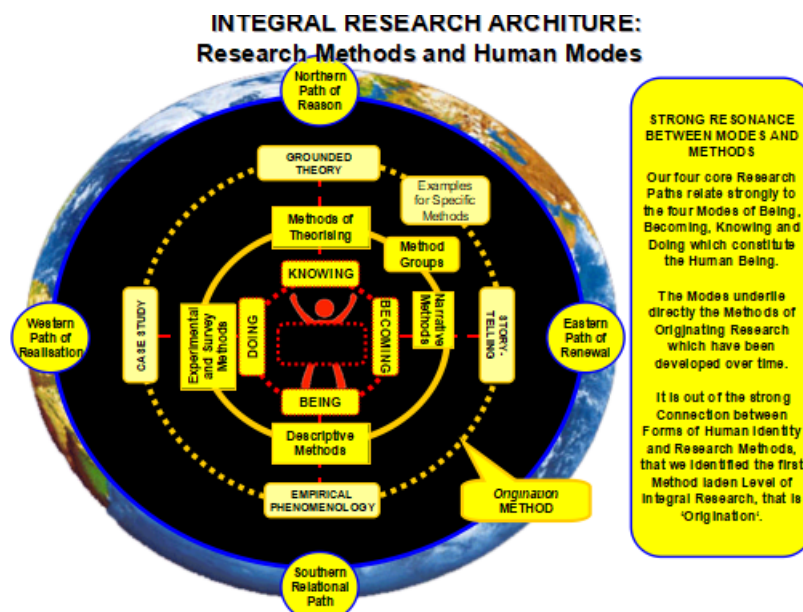


Figure 9: Integral Research Architecture

The research design used was a Focus Group discussions. A Focus Group was setup from the 22 Knowledge Generation institutions in Zimbabwe which participated in the National Indaba on the National Science, Technology and Innovation System of Zimbabwe held on 4th October, 2017 at the HICC, Harare. Focus Group discussions were held by this Working Group on Knowledge Generation every week from 4th October until end of November, 2017. Knowledge generation in Zimbabwe is primarily done by the following entities:

- ❖ Zimbabwe Academy of Sciences
- ❖ Sectorial Research Councils
- ❖ Universities and Colleges
- ❖ Statistical Agencies
- ❖ Standards Measurement Bodies
- ❖ Public Laboratories
- ❖ Research Centres
- ❖ Private Laboratories
- ❖ Intellectual Property Agencies
- ❖ Indigenous Knowledge Systems

The Work Plan for the Working Group on Knowledge Generation (the Focus Group) used this research programme was developed and is as shown below on Table 6, indicating the action steps for each milestone.

Table 6: Work Plan for Knowledge Generation

Milestone	Action step
1 Knowledge Prioritisation	a) Identification of priority areas b) Engagement of institutions c) Identification of two technologies d) Map way forward for commercialisation
2 Advocacy of the National Science Technology and Innovation System of Zimbabwe	a) Develop Communication and publicity Framework b) Interfacing with the media for publicity
3 Knowledge co-creation and collaboration	a) Institution to institution collaboration b) Interministerial collaboration c) Institution to company collaboration d) Country to Country collaboration
4 Knowledge Acquisition	a) Diaspora engagement b) Formulate skills exchange programme in identified priorities c) Identify expired patents for utilisation
5 Knowledge Enterprenuerising and Commercialisation	a) Develop and adopt solid ideation prototype process (SIPP) in 10 days b) Establish and maintain a national support (NSN)network in 10 days c) Develop and disseminate a national scheme to promote, incentives and recognise innovation

	d) Set up and maintain database of ongoing project works
6 Fusion and Adaption	a) Literature review b) Identification of technologies that will address priority areas c) Identify relevant local prayers who can take up the technologies d) Creation of database for commercially viable science e) End user experimentation of identified technologies for adaption f) Development of technology usage tracking mechanism

The Knowledge Generation Working Group (Focus Group) is guided by the following governance structure at the national level.

- ❖ Political leader (Chief Secretary - OPC)
- ❖ Sponsor (Deputy Chief Secretary)
- ❖ Results Leader (RCZ Board Chair)
- ❖ Strategic Leader (Executive Director of RCZ)
- ❖ Team Leader/ Deputy Team Leader
- ❖ Team Secretary
- ❖ Team Members
- ❖ Coach

3.2. Quantitative data collection of the KDD'99 data set for the development a Bayesian Network Model

The research used the KDDCup 1999 intrusion detection benchmark dataset in order to build an efficient network intrusion detection system. The population was the primary data obtained from <http://kdd.ics.uci.edu/databases/kddcup99/kddcup99.html> and comprised about 10 million records with 42 variables (attributes). The data was obtained from the archived source at UCI KDD Archive, Information and Computer Science, University of California, Irvine, Irvine, CA 92697-3425. From this population, a sample of 494,020 records with 42 instances was selected for data analysis.

3.3. Population, sampling and Model for analysis

The researcher gets the required information from a selected sample size of respondents (Kumar, R., 2011). The full set of cases from which a sample is taken from constitute the population (Saunders, et al, 2009). Population signifies the full set which the researcher wishes to study. According to Saunders, et al (2009), the comprehensive list of members of the population from which a sample is drawn is referred to as the sampling frame. The Researcher used the Yamane's formula to compute the sample size using a 95% confidence level (Saunders et al, 2009), since the population size is finite.

The KDD'99 Dataset with 494,020 intrusion detection records was the sample and the purposive sampling method was used. In probability sampling, each item has a nonzero chance of being has an equal probability of being selected from the population. Nonprobability sampling does not give all the participants or units in the population equal chances of being included. When the researcher faces

challenges of limited resources, time and workforce, the nonprobability sampling method would be most appropriate, and it can also be used when the research does not aim to generate results that will be used for generalizations of the entire population. The purposive sampling technique, also called judgment sampling, is the deliberate choice of a participant due to the qualities the participant possesses (Etikan, I., 2016, p.2).

By virtue of knowledge or experience, the researcher decides what needs to be known and sets out to find people who can and are willing to provide the required information. According to Etikan (2016, p.2), purposive sampling is typically used in qualitative research to identify and select the information-rich cases for the most proper utilization of available resources, and often involves selecting targeted groups or individuals as data sources. The Maximum Variation Sampling (MVS) type of purposive sampling was used.

The research population for the purpose of this study consists of all data analytics models for cybersecurity that have been proposed and developed in literature, journals, conference proceedings and working papers. The researcher identified two data analytics models or frameworks from a review of literature and the sample size of 8. Eight participants in total were interviewed. However, while this may be limited data, it will be sufficient for the present needs of this study. The researcher used secondary data in order to investigate the application of data analytics models in cybersecurity. In analyzing the different data analytics models for cybersecurity the researcher makes reference to the characteristics of an ideal data analytics model for cybersecurity. The basic framework for big data analytics model for cybersecurity consists of three major components which are big data, analytics, and insights (Hashem, I.A.T., et al, 2015). This is depicted in Figure 10 below. The first component in the bigdata analytics framework for cybersecurity is the availability of big data about cybersecurity. Traditional sources of big data are systems logs and vulnerability scans (Hashem, I.A.T., et al, 2015). However, sources of big data about cybersecurity have extended to include computer-based data, mobile-based data, physical data of users, human resources data, credentials, one-time passwords, digital certificates, biometrics, and social media data (Truong, T.C., 2020). Basic sources of big data identified for cybersecurity work include business mail, access control systems, CRM system and human resources system, a number of pullers in linked data networks, intranet/ internet and industrial internet of things (IIoT) /IoT, collectors and aggregators in social media networks and external news tapes (Stallings, W., 2015). To address the concerns of big data about cybersecurity, more robust big data analytics models for cybersecurity have been developed in data mining techniques and machine learning (Hashem, I.A.T., et al, 2015). In cybersecurity, big data analytics employs data mining reactors and algorithms, intrusion and malware detection techniques, and support vector machine learning techniques (Hashem, I.A.T., et al, 2015). However, the greatest challenges faced in intrusion detection systems include data nonstationarity, unbounded patterns, individuality, uneven time lags, high false alarm rates, and collusion attacks (Menzes, F.S.D., et al, 2016). This necessitates a multi-layered and multidimensional approach to big data analytics for cybersecurity. In other words an effective big data analytics model for cybersecurity must be able to detect intrusions and malware at every layer in the cybersecurity framework.

The research population for the purpose of this study consists of all data analytics models for cybersecurity that have been proposed and developed in literature, journals, conference proceedings and working papers. The researcher identified two data analytics models or frameworks from a review of literature and the sample size of 8. Eight participants in total were interviewed. However, while this may be limited data, it will be sufficient for the present needs of this study. The researcher used

secondary data in order to investigate the application of data analytics models in cybersecurity. In analyzing the different data analytics models for cybersecurity the researcher makes reference to the characteristics of an ideal data analytics model for cybersecurity. The basic framework for big data analytics model for cybersecurity consists of three major components which are big data, analytics, and insights (Hashem, I.A.T., et al, 2015). This is depicted in Figure 10 below.

The first component in the big data analytics framework for cybersecurity is the availability of big data about cybersecurity. Traditional sources of big data are systems logs and vulnerability scans (Hashem, I.A.T., et al, 2015). However, sources of big data about cybersecurity have extended to include computer-based data, mobile-based data, physical data of users, human resources data, credentials, one-time passwords, digital certificates, biometrics, and social media data (Truong, T.C., 2020). Sources of big data about cybersecurity that have been identified by other researchers include business mail, access control systems, CRM system and human resources system, a number of pullers in linked data networks, intranet/ internet and industrial internet of things (IIoT) /IoT, collectors and aggregators in social media networks and external news tapes (Stallings, W., 2015).

To address the concerns of big data about cybersecurity, more robust big data analytics models for cybersecurity have been developed in data mining techniques and machine learning (Hashem, I.A.T., et al, 2015). Big data analytics employ intrusion and malware detection techniques, data mining reactors and algorithms, and vector machine learning techniques for cybersecurity (Hashem, I.A.T., et al, 2015). However, intrusion detection systems face challenges such as data nonstationarity, collusion attacks, unbounded patterns, uneven time lags, individuality, and high false alarm rates (Menzes, F.S.D., et al, 2016). This necessitates a multi-layered and multi-dimensional approach to big data analytics for cybersecurity. In other words an effective big data analytics model for cybersecurity must be able to detect intrusions and malware at every layer in the cybersecurity framework.

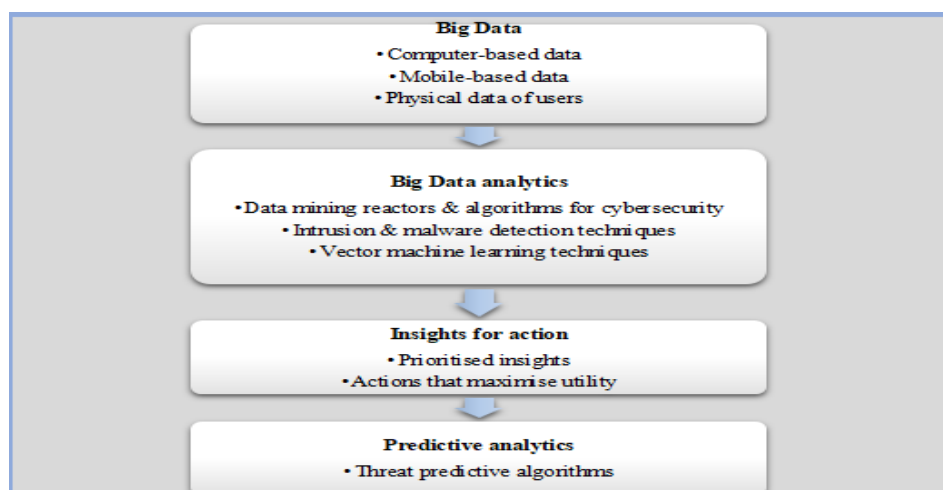


Figure 10: Big Data Analytics Model for Cybersecurity

4. RESULTS AND ANALYSIS

4.1 Knowledge Generation for STI Investment Projects

Research creates knowledge and technology, and the process of innovation goes further to include putting that knowledge into practical use. Knowledge co-creation is a synergetic process of combining selected value-adding content and process from disciplinary traditions to synthesize new ways of knowing. The innovative support instruments may include applied research, idea/proof of concept up to commercialisation stage, venture capital, and foreign direct investment (FDI). The possible contributing factors to why African scientific contribution globally is only 2% may include:

- ❖ Weakness or nonexistence of an environment advantageous for research;
- ❖ Deficient budget dedicated to research;
- ❖ Not rewarding status of the researchers;
- ❖ Rough evaluation of the impact of research on development.

The strategies for knowledge generation through the Rapid Results Initiative (RRI) are:

1. Advocacy of the National Science, Technology and Innovation System
2. Knowledge co-creation and collaboration
3. Knowledge acquisition
4. Knowledge prioritization
5. Knowledge fusion and adaption
6. Knowledge enterprenuerising and commercialization

Following a review of the national research priorities, SDGs and STISA 2024. The following key projects were identified as key projects to be pursued in the next 100 days:

1. Post harvest technologies
2. Small scale mining/mineral value addition/bio mining
3. Clean water alternatives
4. Tiles technologies from mining waste
5. ICT innovations for applications /Cyber security systems
6. Defence technologies (double use technologies, drones, puma vehicle, land mine detectors, etc.)

The critical success factors for the successful knowledge generation, exploitation and commercialisation of priority scientific projects were identified as:

- ❖ High level sponsorship
- ❖ Resources
- ❖ Skills
- ❖ Project management
- ❖ Teamwork
- ❖ Excellence
- ❖ Convergence of mindset

A visit was made to the higher education institutions in Zimbabwe to ascertain the availability and possible exploitation of the new innovations in line with the key priority projects. The best

technologies identified for practical and urgent exploitation at least cost are shown on the schema below on Table 7 below:

Table 7: Key priority technological innovations in Zimbabwe

Priority Technology	Schema
1. Kwekwe Polytechnic Brick Machine	
2. Chinhoyi University of Technology Animal drawn Lime Spreader	
3. Harare Polytechnic Floor tile manufactured	

The growth of the Zimbabwean economy, anchored by the National Development Strategy, is purposed to achieve the following elements:

- i. Improved access to credit and liquidity by key sectors of the economy such as agriculture;
- ii. Establishment of a Sovereign Wealth Fund;
- iii. Improvement of revenue collection from key sectors of the economy such as mining;
- iv. Increased investment in infrastructure such as energy and power development, rail, roads, telecommunication, ICTs, aviation, water and sanitation, through acceleration in the

- implementation of Public Private Partnerships (PPPs) and other private sector driven initiatives;
- v. Increased Foreign Direct Investment (FDI);
- vi. Establishment of Special Economic Zones;
- vii. Continued use of the multi-currency system;
- viii. Implementation of effective Value Addition policies and strategies; and
- ix. Improved supply of electricity and water.

The National Development Strategy must therefore be premised on the scientific technological solutions and economic framework inferred by the national science, technology and innovation system of Zimbabwe. Special attention should be given to the following sectors of the economy:

1. Agriculture
2. Mining
3. ICT
4. Research and Development
5. Manufacturing
6. Construction
7. Human capital development
8. Health and social services
9. Defence and security

The financing model for the National Development Strategy has the following elements:

- 1) National resource mobilisation programmes from domestic resources driven by the Government of Zimbabwe
- 2) The establishment of the National Wealth and Innovation Fund
- 3) Collaboration with all the development partners in STEM-related projects and programmes
- 4) Public Private Partnerships (PPPs) in the proposed Special Economic Zones.

However, it is envisaged that a number of fiscal reform measures shall be undertaken in order to improve fiscal policy management and financial sector stability. Progress shall be monitored against registered re-engagement process with the International Financial Institutions (IFIs) and creditors through various strategies.

4.2 Performance of Machine Learning Algorithms

The gross inadequacies of classical security measures have been overtly exposed. Therefore, effective solutions for a dynamic and adaptive network defence mechanism should be determined. Intrusion attack classification requires optimization and enhancement of the efficiency of data mining techniques. Table 8 shows a comparison of the data mining techniques that can be used in intrusion detection.

TABLE 8: Advantages and disadvantages of data mining techniques (Source: Berman, D.S., et al, 2019)

Technique	Advantages	Disadvantages
Genetic Algorithm	❖ Finding a solution for any	❖ Complexity to propose a

	optimization problem ❖ Handling multiple solution search spaces	problem space ❖ Complexity to select the optimal parameters ❖ The need to have local searching technique for effective functioning
Artificial Neural Network	❖ Adapts its structure during training without the need to program it	❖ Not accurate results with test data as with training data
Naive Bayes Classifier	❖ Very simple structure ❖ Easy to update	❖ Not effective when there are high dependency between features
Decision Tree	❖ Easy to understand ❖ Easy to implement	❖ Works effectively only with attributes having discrete values
K Mean	❖ Very easy to understand ❖ Very simple to implement in solving clustering problems	❖ Number of clusters is not automatically calculated ❖ High dependency on initial centroids.

An intrusion detection system determines if an intrusion has occurred, and so monitors computer systems and networks, and the IDS raises an alert when necessary (Bloice, M., and Holzinger, A., 2018). However, Bloice, M., and Holzinger, A. (2018) addressed the problems of Anomaly Based Signature (ABS) which reduces false positives by allowing a user to interact with the detection engine and raising classified alerts. The advantages and disadvantages of ABSs and SBSs are summarised on table, Table 3, below.

TABLE 3: Advantages and disadvantages of ABSs and SBSs models (Source: Bloice, M., and Holzinger, A., 2018).

Detection model	Advantages	Disadvantages
Signature-based	❖ Low false positive rate ❖ Does not require training ❖ Classified alerts	❖ Cannot detect new attacks ❖ Requires continuous updates ❖ Training could be a thorny task
Anomaly-based	❖ Can detect new attacks ❖ Self-learning	❖ Prone to raise false positives ❖ Black-box approach ❖ Unclassified alerts ❖ Require initial training

The performance of each of the classical Machine Learning algorithms is presented below from Figure 11.

4.2.1. Classification and Regression Trees (CART)

Table 4 below shows the performance results of our CART algorithm in predicting bank failure on the training set. The algorithm's level of accuracy on the training dataset was 82.8%. The best tune or complexity parameter of our optimal model was 0.068. On the test dataset, the algorithm achieved an accuracy level of 92.5% and a kappa of 88.72%. The algorithm only misclassified 2 instance as moderate and 1 as satisfactory.

Complexity Parameter	Accuracy	Kappa	AccuracySD	KappaSD
0.06849315	0.8275092	0.7519499	0.04976459	0.07072572
0.15753425	0.7783150	0.6683229	0.07720896	0.14039942
0.42465753	0.5222344	0.1148591	0.08183351	0.18732422

TABLE 4: CART model performance.

The accuracy of the CART model based on the complexity parameters of different test runs is shown on Figure 11 below. The complexity parameter or the best tune parameter of 0.068 optimized the model performance.

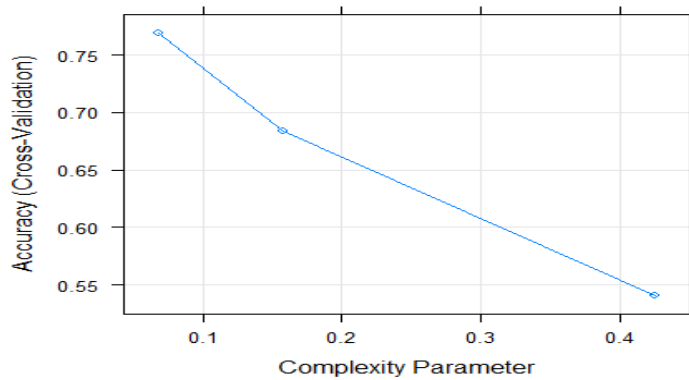


FIGURE 11: CART accuracy curve.

4.2.2. Support Vector Machine

The accuracy level of the SVM model on the training dataset was 79.1% in predicting bank solvency as shown in table 5. The best tune sigma and cost values of our highly performing model where 0.05 and 1 as shown on Figure 12 below. The Kappa statistic and the Kappa SD where 67.9% and 0.13 respectively. On the test dataset, the algorithm achieved an accuracy level of 92.5% and a kappa of 88.54%. The algorithm only misclassified 3 instance as moderate in comparison to the CART algorithm.

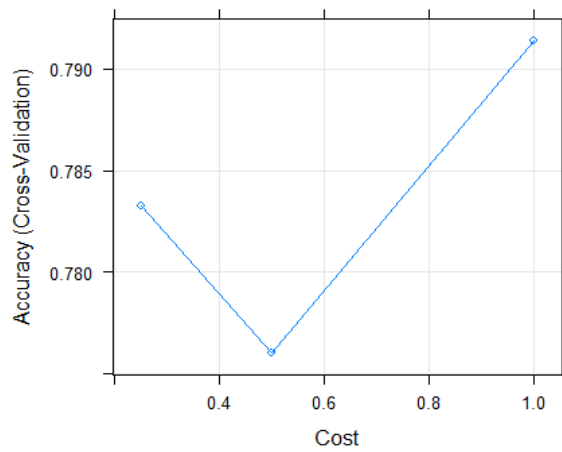


FIGURE 12: SVM accuracy curve

sigma	c	Accuracy	Kappa	AccuracySD	KappaSD
0.050398	0.25	0.783223	0.678536	0.095598	0.140312
0.050398	0.50	0.776007	0.661354	0.087866	0.132552
0.050398	1.00	0.791391	0.678694	0.080339	0.126466

TABLE 5: Support Vector Machine performance

4.2.3. Linear Discriminant Algorithm

Accuracy	Kappa	AccuracySD	KappaSD
0.8042399	0.7038131	0.1016816	0.159307

TABLE 6: Linear Discriminant algorithm performance

On the training dataset, the LDA achieved an accuracy level of 80% as in table 6. The Kappa statistic and the Kappa SD were 70% and 0.16 respectively. On the test dataset, the algorithm achieved an accuracy level of 90% and a kappa of 84.64%. The algorithm only misclassified 4 instance as moderate whose performance is poor in comparison to the CART algorithm.

4.2.4. K-Nearest Neighbor

Table 7 shows the K-NN algorithm performance and confusion accuracy on Figure 10.

K	Accuracy	Kappa	AccuracySD	KappaSD
5	0.5988645	0.3698931	0.1280376	0.2158109
7	0.6268864	0.4072928	0.1564920	0.2703504

9	0.6621978	0.4715556	0.1747903	0.2881390

TABLE 7: K-NN algorithm performance

The level of accuracy on the training dataset was 66.2%. The best tune parameter for our model was k=9 or 9 neighbors as shown on the accuracy curve in Figure 13 below. The Kappa statistic and the Kappa SD where 47.2% and 0.17 respectively. On the test dataset, the algorithm achieved an accuracy level of 67.5% and a kappa of 49%. The algorithm was not highly effective in classifying bank performance in comparison to other algorithms.

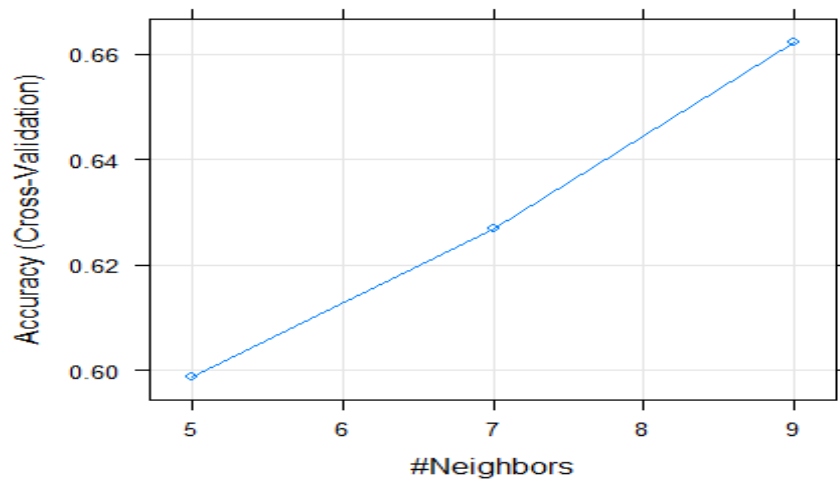


FIGURE 13: K-NN confusion accuracy graph

4.2.5. Random Forest

TABLE

8:

mtry	Accuracy	Kappa	AccuracySD	KappaSD
2	0.8272527	0.7421420	0.10396454	0.15420079
14	0.8554212	0.7829891	0.06069716	0.09303130
16	0.8482784	0.7718935	0.06455248	0.09881991

Random Forest performance

On the training set, the accuracy of our random forest was 85.5% as designated in table 8. The best tune parameter for our model was the mtry of 14 which is the number of randomly selected predictors

in constructing trees as shown on Figure 14. The Kappa statistic and the Kappa SD where 78.3% and 0.09 respectively. On the test dataset, the algorithm achieved an accuracy level of 96% and a kappa of 96%. The algorithm was highly effective in classifying bank performance in comparison to all algorithms.

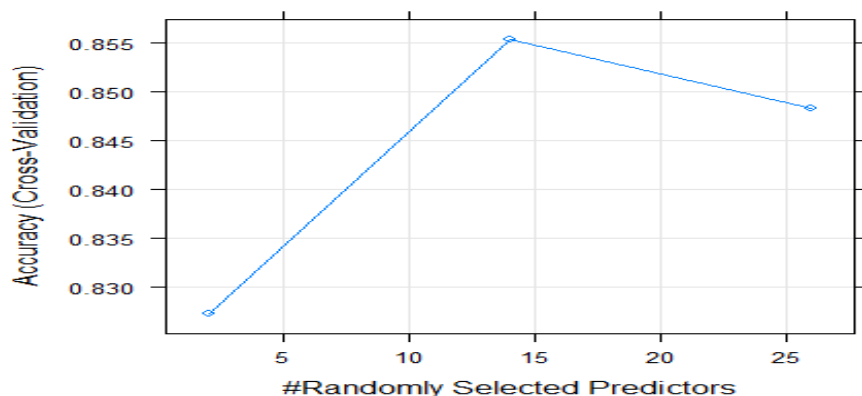


FIGURE 14: Random forest accuracy graph

4.2.6. Challenges and Future Direction

As the number of banking activities increase, also it implies that the data submission to the Reserve Bank of Zimbabwe continues to grow exponentially. This challenging situation in combination with advances in machine learning (ML) and artificial intelligence (AI) presents unlimited opportunities to apply neural network-based deep learning (DL) approaches to predict Zimbabwean Bank's solvency. Future work will focus on identifying more features that could possibly lead to poor bank performance and incorporate these in our models to develop a robust early warning supervisory tool based on big data analytics, machine learning and artificial intelligence.

The researcher analyses the two models that have been proposed in literature with reference to an ideal data analytics model for cybersecurity presented in Section 3.

4.2.7. Model 1: Experimental/ Prototype Model

In the first case the researcher makes reference to the model presented in Stallings, W. (2015) which although developed in the context of the public sector can be applied to the private sector organizations. Table 9 below summarizes the main characteristics of the experimental model. [The reader is referred to the prototype model also demonstrated in Stallings, W. (2015).

TABLE 9: EXPERIMENTAL BIG DATA ANALYTICS MODEL FOR CYBERSECURITY

MODEL ATTRIBUTES	DESCRIPTION
HBase working on HDFS (Hadoop Distributed File System)	<ul style="list-style-type: none"> HBase, a non-relational database, facilitates analytical and predictive operations Enables users to assess cyber-threats and the dependability of critical infrastructure

Analytical data processing module	<ul style="list-style-type: none"> Processes large amounts of data, interacts with standard configurations servers and is implemented at C language Special interactive tools (based on JavaScript/ CSS/ DHTML) and libraries (for example jQuery) developed to work with content of the proper provision of cybersecurity
Special interactive tools and libraries	<ul style="list-style-type: none"> Interactive tools based on JavaScript/ CSS/ DHTML Libraries for example jQuery developed to work with content for Designed to ensure the proper provision of cybersecurity
Data store for example (MySQL)	<ul style="list-style-type: none"> Percona Server with the ExtraDB engine DB servers are integrated into a multi-master cluster using the Galera Cluster.
Task queues and data caching	<ul style="list-style-type: none"> Redis
Database servers balancer	<ul style="list-style-type: none"> Haproxy
Web server	<ul style="list-style-type: none"> nginx , involved PHP-FPM with APC enabled
HTTP requests balancer	<ul style="list-style-type: none"> DNS (Multiple A-records)
Development of special client applications running Apple iOS	<ul style="list-style-type: none"> Programming languages are used: Objective C, C++, Apple iOS SDK based on Cocoa Touch, CoreData, and UIKit.
Development of applications running Android OS	<ul style="list-style-type: none"> Google SDK
Software development for the web platform	<ul style="list-style-type: none"> PHP and JavaScript.
Speed of the service and protection from DoS attacks	<ul style="list-style-type: none"> CloudFare (through the use of CDN)

(Source: Stallings, W. , 2015).

4.2.8 Model 2: Cloud computing/Outsourcing

The second model involves an organization outsourcing its data to a cloud computing service provider. Cloud computing service providers usually have advanced big data analytics models, with advanced detection and prediction algorithms and better state of the art cybersecurity technologies and better protocols because they specialize in data and networks. However, it is to be noted that cloud computing service providers are neither exempt nor immune from cyber-threats and attacks.

4.3. Development of the Bayesian Network Model

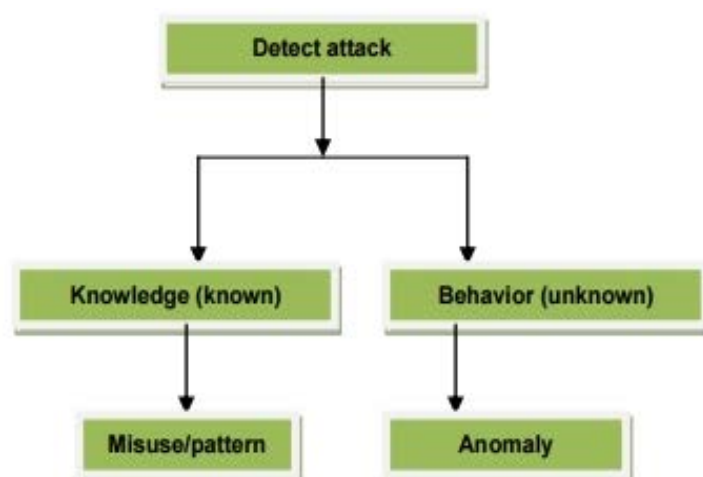
Bayesian networks allow for prediction, generalization, and planning. It must be noted that network traffic behaviour as well as payload protocol lexical and syntactical patterns may differ substantially depending on the sort of service provided from each specific equipment, i.e. from each different IP address and from each specific TCP destination port. Bringas, P.B., and Santos, I., (2010) proposed the use of a multi-instance schema, with several Dynamic Bayesian Networks, one for each combination of TCP destination address and port. Bringas, P.B., and Santos, I., (2010) argue that it must be able to simultaneously offer efficient response against both well-known and zero-day attacks. Bayesian networks require many computational resources. Hence, several of the tasks to be performed must be designed in a parallel way to accelerate it (Bringas, P.B., and Santos, I., 2010, p.240).

Adjusting the whole behaviour of the Network Intrusion Detection System to special needs or configurations has a high degree of complexity in Bayesian structures and conditional probability

parameters. The dynamic regulation of knowledge representation model can be accomplished by using the sensibility analysis so as to avoid denial of service attacks, automatically enabling or disabling expert modules by means of one combined heuristic measure which considers specific throughputs and representative features (Bringas, P.B., and Santos, I., 2010, p.242). Furthermore, it is also possible to perform model optimization, to obtain the minimal set of representative parameters, and also the minimal set of edges among them, with the subsequent increase of the general performance. In order to improve inference and adaptation time of response, approximate evidence propagation methods can also be applied.

Intrusion Detection System (IDS) operates differently from a firewall and antivirus. Firewall and antivirus software can be bypassed, and does not stop internal intrusion and as well as external attacks efficiently. Firewall generally works on static rules via which it filters traffic but never has ability to detect intrusion. IDS detects intrusion after its first occurrence in order to prevent such future attacks (Murugan, S., and Rajan, M.S., 2014, p.1). The simple rules for the analysis of attack are shown on Figure 15 below. Any kind of unusual behaviour on the network triggers an alarm on the IDS for the anomaly-based intrusion detection method.

Figure 15: Analysis of Attack [\(Source: Murugan, S., and Rajan, M.S., 2014, p.2\)](#)



Defence security agencies and other militarily related organizations are highly concerned about the confidentiality and access control of the stored data. Therefore, it is really important to investigate on Intrusion Detection System (IDS) to detect and prevent cybercrimes to protect these systems (Alocious, C., *et al*, 2014, p.1). Distributed Denial of Service (DDoS) attacks are carried out to make system inaccessible by flooding the server's network and end user systems with fake generated traffic. In this way, legitimate users would be prevented from accessing the system resources. Signature based detection also called as rule based detection determines the user behavior with a comparison of some rules defined related to legitimize the user's behavior. Signature based IDS consists of a database of known signatures of known attacks, these attacks are predefined based on the attack analysis. There is a phenomenal growth in modern cyber-attacks and their formations are changing regularly. A genetic algorithm is a computational model, where the basic concepts behind genetic algorithm is an

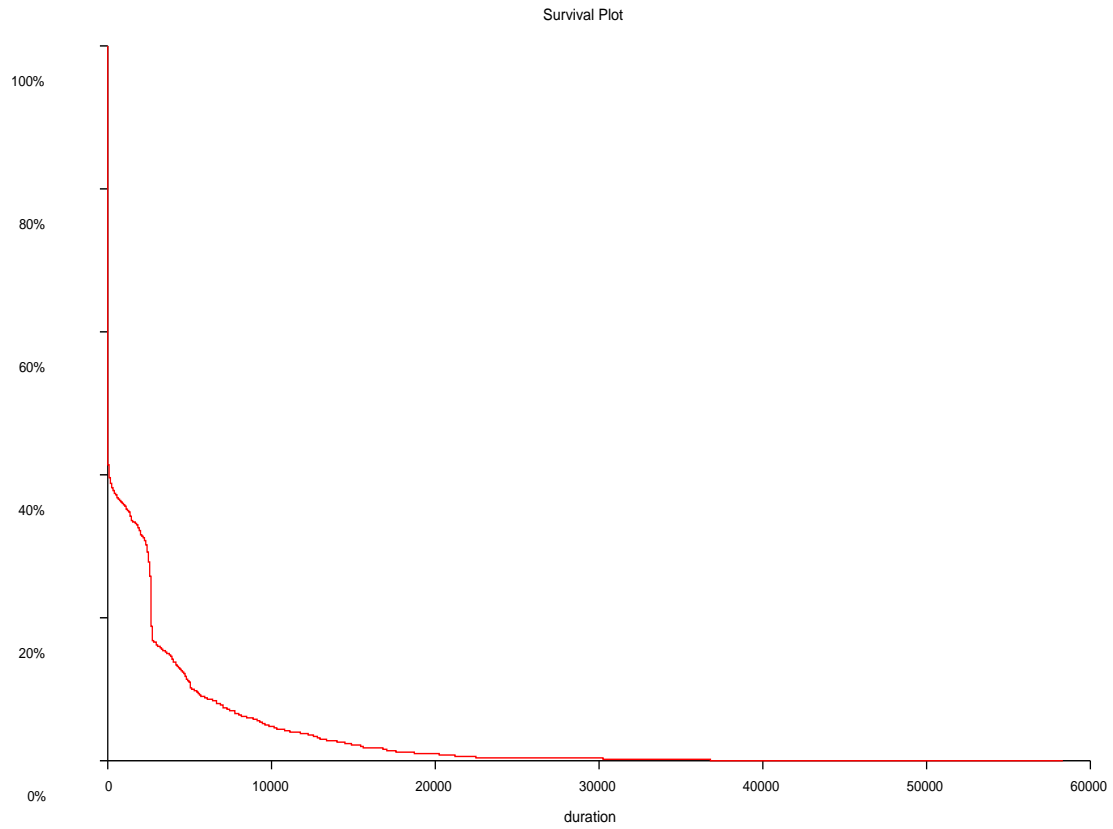
evaluation and natural selection. This means only the fittest will be survived in the process of natural selection. Genetic algorithms are used by creating a set of rules for network data.

The analysis of the quantitative data was done using the SNORT open source software and other Bayesian Network supportive platforms such as NCSS 2019, Pass 2019, GeNIe 2.3, WinBUGS14, BayES and Analytica 5.1. In sniffer mode, the program will read network packets and display them on the console. In packet logger mode, the program will log packets to the disk. The SNORT IDS mode was used to illustrate the results of the research. SNORT was chosen due to the following reasons:

- ❖ Support multiple packet processing threads
- ❖ Shared configuration and attribute table
- ❖ Use a simple, scriptable configuration
- ❖ Make key components pluggable
- ❖ Autodetect services for portless configuration
- ❖ Support sticky buffers in rules
- ❖ Autogenerate reference documentation
- ❖ Provide better cross platform support
- ❖ Development for the project will be fast paced

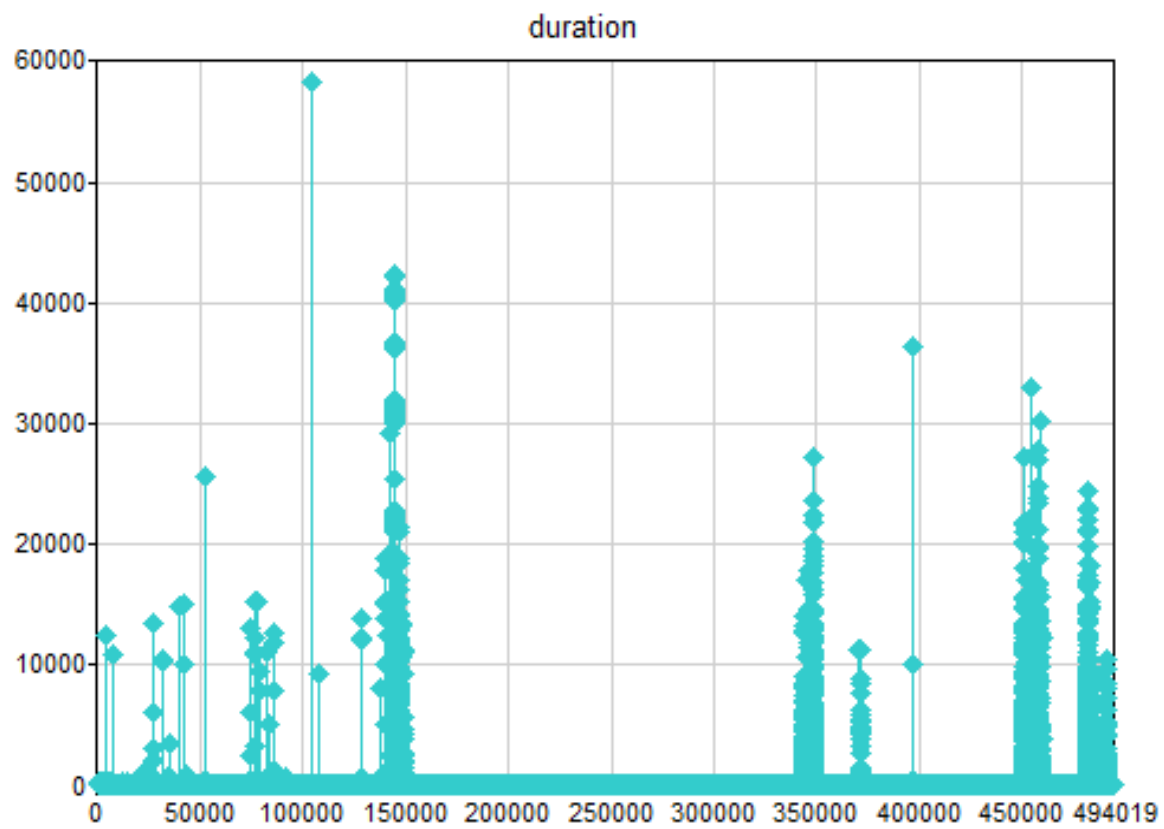
A non-parametric Survival Analysis was conducted with respect to one variable, *duration*, and the result is shown as Kaplan-Meier Survival Curves on Figure 16 below.

Figure 16: Kaplan-Meier Survival Curve(s)



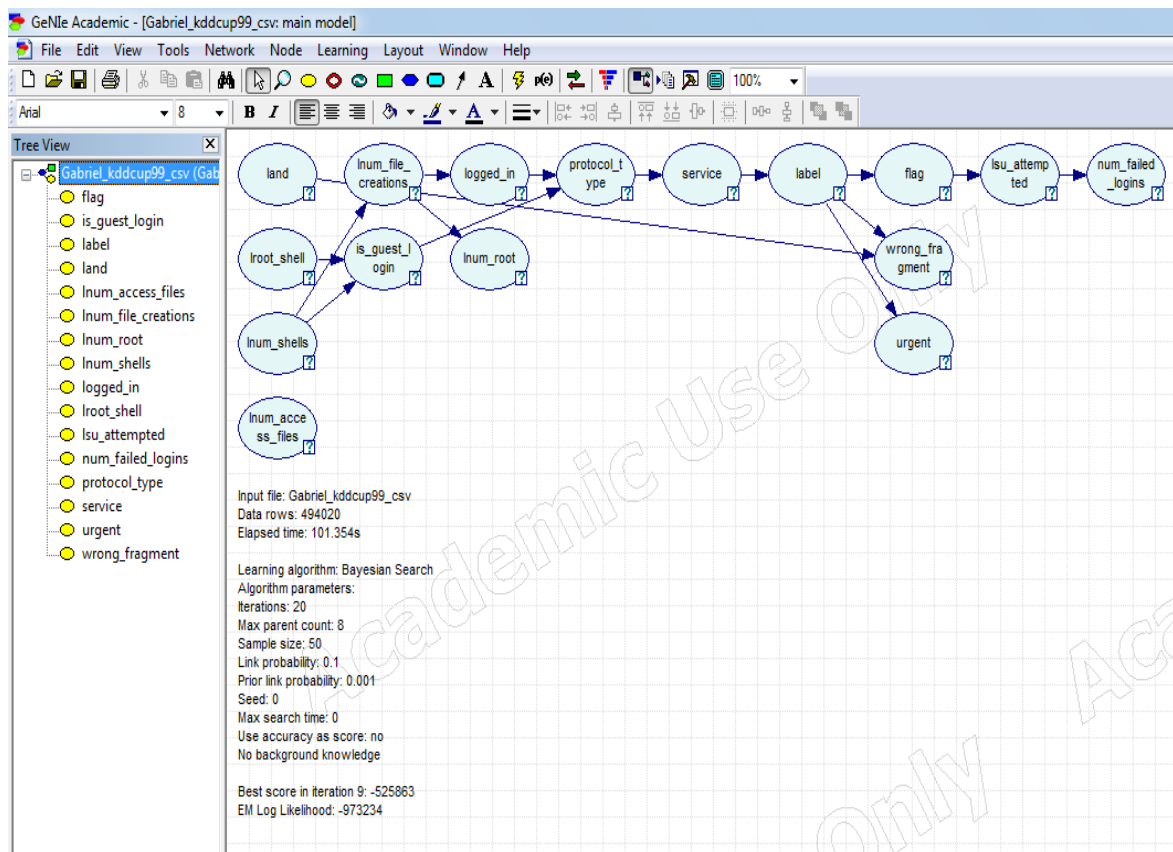
Time Series analysis of the variable *Duration* is shown on Figure 17 below.

Figure 17: Time Series of the variable Duration



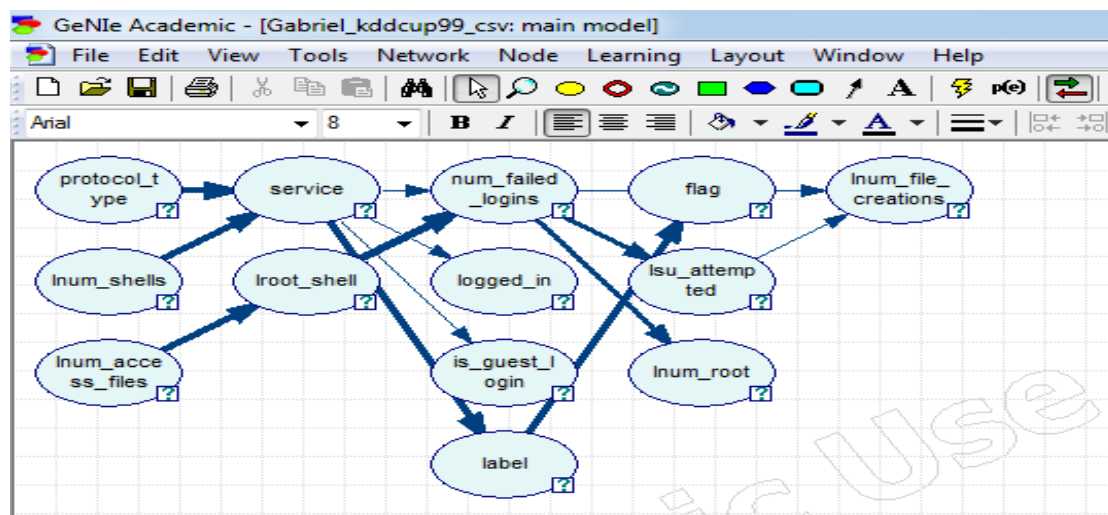
A new Bayesian Network model was created from the dataset and is shown on Figure 18 below.

Figure 18: Bayesian Network Structure



The Strength of Influence of the Bayesian Network is shown on Figure 19 below.

Figure 19: Strength of Influence



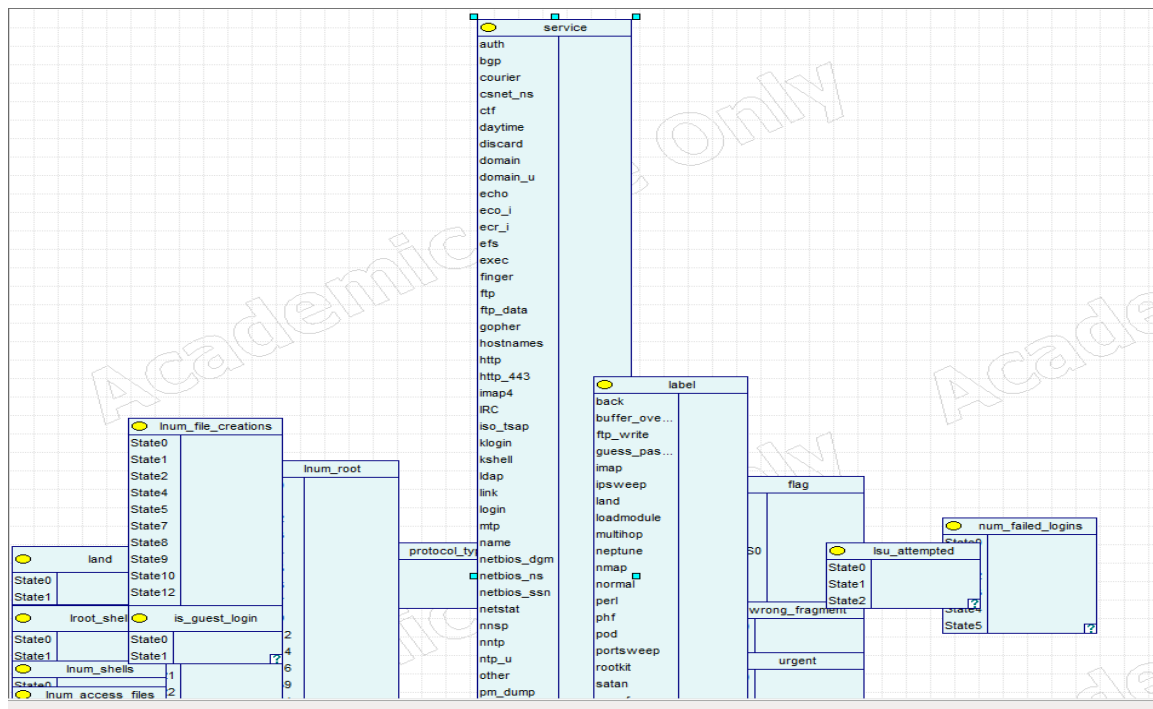
The Adjacency Matrix was computed for the network and is shown on Figure 20 below.

Figure 20: Adjacency Matrix

Adjacency Matrix												
	is_guest_login	lnum_file_creations	lnum_root	lsu_attempted	logged_in	num_failed_logins	lrout_shell	lnum_access_files	flag	label	service	lnum_shells
is_guest_login												
lnum_file_creations												
lnum_root												
lsu_attempted		X										
logged_in												
num_failed_logins			X	X								
lrout_shell						X						
lnum_access_files							X					
flag		X										
label								X				
service	X	X			X	X			X	X		
lnum_shells											X	
protocol_type											X	

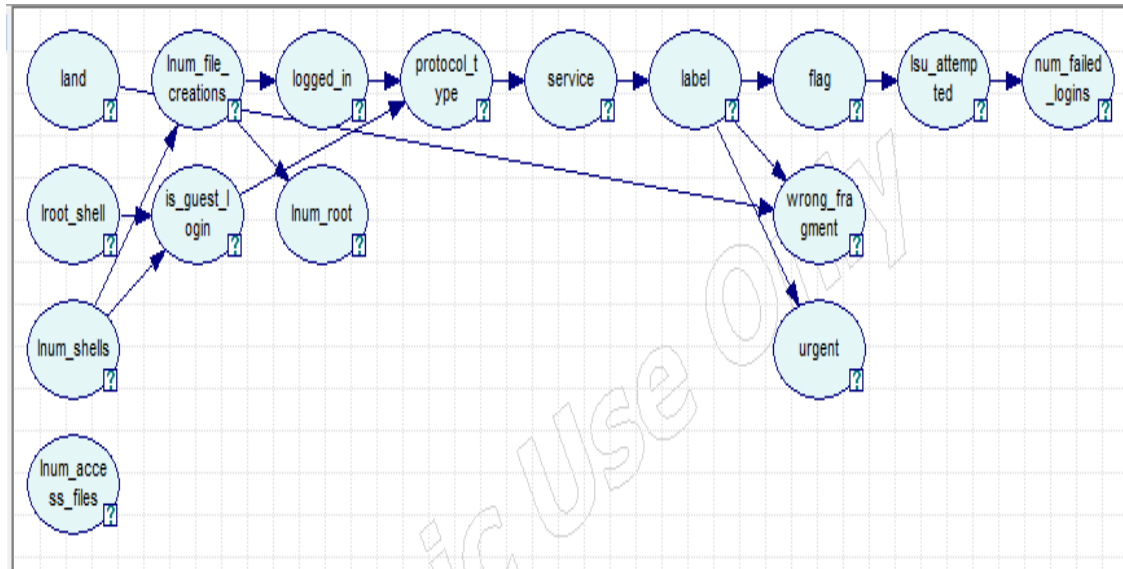
The bar chart of the Node Properties are shown on Figure 21 below.

Figure 21: Bar Chart of Node Properties



The Bayesian Network structure derived from the dataset is shown below on Figure 22.

Figure 22: Bayesian Network structure



There are several problems in the use of Bayesian Networks, one of which is about the correspondence between the graphical structure and associated probabilistic structure which allows us to reduce all the problems of inference problems in graph theory. The other problem is in the operation for transposition of the causal graph to a probabilistic representation. However, BNs have been applied in anomaly detection in different ways, one of which is the Naive Bayes, which is a two-layer Bayesian network that assumes complete independency between the nodes.

From the sample dataset of 494,020 instances with 42 variables analysed, the mean values of the key variables are shown on the table below, Table 10:

Table 10: Mean values of the selected key variables

Variable	Mean
duration	47.9794
protocol_type	tcp
service	http
flag	SF

src_bytes	3025.62
dst_bytes	868.531
land	4.45E-05
wrong_fragment	0.00643294
urgent	1.42E-05
hot	0.0345188
num_failed_log+	0.000151816
logged_in	0.148245
Inum_compromis+	0.0102121
lroot_shell	0.000111332
lsu_attempted	3.64E-05
Inum_root	0.0113518
Inum_file_crea+	0.00108295
Inum_shells	0.000109307
Inum_access_fi+	0.00100806
is_guest_login	0.00138658
count	332.286
srv_count	292.907
serror_rate	0.176687
srv_serror_rate	0.176609
rerror_rate	0.0574335
srv_rerror_rate	0.0577191
same_srv_rate	0.791547
diff_srv_rate	0.0209824
srv_diff_host_+	0.0289962
dst_host_count	232.471
dst_host_srv_c+	188.666
dst_host_same_+	0.753781
dst_host_diff_+	0.0309058

dst_host_same_+	0.601936
dst_host_srv_d+	0.00668351
dst_host_serro+	0.176754
dst_host_srv_s+	0.176443
dst_host_rerro+	0.0581177
dst_host_srv_r+	0.0574118
label	normal

It's possible that abnormal behavior can happen as also be a result of factors such as policy changes or the offering of new services by a site. The solution to these two problems is the introduction of a hybrid detection which takes advantage of misuse detection to have a high detection rate on known attacks and capacity to detect unknown attacks. The most common type of hybrid system is to combine a misuse detection and an anomaly detection together. Arguably, a hybrid IDS can be used by combining both misuse detection and anomaly detection components, in which a random forest algorithm was applied firstly in the misuse detection module to detect known intrusions. Evaluations with a part of the KDDCUP'99 data set used in this research showed that the misuse detection module generated a high detection rate with a low false positive rate, and at the same time the anomaly detection component had the potential to find novel intrusions.

Zekrifa, D.M.S. (2014, p.17) proposed a two-stage hybrid intrusion detection and visualization system that leverages the advantages of signature-based and anomaly detection methods, which potentially could identify both known and unknown attacks on system calls. A suggested improved IDS would be a novel hybrid IDS system consisting of an anomaly detection module, a misuse detection module, and a decision support system. The decision support system would be used to combine the results of the two previous detection modules. Instead of combining signature detection techniques and anomaly detection techniques, some other hybrid systems fuse multiple anomaly detection systems according to some specific criteria considering that the detection capability for each anomaly detection technique is different. The hybrid system is purposed to keep an acceptable detection rate and reduce the large number of false alerts generated by current anomaly detection approaches (Zekrifa, D.M.S., 2014, p.19).

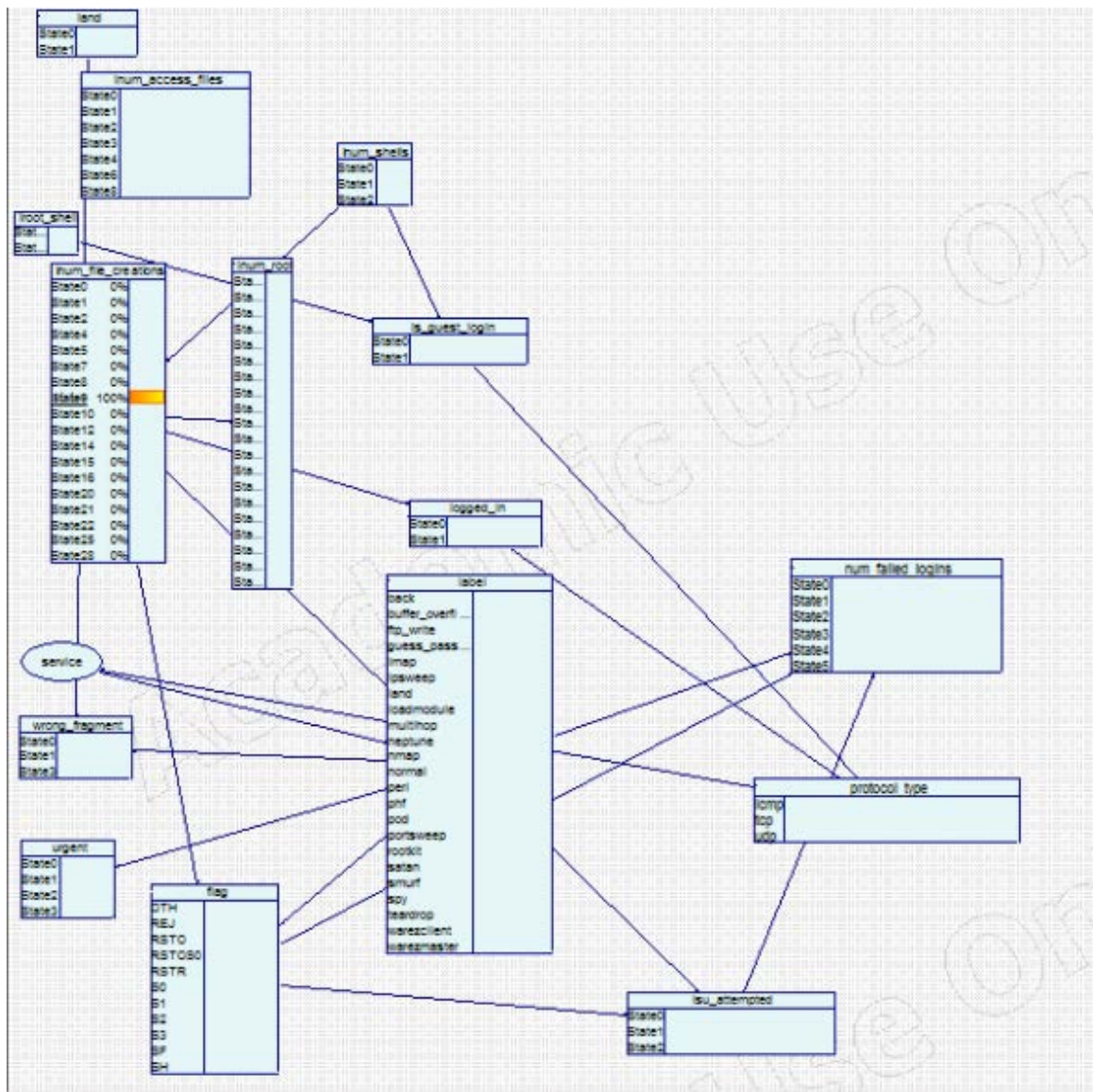
The cybersecurity challenges that are being faced in developing countries, like Zimbabwe, include the following:

- Infrastructure
- Legal frameworks
- 3. Harmonization of legislation
- 4. Balancing harmonization and country specific needs
- 5. Systems
- 6. Education and awareness
- 7. Cybersecurity knowledge
- 8. Affordability and funding
- 9. Perceived low susceptibility to attacks
- 10. Lack of adequate frameworks that speak to their cybersecurity needs
- 11. Reporting cybercrime

12. Data sharing

The Bayesian Network Model developed is shown on Figure 23 below.

Figure 23: The Bayesian Network Model developed



A majority of the currently available network security techniques cannot cope with the dynamic and increasingly complex nature of the attacks on distributed computer systems. Hence, it becomes necessary to construct an automated and adaptive defensive tool for computer networks. Existing

techniques for preventing intrusions start with encryption and firewalls, then followed by Intrusion Detection System (IDS) technology which is able to detect unauthorized access and abuse of computer systems from both internal users and external offenders (Tran, T.P., 2009, p.iv). Artificial Intelligence (AI) technologies such as Artificial Neural Networks (ANN) have been adopted to improve detection performance. However, ANN is computationally expensive.

Updating the probabilities in the network structure requires learning the structure of the Bayesian network and use of prior knowledge and data (Soberanis, I.V.D., 2010, p.66). In sequential update of Bayesian Networks the learning procedure receives the data as stream of observations and there is an output model from the learning procedure, based on the data observed thus far. There are various Sequential Update approaches: naive approach, maximum a-posteriori probability (MAP), and the incremental approaches (Soberanis, I.V.D., 2010, p.67). However, the huge amount of data requires a lot of memory. In order to deal with the large data set issue, the MAP approach stores all the previous data by summarizing the data used in the model so far assuming that the data being summarized has a probability distribution based on the current model. Bayesian updating can be recursively and incrementally updated. The wonderful thing about recursive bayesian updating is that it is simple and has a wide variety of applications. The methodical and efficient method of clustering is provided by the junction tree algorithm. This method involves performing bayesian propagation on an updated graph called a junction tree. The Junction tree approach eliminates cycles in a network by clustering them into single nodes (Soberanis, I.V.D., 2010, p.70). Reasoning with Bayesian network is done by updating the probabilities, which involves using new information or evidence to compute the posterior probability distributions. Bayesian updating for any probabilistic inference is the computation of the posterior probability distribution for a set of query nodes, given values for some evidence nodes.

Learning can be assisted by the use of existing knowledge, which we can refer to as the training data. In fact, prior knowledge can be enormously useful in learning. The knowledge that we compile or is given can greatly aid in the speeding up the decision making process. There are a variety of learning techniques that can be utilized based on the data. The learning method can be supervised, unsupervised or reinforced. Supervised learning is the adjustment of the state of the network in response to the data generated in the environment (Soberanis, I.V.D., 2010, p.74). In unsupervised training, the network is provided with inputs but not with desired outputs, that is the training data is provided and the likely or unlikely data is derived. The system itself must then decide what features it will use to group the input data or the network has to make sense of the inputs without outside help. Soberanis, I.V.D. (2010, p.126) proposed an online traffic classification method, in which the unigram payload distribution model is applied to extract the required set of features. Thereafter the J48 decision tree is employed to classify the network applications based on the unigram features, and observed that the signatures are present in some designated positions in the payload. It is important to place more weight on the features that appear in these more important positions through a weighted scheme over the features using a genetic algorithm.

Almutairi, A., (2016) identified two main challenges; the first one is that signature-based intrusion detection systems such as SNORT lack the capability of detecting attacks with new signatures without human intervention. The other challenge is related to multi-stage attack detection, it has been found that signature-based is not efficient in this area. Almutairi, A. (2016) handled the first challenge by developing a multi-layer classification methodology. The first layer was premised on a decision tree and the second layer was derived from a hybrid module which uses neural network and fuzzy logic as the the two data mining techniques. The second layer was purposed to detect new attacks in case the

first one fails to detect. This system detects attacks with new signatures, and then updates the SNORT signature holder automatically, without any human intervention. The obtained results showed that a high detection rate was obtained with attacks having new signatures. However, it has observed that the false positive rate needs to be lowered. The second challenge was approached by evaluating IP information using fuzzy logic. This approach looked at the identity of participants in the traffic, rather than the sequence and contents of the traffic. The results showed that this approach can help in predicting attacks at very early stages in some scenarios. Almutairi, A. (2016) conceded to the fact that combining this approach with a different approach that looks at the sequence and contents of the traffic, such as event- correlation, will achieve a better performance than each approach individually. However, building an effective solution using data mining faces some major challenges, one of which is the massive increase in the amount and complexity of data to be analysed. This makes data mining quite expensive in terms of computation, and so data mining in may consume a lot of CPU and memory resources that are expensive or not available. Hence, carrying out analysis on network traffic using a sample of the data and not all of them for the purpose of generating profiles may cause false conclusions.

5. CONCLUSION

It is envisaged that the national programme would:

- make Zimbabwe's innovation system truly international, by supporting partnerships, collaboration and foreign investment in Zimbabwean R&D;
- build a culture of innovation and new ideas by strengthening investment in creativity and knowledge generation;
- accelerate the take up of new technology, so Zimbabwean firms can access the best ideas from around Zimbabwe and the rest of the world;
- focus incentives for business R&D to promote global competitiveness, delivering the best outcomes for exports and economic growth; and
- Enable resource mobilisation for the specific national innovations and industrialisation programmes which are STEM-related.

Advocacy and publicity work was achieved through a series of communication/ visibility events; media coverage; social media profile; promotional materials; and research publications and bulletins.

Machine learning algorithms as part of Artificial Intelligence can be clustered into supervised, unsupervised, semi-supervised, and reinforcement learning algorithms. The main characteristic of ML is the automatic data analysis of large data sets and production of models for the general relationships found among data. Big data analytics is not only about the size of data but also clinches on volume, variety and velocity of data.

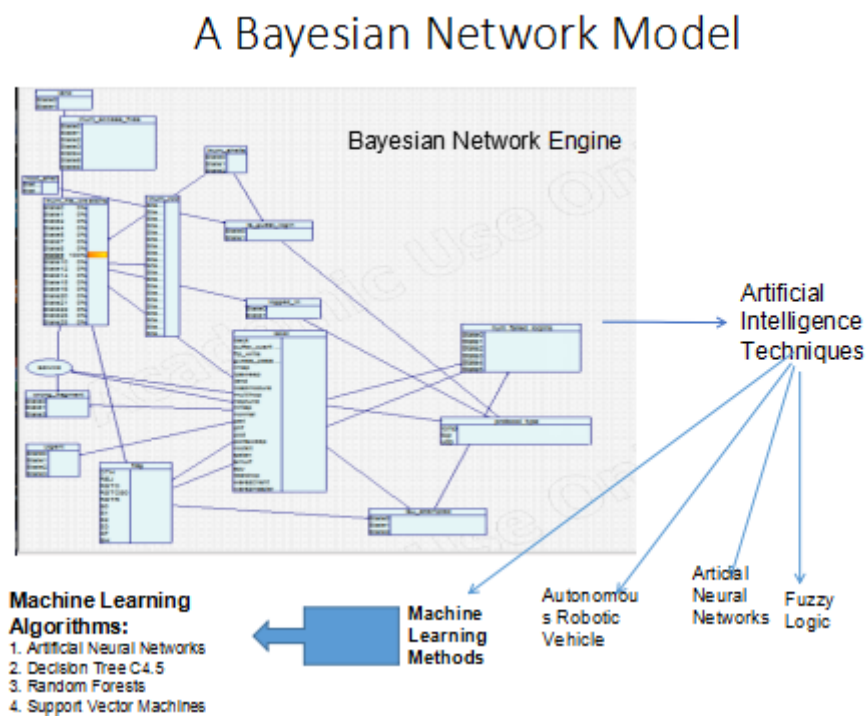
While the review of literature showed that institutions and countries adopt different big data analytics models for cybersecurity, the researcher also demonstrated that beside the unique requirements these models share major common characteristics for example reactors and detection algorithms are usually present in every model but differ in terms of complexity. Further, using the models presented in this section it is worthy of note that many small organizations will usually adopt Model 2 whereas very large organizations and sensitive public sector organizations will adopt Model 1. This may also explain why models used may differ although the framework used in designing a data analytics model for cybersecurity in a cloud computing services provider may share similar characteristics with that developed by an institution on its own.

In this section the researcher presented two models for adopting data analytics models to cybersecurity. The first experimental or prototype model involves the design, and implementation of a prototype by an institution and the second model involves the use serviced provided by cloud computing companies.

Future research work is envisaged to focus on new algorithmic performance in ML and applications in responsible AI for e-learning.

The final Bayesian Network model developed is shown on the diagram below on Figure 24.

Figure 24: The Final Bayesian Network model



However, the Bayesian Network must be supported by the Artificial Intelligence paradigms for network detection and prevention systems that include machine learning methods, autonomous robotic vehicle, artificial neural networks, and fuzzy logic. Furthermore, these algorithms ought to be used in the basic network intrusion detection and prevention system:

- ❖ Support Vector Machines,
- ❖ Artificial Neural Network,
- ❖ K-Nearest Neighbour,

- ❖ Naive-Bayes and
- ❖ Decision Tree Algorithms

Alternative improved solutions include the use of machine learning algorithms specifically Artificial Neural Networks (ANN), Decision Tree C4.5, Random Forests and Support Vector Machines (SVM). However, the use of Bayesian Networks has its own limitations which include the fact that the correspondence between the graphical structure and associated probabilistic structure will allow to reduce all the problems of inference problems in graph theory, which requires further research. However, these problems are relatively complex and give rise to much research. There is also a challenge in the operation for transposition of the causal graph to a probabilistic representation.

REFERENCES

- BERMAN, D.S., Buczak, A.L., Chavis, J.S., and Corbett, C.L. (2019). “Survey of Deep Learning Methods for Cyber Security”, *Information* **2019**, 10, 122; doi:10.3390/info10040122.
- BLOICE, M. & Holzinger, A., (2018), *A Tutorial on Machine Learning and Data Science Tools with Python*. Graz, Austria: s.n.
- BURT, D., Nicholas, P., Sullivan, K., & Scoles, T. (2013). Cybersecurity Risk Paradox. *Microsoft SIR*.
- Government of Zimbabwe’s Various Budget Statements by the Ministry of Finance, accessed at <http://www.zimtreasury.gov.zw/>
- Government of Zimbabwe’s Various Monetary Policy Statements by the Reserve Bank of Zimbabwe, accessed at <http://www.rbz.co.zw/>
- HASHEM, I. A. T., Yaqoob, I., Anuar, N. B., Mokhtar, S., Gani, A., & Ullah Khan, S. (2015). The rise of “big data” on cloud computing: Review and open research issues. In *Information Systems*. <https://doi.org/10.1016/j.is.2014.07.006>.
- JONES, C. (1998) *Introduction to Economic Growth*, (W.W. Norton, 1998 First Edition, 2002 Second Edition).
- KABANDA, G. (2013), “African context for technological futures for digital learning and the endogenous growth of a knowledge economy”, *Basic Research Journal of Engineering Innovation (BRJENG)*, Volume 1(2), April 2013, pages 32-52, <http://basicresearchjournals.org/engineering/pdf/Kabanda.pdf>

- KABANDA, G., (2021), “*Performance of Machine Learning and Big Data Analytics paradigms in Cybersecurity and Cloud Computing platforms*”, Global Journal of Computer Science and Technology: G Interdisciplinary Volume 21, Issue 2, Version 1.0, Year 2021; Type: Double Blind Peer Reviewed International Research Journal; Publisher: Global Journals Online ISSN: 0975-4172 & Print ISSN: 0975-4350; [Performance of Machine Learning and Big Data Analytics Paradigms in Cybersecurity and Cloud Computing Platforms \(globaljournals.org\)](https://doi.org/10.1016/j.eswa.2016.08.014).
- KOTHARI , C.R.(2004) .Research Methodology Methods and Techniques 2nd Revised Edition .New Age International Publishers
- MAZUMDAR, S., and Wang, J., (2018). Big Data and Cyber security: A visual Analytics perspective in S. Parkinson et al (Eds), Guide to Vulnerability Analysis for Computer Networks and Systems.
- MENZES, F.S.D., Liska, G.R., Cirillo, M.A. and Vivanco, M.J.F. (2016) Data Classification with Binary Response through the Boosting Algorithm and Logistic Regression. Expert Systems with Applications, 69, 62-73. <https://doi.org/10.1016/j.eswa.2016.08.014>
- MURUGAN, S., and Rajan, M.S., (2014). Detecting Anomaly IDS in Network using Bayesian Network, IOSR Journal of Computer Engineering (IOSR-JCE), e-ISSN: 2278-0661, p-ISSN: 2278-8727, Volume 16, Issue 1, Ver. III (Jan. 2014), PP 01-07, www.iosrjournals.org
- NAPANDA, K., Shah, H., and Kurup, L., (2015). Artificial Intelligence Techniques for Network Intrusion Detection, International Journal of Engineering Research & Technology (IJERT), ISSN: 2278-0181, IJERTV4IS110283 www.ijert.org, Vol. 4 Issue 11, November-2015.
- NIELSEN, R. (2015). CS651 Computer Systems Security Foundations 3d Imagination Cyber Security Management Plan, Technical Report January 2015, Los Alamos National Laboratory, USA.
- SARKER, I. H., Kayes, A. S. M., Badsha, S., Alqahtani, H., Watters, P., & Ng, A. (2020). Cybersecurity data science: an overview from machine learning perspective. *Journal of Big Data*. <https://doi.org/10.1186/s40537-020-00318-5>
- SITI Nurul Mahfuzah, M., Sazilah, S., & Norasiken, B. (2017). An Analysis of Gamification Elements in Online Learning To Enhance Learning Engagement. *6th International Conference on Computing & Informatics*.
- STALLINGS, W., (2015). Operating System Stability. Accessed on 27th March, 2019. <https://www.unf.edu/public/cop4610/ree/Notes/PPT/PPT8E/CH15-OS8e.pdf>

- THOMAS, E. M., Temko, A., Marnane, W. P., Boylan, G. B., & Lightbody, G. (2013). Discriminative and generative classification techniques applied to automated neonatal seizure detection. *IEEE Journal of Biomedical and Health Informatics*.
<https://doi.org/10.1109/JBHI.2012.2237035>
- TRUONG, T.C; Diep, Q.B.; & Zelinka, I. (2020). Artificial Intelligence in the Cyber Domain: Offense and Defense. *Symmetry* 2020, 12, 410.
- UMAMAHESWARI, K., and Sujatha, S., (2017). Impregnable Defence Architecture using Dynamic Correlation-based Graded Intrusion Detection System for Cloud, *Defence Science Journal*, Vol. 67, No. 6, November 2017, pp. 645-653, DOI : 10.14429/dsj.67.11118.
- WILSON, B. M. R., Khazaei, B., & Hirsch, L. (2015, November). Enablers and barriers of cloud adoption among Small and Medium Enterprises in Tamil Nadu. In: 2015 IEEE International Conference on Cloud Computing in Emerging Markets (CCEM) (pp. 140-145). IEEE.