ISSN Application No. 21723

Journal of Current Development in Artificial Intelligence

Volume No. 11

Issue No. 2

May - August 2023

ENRICHED PUBLICATIONS

ENRICHED PUBLICATIONS PVT. LTD

S-9, IInd FLOOR, MLU POCKET,
MANISH ABHINAV PLAZA-II, ABOVE FEDERAL BANK,
PLOT NO-5, SECTOR-5, DWARKA, NEW DELHI, INDIA-110075,
PHONE: - + (91)-(11)-47026006

Journal of Current Development in Artificial Intelligence

Aims and Scope

Journal of Current Development in Artificial Intelligence is a Journal addresses concerns in applied research and applications of artificial intelligence (AI). the journal also acts as a medium for exchanging ideas and thoughts about impacts of AI research. Articles highlight advances in uses of AI systems for solving tasks in management, industry, engineering, administration, and education evaluations of existing AI systems and tools, emphasizing comparative studies and user experiences and the economic, social, and cultural impacts of AI. Papers on key applications, highlighting methods, time schedules, person months needed, and other relevant material are welcome.

Journal of Current Development in Artificial Intelligence

Journal of Current Development in Artificial Intelligence

(Volume No. 11, Issue No. 2, May - August 2023)

Contents

To Study University Smart Campus to Matching Requirements of Industrial Revolution 4.0 in Education Case Study: 'Al-Madinah International University Mediu'

[1] Shadi M. S. Hilles, [2] Abdallah Mahmoud Mousa Altrad, [3] Barjoyai Bardai

[1,2]Faculty of Computer & Info. Tech,

[1,2,3,]Al-Madinah International University (MEDIU)

E-mail: [1]shadihilless@gmail.com

ABSTRACT

Cyber-physics system is combination of physics with cyber components potentially networked and interconnected devices, individually any of these would have a large impact. together, it is a perfect storm of change, Cloud computing uses remote servers to store, manage and process data for faster processes, 3D printing lab, robotics and IoT lab in smart University, Augmented reality is impact smart classroom to allow student to study theoretical subjects with interaction in 3D data, Big data can determine future actions and improve evaluation and analyzing and cyber Security protects smart campus most valuable data, Internet of things IoT which connected internet to machine to send, receive and process data, Simulation ideal for training university staff and scenario planning, Autonomous systems program machinery and robots to act autonomously, Therefore, the 4.0 IR makes a University interoperability which is machine, devices and people connect and communicate with one another. The aim of this study is to propose conceptual model of smart university campus and smart education based on new method of teaching and advance technology and identify the main distinctive features, components, technologies and systems of smart university, the qualitative study and interview shows that the five academic programs from information Technology faculty and Business faculty small percentage of IR 4.0 areas are included in the curriculum used nine pillars of cyber-physical systems.

I. INTRODUCTION

Nowadays, the increased and rapid changed on technology in several sectors such as smart home, smart campus due to interconnected devices and high bandwidth of internet speed, where is data store in cloud computing in real time, and due to data generated from several resources while Artificial Intelligent became much more complexity algorithm, therefore, the new era which is represented the advance technology in embedding systems or it's called cyber-physical systems the era is called Industrial revolution 4.0 IR 4.0 or Industry 4.0 [1], the IR 4.0 has potential to disrupted the existing conventional approach in enterprise, manufactories and as well university campus, therefore, there is interoperable systems decentralization, transparency and making decision systems, those allow to coordinated more effort among industrial, enterprise, universities and even countries [2], the Internet communication technology ICT facilities smart machines, storage systems and production facilities capable of autonomously exchanging information and also called smart factory, The term of IR 4.0 has been introduced in 2013 by Germany group in academic, business and politics to adopted new era of high tech- strategy for 2020, and was revived in 2011 at the Hannover Fair. There are number of technologies making up IR4.0 such as AI, Robotics, Automation, genetic engineering/biotechnology, Nano technology, Most of these are autonomous, things that happen without human intervention.

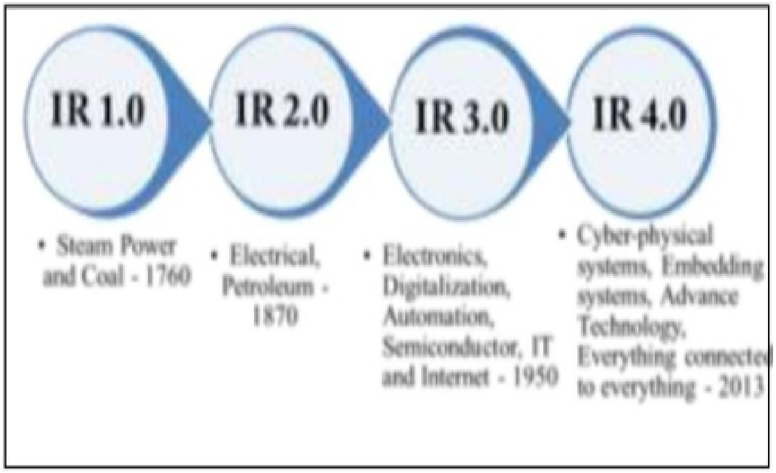

IR 1.0
IR 2.0
IR 3.0
IR 4.0
• Steam Power and Coal - 1760
• Electrical, Petroleum - 1870
• Electronics, Digitalization, Automation, Semiconductor, IT and Internet - 1950
• Cyber-physical systems, Embedding systems, Advance Technology, Everything connected to everything - 2013

Fig1. History of Industrial Revolution

Figure 1 is represented the historical of industrial revolution where is first industrial is mechanical industrial and it was based on steam power and coal, the second is electrical industrial and represented by steal, petroleum and electrical, the third industrial has several name such as electronic, digital automation production and sustainable, digitalization, semiconductors, information technology and internet, while the fourth industrial is called cyber-physical systems means everything is connected to everything digital, physical and biological[3].

Cyber-physics system is combination of physics with cyber components potentially networked and interconnected devices.

Individually any of these would have a large impact. together, it is a perfect storm of change, Cloud computing uses remote servers to store, manage and process data for faster processes, Adaptive manufacturing which is digital 3D design data builds in layers by depositing materials, Augmented reality save expense of a physical trail to show a product, Big data can determine future actions and improve processes and cyber Security protects manufacturer's most valuable data, Internet of things IoT which connected internet to machine to send, receive and process data, Simulation ideal for training employees and scenario planning, Autonomous systems program machinery and robots to act autonomously, Therefore, the 4.0 IR makes a factory interoperability which is machine, devices and people connect and communicate with one another, information transparency system: create virtually copy of physical world through sensor data, decentralization decision- making which is cyber-physical systems make simple decisions become as autonomous as possible, technical assistance: the system support humans in making decisions and solving problems and assist with tasks, and every industrial is planning substantial investments.

What are the challenges for the workforce? Increased expectations regarding individual flexibility, Change of necessary qualification, Improvement through the use of robots, Create of new jobs for high skilled workers.

The proposed project is based on improvement of building services efficiency, smart services and smart tools.

The IR 4.0 consist nine pillars of advance technology which are impact to education and new method of teaching. Those pillars as shown on figure below:

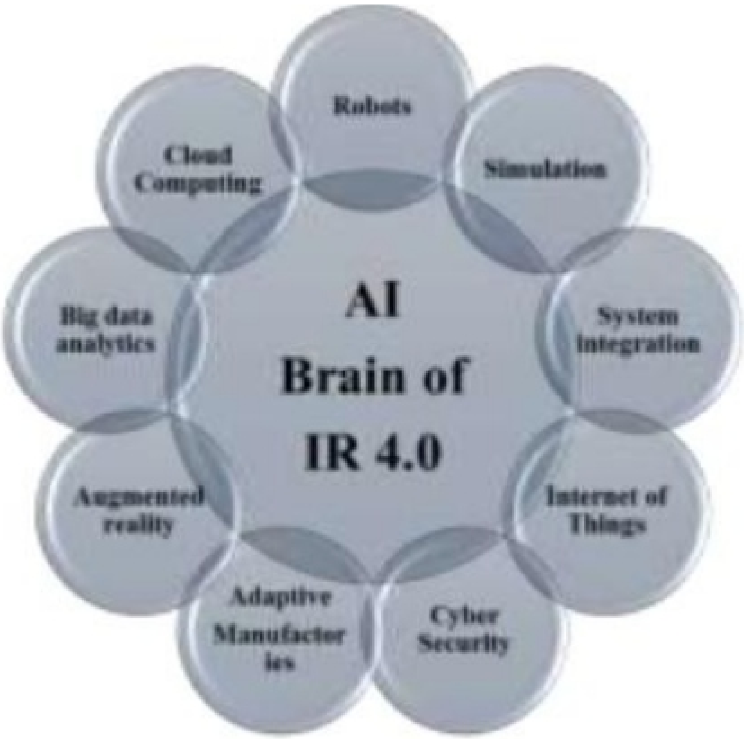


Fig.2. IR 4.0 and nine pillars of advance technology Objectives of smart University

One of requirements to enable smart university is smart education, which is needed to enhancing curriculum and activities in order to offered IR 4.0 advance technology in curriculum, these technologies as illustrated on figure 2 play important role education to increase effectiveness of education in classroom and eLearning and for knowledge to be effectively on future for student by sharing and training, the innovative and advance technology such as augmented and virtual reality, cloud computing which allow to enhance the effectiveness of data management in smart campus [5].

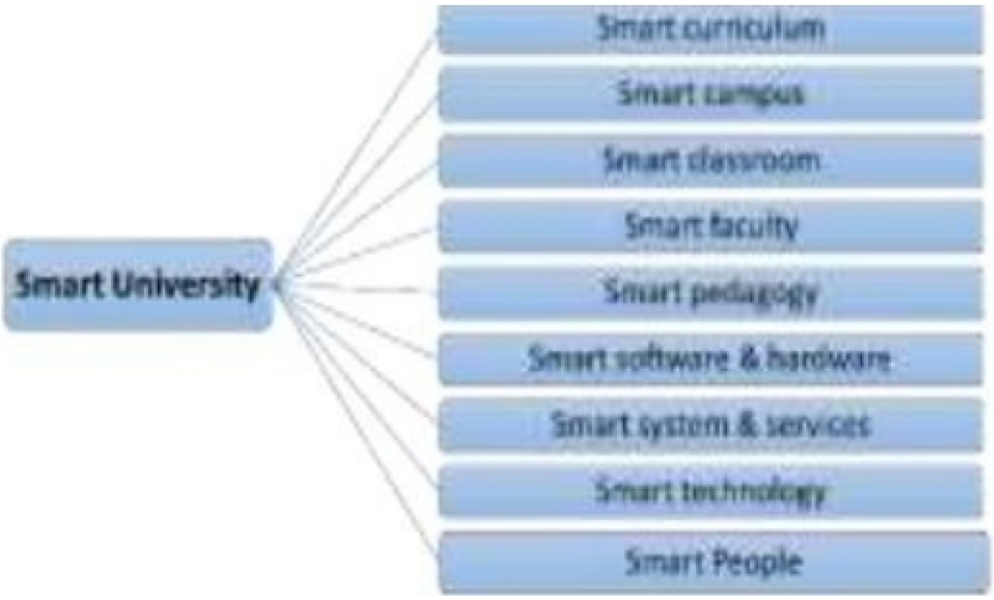


Fig. 3. University Smart Campus

In Figure 3. illustrated University smart campus and figure 4 presented framework of education 4.0, teaching factors to upgrade traditional learning factors by emerging digital technologies [5], the Learning factory module is presented and has been discussed [6], To make university services and facilities much more agile and flexible, ease and fast, responsive to students needs and promote competitive of advance technology IR 4.0 to able to move to smart campus, it possible to create smart campus by using wireless sensors and many other advance technology IR 4.0 used[4].New infrastructure of interconnected devises IoT, and application of IoT. In this paper present smart campus framework as shown on the figure 3.

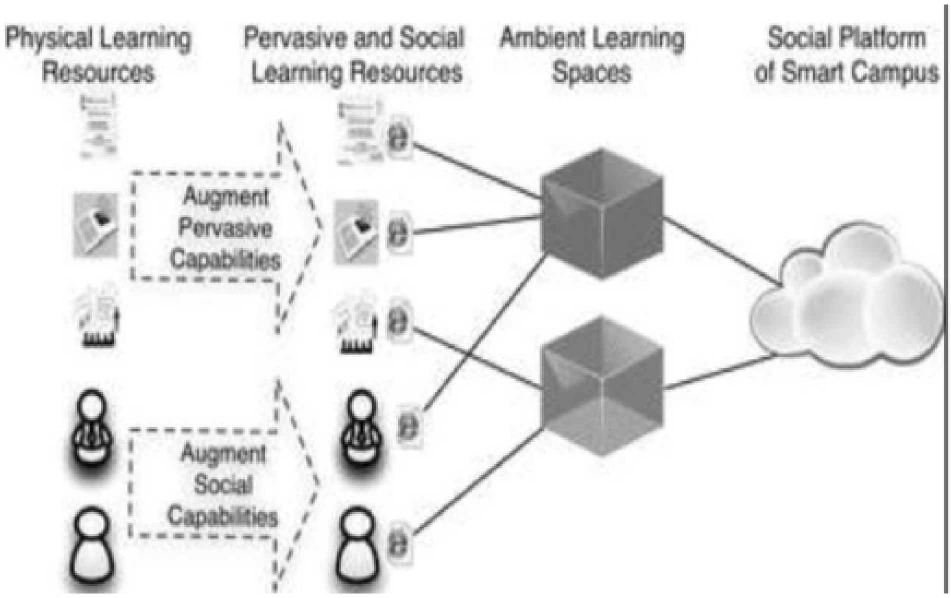
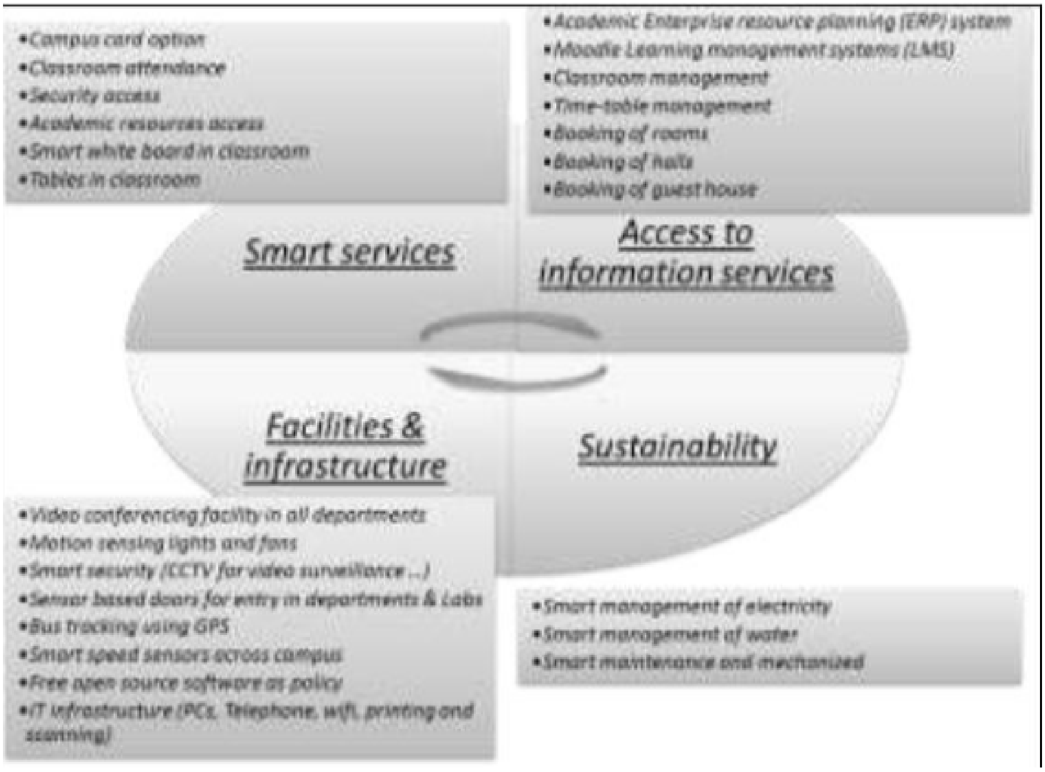

Physical Learning Resources
Pervasive and Social Learning Resources
Ambient Learning Spaces
Social Platform of Smart Campus
Augment Pervasive Capabilities
Augment Social Capabilities

Fig4. Smart campus framework

II. PROPOSED CONCEPTUALAPPROACH OF UNIVERSITY SMART CAMPUS

Objective of research paper is to propose conceptual model of universities smart campus used smart campus framework as a system based on smart university framework, level of a smart system, smart classrooms, smart faculty, smart pedagogy, smart software and hardware systems, smart technology, smart curriculum, smart campus and smart services, and to identify the main distinctive features, components, technologies and systems of smart university. Ref toMalaysia Education Blueprint 2015-2025 (Higher Education) [7, 8]



•Campus card option
•Classroom attendance
•Security access
•Academic resources access
•Smart white board in classroom
•Tables in classroom
•Academic Enterprise resource planning (ERP) system
•Moodle Learning management systems (LMS)
•Classroom management
•Time-table management
•Booking of rooms
•Booking of halls
•Booking of guest house
Smart services
Access to information services
Facilities & infrastructure
Sustainability
•Video conferencing facility in all departments
•Motion sensing lights and fans
•Smart security (CCTV for video surveillance ...)
•Sensor based doors for entry in departments & Labs
•Bus tracking using GPS
•Smart speed sensors across campus
•Free open source software as policy
•IT infrastructure (PCs, Telephone, wifi, printing and scanning)
•Smart management of electricity
•Smart management of water
•Smart maintenance and mechanized

Fig. 5 Architecture of the conceptual proposed approach of university smart campus

III. SMART UNIVERSITY TAXONOMY: FEATURES, COMPONENTSAND SYSTEMS

The idea of next generation of smart classroom system that should significant emphasizes not only software/hardware features but also smart features and functionality of smart systems. Therefore, next generation of smart classrooms should pay more attention to implementation of smartness to create taxonomy of a smart university, smart features, smart services, and identify components and systems that go well beyond those a traditional university face-to-face classes and learning activities.

As illustrated on figure 4, the new approach of university smart campus is represented the four important factors such as smart services, access to information services, facilities & infrastructure and sustainability.

Identification features of Software systems to be developed as following:

Web-lecturing systems (with video capturing and computer screen capturing function) for learning content Smart classroom in-class activities recording systems Smart cameraman software systems Systems for seamless collaborative learning for both OC and OL in smart classroom and sharing learning content/ documents.

Collaborative web-based audio/video one-to-one and many-to-many communication systems

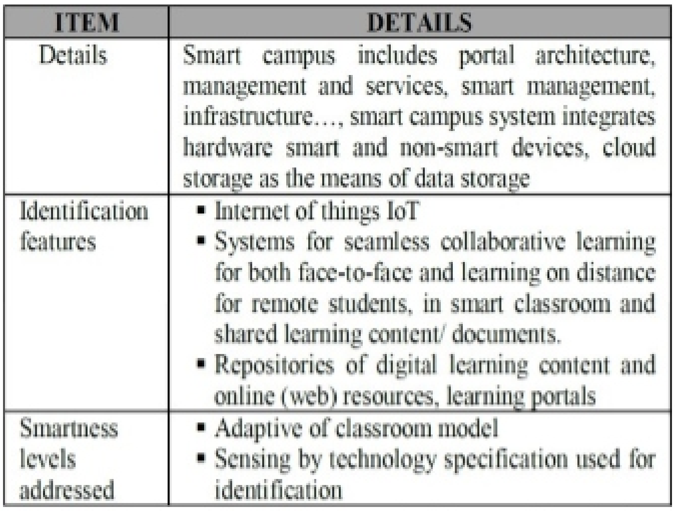
System to reply automatically recorded class activities and lectures.

Repository of digital learning content and online Smart learning analytics and smart teaching analytics systems

- Speaker/ instructor motion tracking systems
- Speech/ voice recognition systems
- Speech-to-text systems
- Text-to-voice systems.
- Face recognition systems
- Emotion recognition systems
- Various smart software agents
- Automatic translation systems

A. Construct smart campus based on cloud computing platform and internet of things IoT

The approach used is Smart campus development approach is based is based on higher stage of education information systems that connect everything through RFID, technology, sensors and Internet of things IoT.

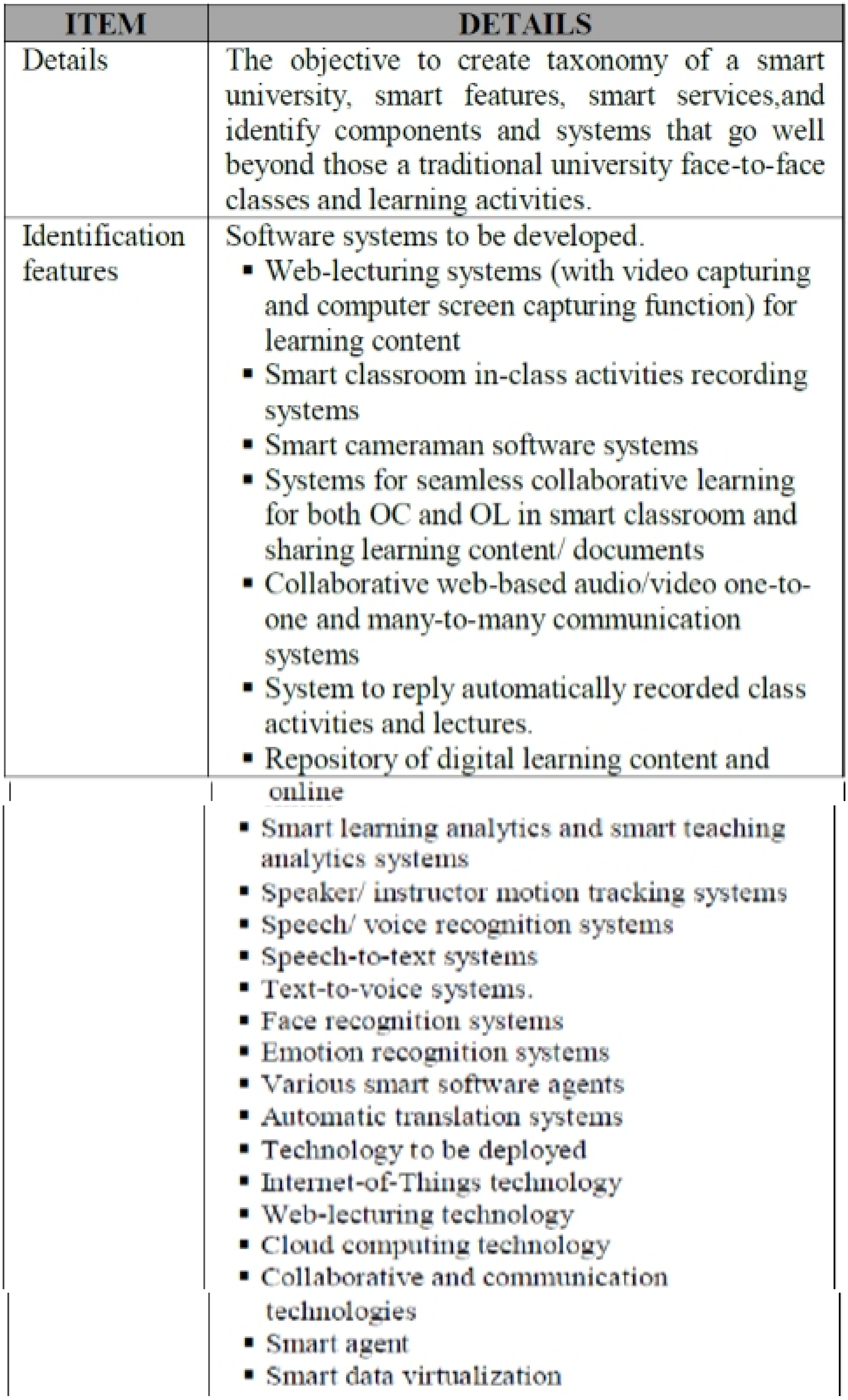


ITEM	DETAILS
Details	Smart campus includes portal architecture, management and services, smart management, infrastructure…, smart campus system integrates hardware smart and non-smart devices, cloud storage as the means of data storage
Identification features	▪ Internet of things IoT ▪ Systems for seamless collaborative learning for both face-to-face and learning on distance for remote students, in smart classroom and shared learning content/ documents. ▪ Repositories of digital learning content and online (web) resources, learning portals
Smartness levels addressed	▪ Adaptive of classroom model ▪ Sensing by technology specification used for identification

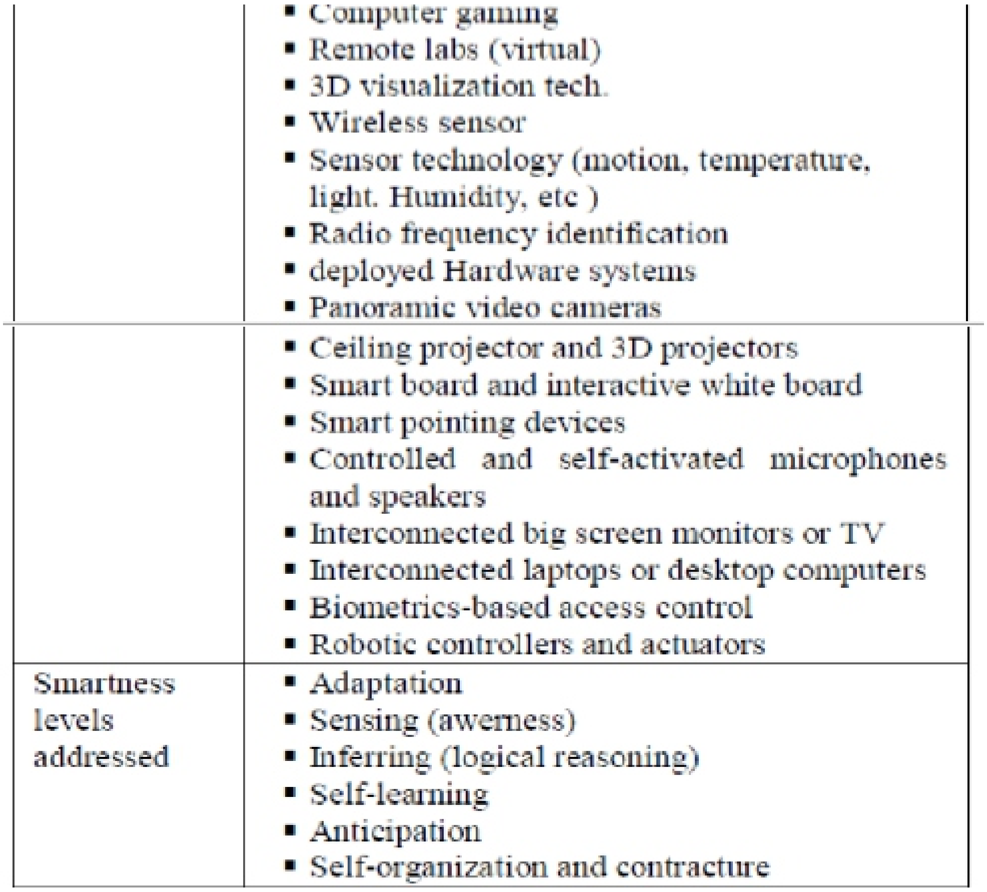
TABLE I: Smart campus based on cloud computing platform and internet of things IoT

B. Smart University taxonomy: features, components and systems.

The approach used the idea of next generation of smart classroom system that should significant emphasize not only software/hardware features but also smart features and functionality of smart systems. Therefore, next generation of smart classrooms should pay more attention to implementation of smartness.



ITEM	DETAILS
Details	The objective to create taxonomy of a smart university, smart features, smart services,and identify components and systems that go well beyond those a traditional university face-to-face classes and learning activities.
Identification features	Software systems to be developed. ▪ Web-lecturing systems (with video capturing and computer screen capturing function) for learning content ▪ Smart classroom in-class activities recording systems ▪ Smart cameraman software systems ▪ Systems for seamless collaborative learning for both OC and OL in smart classroom and sharing learning content/ documents ▪ Collaborative web-based audio/video one-to-one and many-to-many communication systems ▪ System to reply automatically recorded class activities and lectures. ▪ Repository of digital learning content and online ▪ Smart learning analytics and smart teaching analytics systems ▪ Speaker/ instructor motion tracking systems ▪ Speech/ voice recognition systems ▪ Speech-to-text systems ▪ Text-to-voice systems. ▪ Face recognition systems ▪ Emotion recognition systems ▪ Various smart software agents ▪ Automatic translation systems ▪ Technology to be deployed ▪ Internet-of-Things technology ▪ Web-lecturing technology ▪ Cloud computing technology ▪ Collaborative and communication technologies ▪ Smart agent ▪ Smart data virtualization

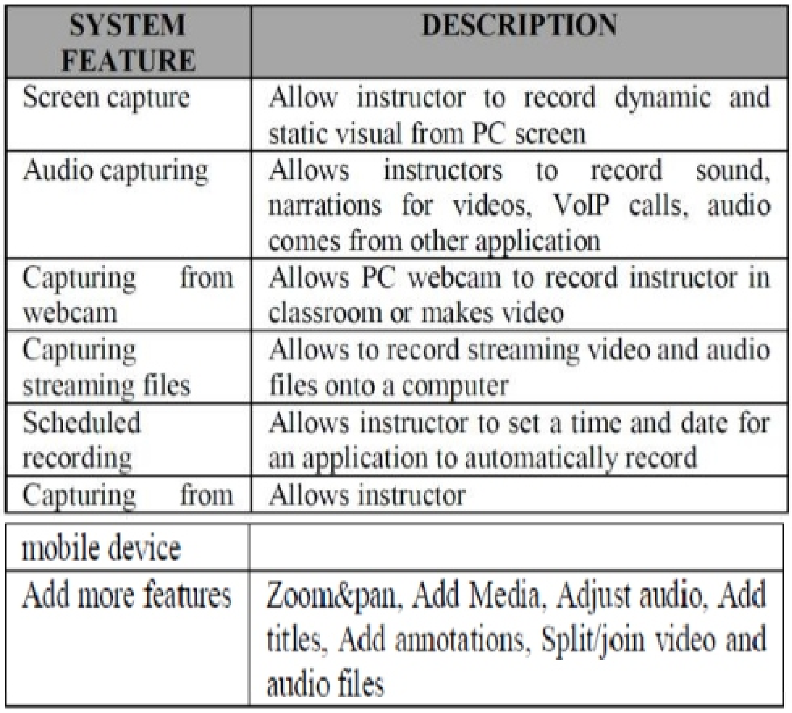


	▪ Computer gaming ▪ Remote labs (virtual) ▪ 3D visualization tech. ▪ Wireless sensor ▪ Sensor technology (motion, temperature, light. Humidity, etc) ▪ Radio frequency identification ▪ deployed Hardware systems ▪ Panoramic video cameras
	▪ Ceiling projector and 3D projectors ▪ Smart board and interactive white board ▪ Smart pointing devices ▪ Controlled and self-activated microphones and speakers ▪ Interconnected big screen monitors or TV ▪ Interconnected laptops or desktop computers ▪ Biometrics-based access control ▪ Robotic controllers and actuators
Smartness levels addressed	▪ Adaptation ▪ Sensing (awerness) ▪ Inferring (logical reasoning) ▪ Self-learning ▪ Anticipation ▪ Self-organization and contracture

TABLE 2: Smart campus based on cloud computing platform and internet of things IoT

C. System Features needed

The features needed in smart classroom as shown on the table below.



SYSTEM FEATURE	DESCRIPTION
Screen capture	Allow instructor to record dynamic and static visual from PC screen
Audio capturing	Allows instructors to record sound, narrations for videos, VoIP calls, audio comes from other application
Capturing from webcam	Allows PC webcam to record instructor in classroom or makes video
Capturing streaming files	Allows to record streaming video and audio files onto a computer
Scheduled recording	Allows instructor to set a time and date for an application to automatically record
Capturing from mobile device	Allows instructor
Add more features	Zoom&pan, Add Media, Adjust audio, Add titles, Add annotations, Split/join video and audio files

TABLE 3: Systems Features Needed

IV. DATA COLLECTION

The data collection stage consisted of 12 individual interviews carried out with Computer and business professors from various disciplines, who teach design courses in the Faculties, the interviews were one hour long included written questionnaire based on method of teaching and offered subjects related to advance technologies included cognitive in blooms taxonomy.

V. RESULT SURVEY AND QUALITATIVE

In this section shows samples of questionnaire of MEDIU head departments and staff lecturer faculties of Computer Science & Info Tech. and also from faculty of business and administration.

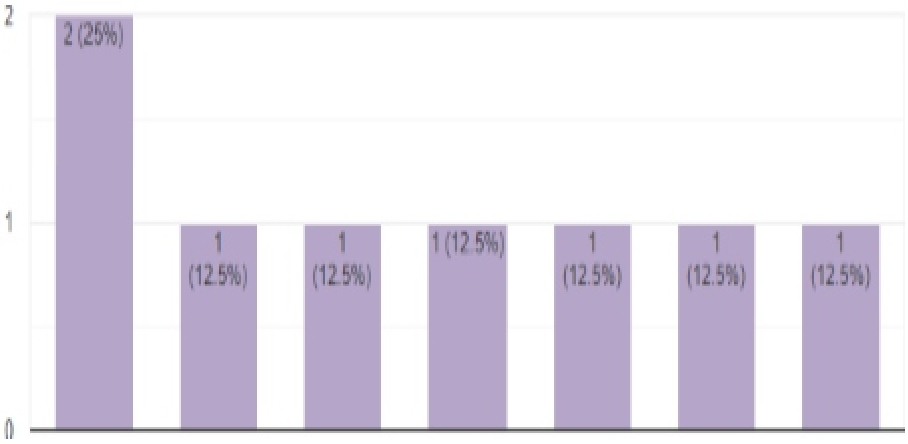
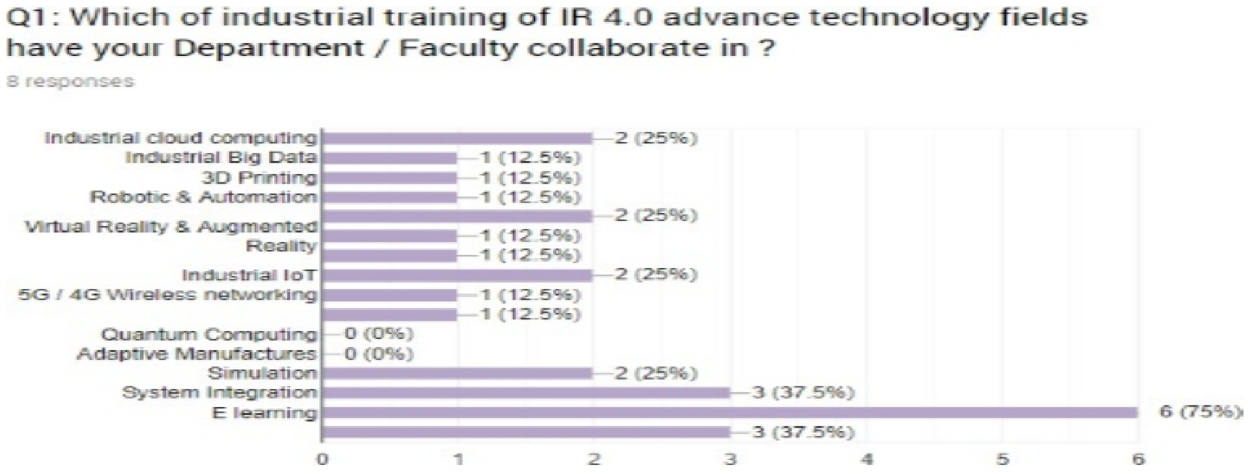


Figure 5. Illustrated percentage of teaching subjects in areas of advance technologies

The questions samples in section of department vision

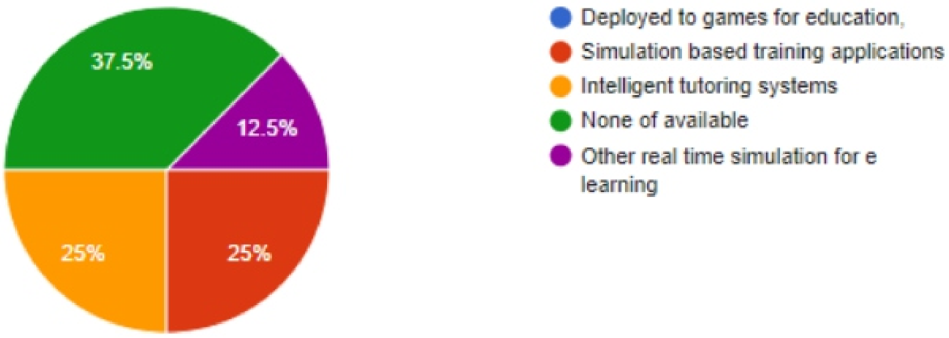
First Question sample:


Q1: Which of industrial training of IR 4.0 advance technology fields have your Department / Faculty collaborate in ?

8 responses

Q7: Which method of teaching and E-learning for modelling and simulation in real time is used ?.

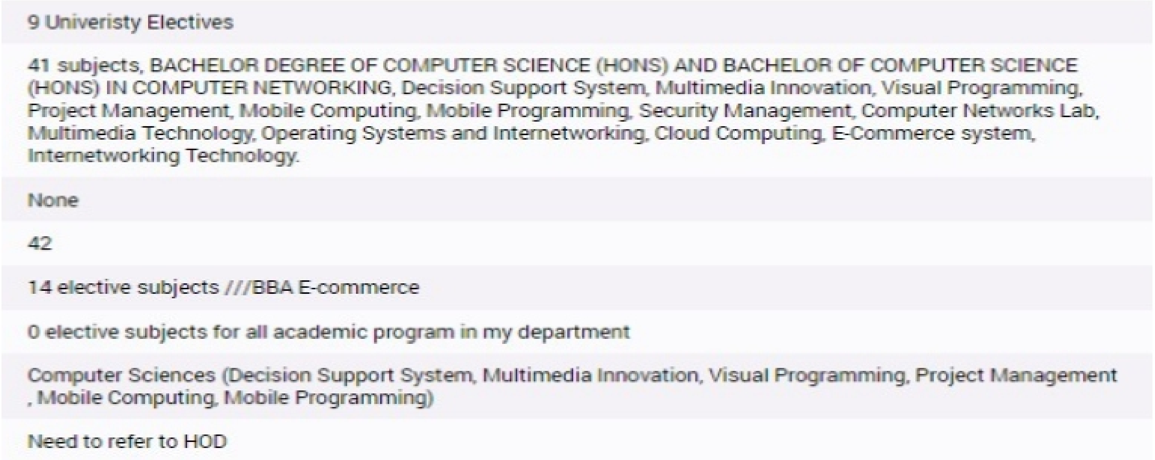
8 responses



First question sample shows percentage of teachning subject in areas of advance technologies

Q1: How many elective subjects in Academic programs are offered for Bachelor degree in your department? Please put the name of academic program and elective subjects

8 responses



9 Univeristy Electives

41 subjects, BACHELOR DEGREE OF COMPUTER SCIENCE (HONS) AND BACHELOR OF COMPUTER SCIENCE (HONS) IN COMPUTER NETWORKING, Decision Support System, Multimedia Innovation, Visual Programming, Project Management, Mobile Computing, Mobile Programming, Security Management, Computer Networks Lab, Multimedia Technology, Operating Systems and Internetworking, Cloud Computing, E-Commerce system, Internetworking Technology.

None

42

14 elective subjects ///BBA E-commerce

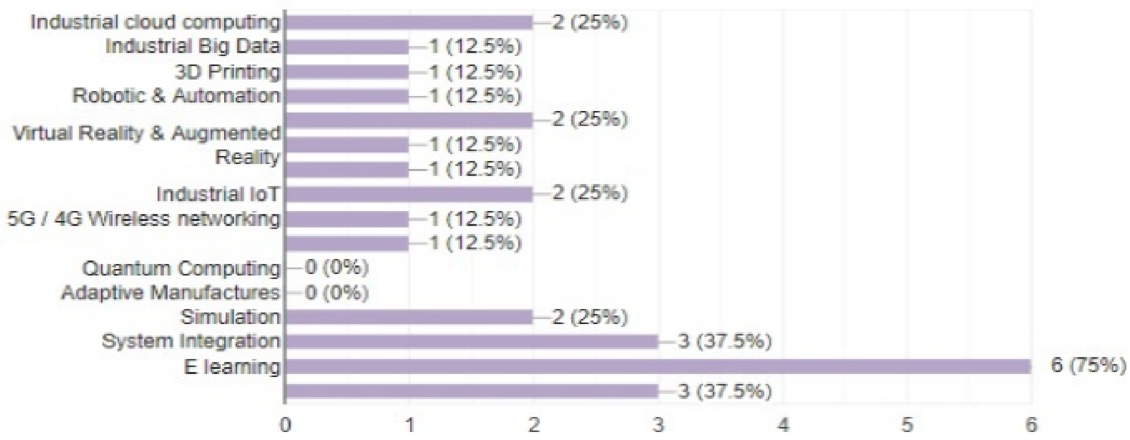
0 elective subjects for all academic program in my department

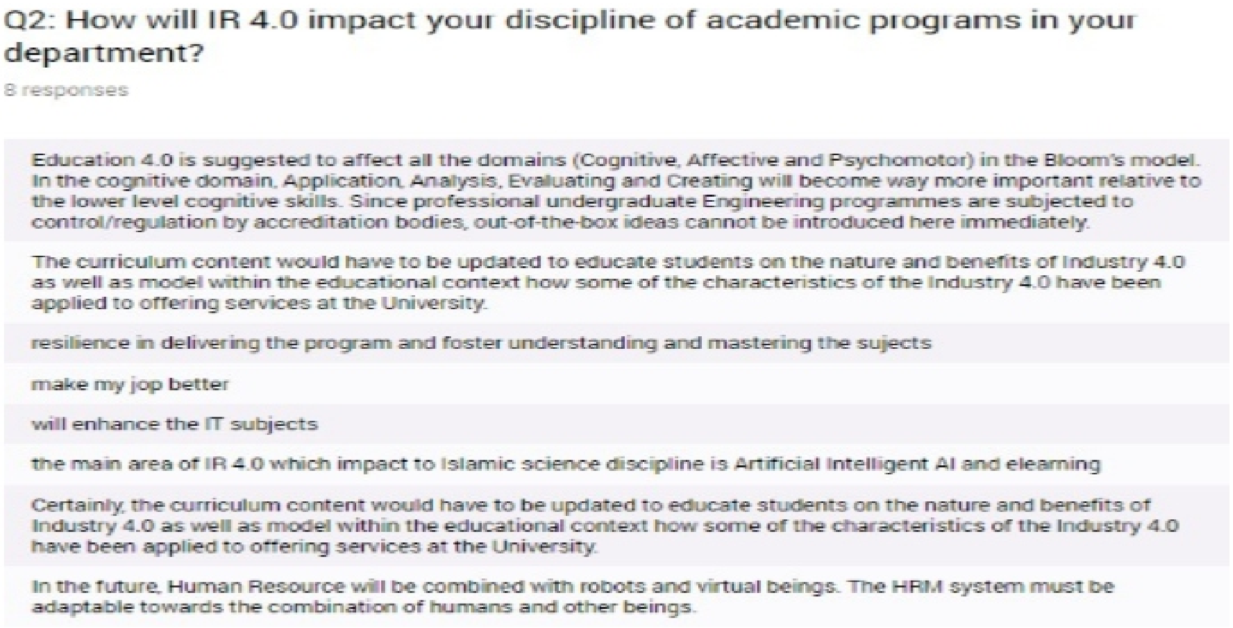
Computer Sciences (Decision Support System, Multimedia Innovation, Visual Programming, Project Management , Mobile Computing, Mobile Programming)

Need to refer to HOD

Q1: Which of industrial training of IR 4.0 advance technology fields have your Department / Faculty collaborate in ?

8 responses





Q2: How will IR 4.0 impact your discipline of academic programs in your department?

8 responses

Education 4.0 is suggested to affect all the domains (Cognitive, Affective and Psychomotor) in the Bloom's model. In the cognitive domain, Application, Analysis, Evaluating and Creating will become way more important relative to the lower level cognitive skills. Since professional undergraduate Engineering programmes are subjected to control/regulation by accreditation bodies, out-of-the-box ideas cannot be introduced here immediately.

The curriculum content would have to be updated to educate students on the nature and benefits of Industry 4.0 as well as model within the educational context how some of the characteristics of the Industry 4.0 have been applied to offering services at the University.

resilience in delivering the program and foster understanding and mastering the sujects

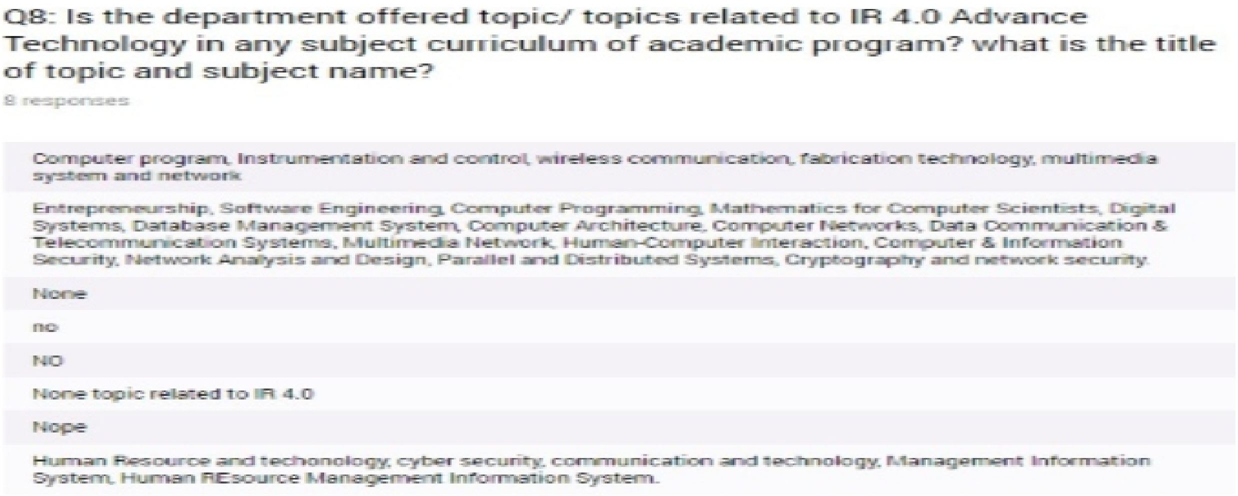
make my jop better

will enhance the IT subjects

the main area of IR 4.0 which impact to Islamic science discipline is Artificial Intelligent AI and elearning

Certainly, the curriculum content would have to be updated to educate students on the nature and benefits of Industry 4.0 as well as model within the educational context how some of the characteristics of the Industry 4.0 have been applied to offering services at the University.

In the future, Human Resource will be combined with robots and virtual beings. The HRM system must be adaptable towards the combination of humans and other beings.



Q8: Is the department offered topic/ topics related to IR 4.0 Advance Technology in any subject curriculum of academic program? what is the title of topic and subject name?

8 responses

Computer program, Instrumentation and control, wireless communication, fabrication technology, multimedia system and network

Entrepreneurship, Software Engineering, Computer Programming, Mathematics for Computer Scientists, Digital Systems, Database Management System, Computer Architecture, Computer Networks, Data Communication & Telecommunication Systems, Multimedia Network, Human-Computer Interaction, Computer & Information Security, Network Analysis and Design, Parallel and Distributed Systems, Cryptography and network security.

None

no

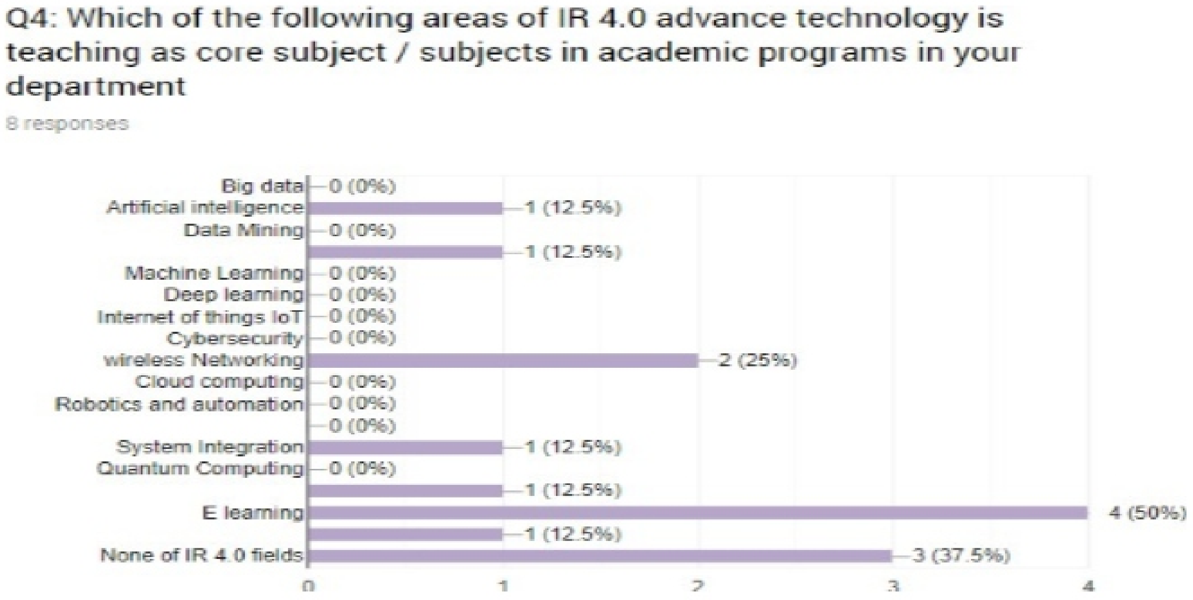
NO

None topic related to IR 4.0

Nope

Human Resource and techonology, cyber security, communication and technology, Management Information System, Human REsource Management Information System.

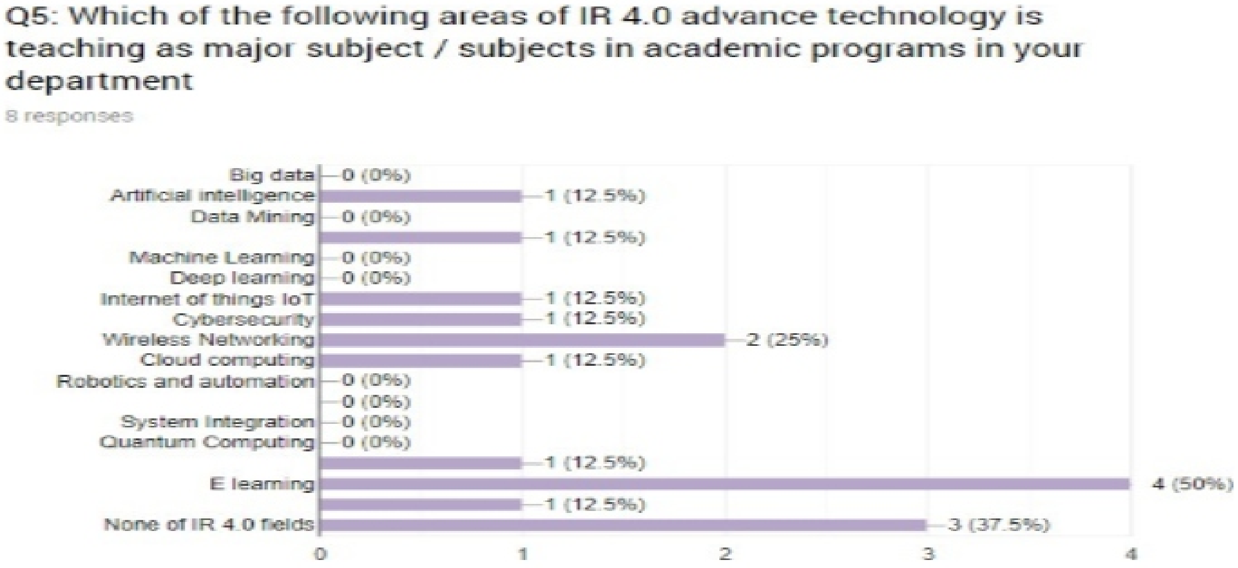
Figure shows teaching subjects in areas of IR 4.0 as core subject in academic program.



Q4: Which of the following areas of IR 4.0 advance technology is teaching as core subject / subjects in academic programs in your department

8 responses

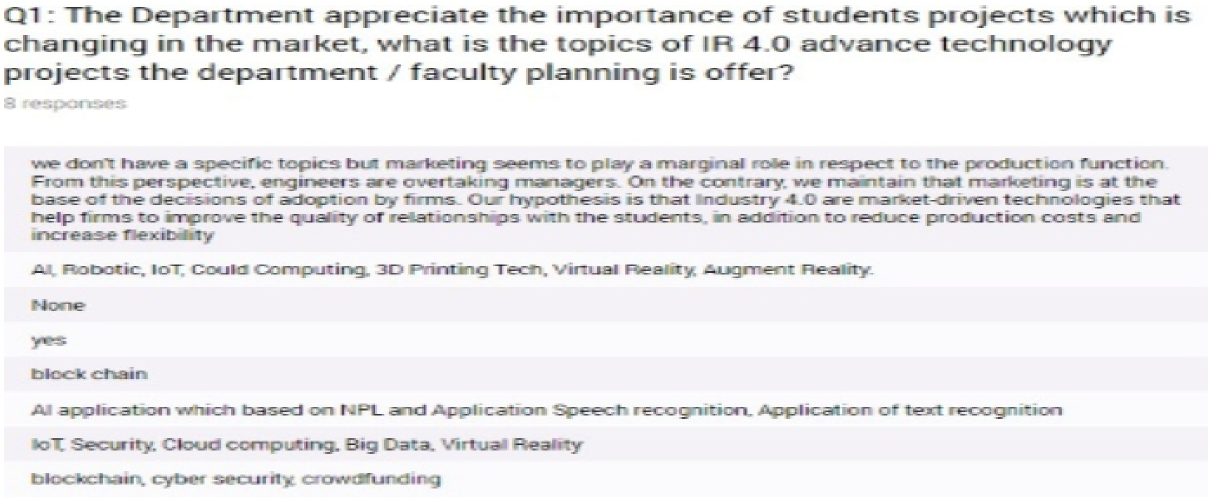
The question sample shows teaching major subjects in area of IR 4.0,



Q5: Which of the following areas of IR 4.0 advance technology is teaching as major subject / subjects in academic programs in your department

8 responses

Below question sample in department vision of teaching subjects in areas of advance technologies.



Q1: The Department appreciate the importance of students projects which is changing in the market, what is the topics of IR 4.0 advance technology projects the department / faculty planning is offer?

8 responses

we don't have a specific topics but marketing seems to play a marginal role in respect to the production function. From this perspective, engineers are overtaking managers. On the contrary, we maintain that marketing is at the base of the decisions of adoption by firms. Our hypothesis is that Industry 4.0 are market-driven technologies that help firms to improve the quality of relationships with the students, in addition to reduce production costs and increase flexibility

AI, Robotic, IoT, Could Computing, 3D Printing Tech, Virtual Reality, Augment Reality.

None

yes

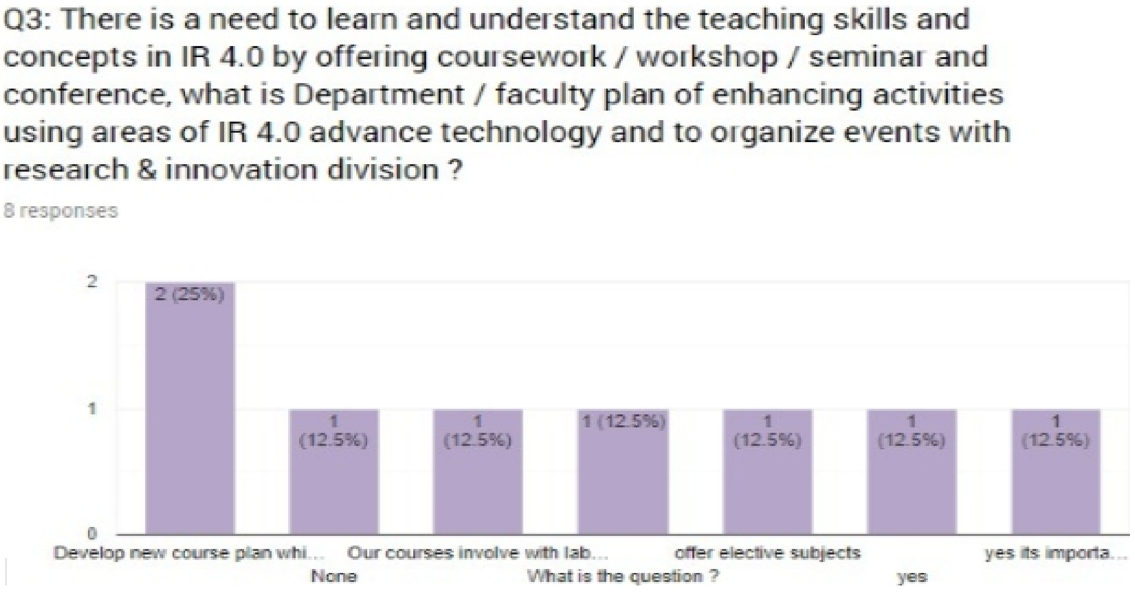
block chain

AI application which based on NPL and Application Speech recognition, Application of text recognition

IoT, Security, Cloud computing, Big Data, Virtual Reality

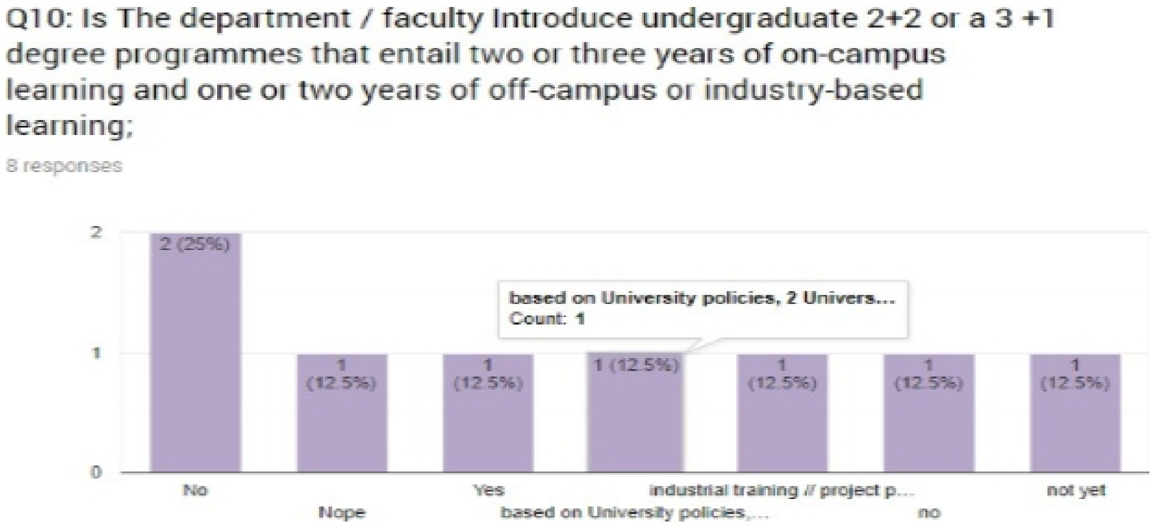
blockchain, cyber security, crowdfunding

Question shows department plan and vision in organize events workshop, seminar



Q3: There is a need to learn and understand the teaching skills and concepts in IR 4.0 by offering coursework / workshop / seminar and conference, what is Department / faculty plan of enhancing activities using areas of IR 4.0 advance technology and to organize events with research & innovation division ?

8 responses


Q10: Is The department / faculty Introduce undergraduate 2+2 or a 3 +1 degree programmes that entail two or three years of on-campus learning and one or two years of off-campus or industry-based learning;

8 responses

VI. CONCLUSION

This research paper presented cyber-physical systems and nine pillars of industrial revolution 4.0, the requirement of university smart campus features, proposed conceptual model for smart campus where studies the needs services and devices for smart lab, smart classroom and smart services all important aspects to enable smart campus in the university.

This paper presented the collaboration of human and machine in education systems. And skills need for handling combined task elements. Collaboration of human and intelligent machines triggers mutual learning.

REFERENCE

[1] Santos, C., Mehrsai, A., Barros, A. C., Araújo, M., & Ares, E. (2017). Towards Industry 4.0: an overview of European strategic roadmaps. Procedia Manufacturing, 13, 972-979.

[2] Lee, J., Bagheri, B., & Kao, H. A. (2015). A cyber-physical systems architecture for industry 4.0-based manufacturing systems. Manuf Lett 3: 18–23.

[3] Vaidya, S., Ambad, P., &Bhosle, S. (2018). Industry 4.0–a glimpse. Procedia Manufacturing, 20, 233-238.

[4] Tabaa, M., Chouri, B., Saadaoui, S., &Alami, K. (2018). Industrial Communication based on Modbus and Node-RED. Procedia Comput. Sci, 130, 583-588.

[5] Mourtzis, D., Vlachou, E., Dimitrakopoulos, G., &Zogopoulos, V. (2018). Cyber-Physical Systems and Education 4.0-The Teaching Factory 4.0 Concept. Procedia Manuf, 23, 129-134.

[6] Wienbruch, T., Leineweber, S., Kreimeier, D., &Kuhlenkötter, B. (2018). Evolution of SMEs towards Industrie 4.0 through a scenario based learning factory training. Procedia Manufacturing, 23, 141-146.

[7] Karim, M. S. A. (2016). Entrepreneurship education in an engineering curriculum. Procedia Economics and Finance, 35, 379-387.

[8] Ministry of education, 2015, https://www.kooperation-international.de/uploads/media/3._Malaysia_Education_Blue print_2015-2025 Higher_Education__.pdf.

Development of an Intelligent System for Detecting Mobile App Install Fraud

Tetiana Polhul
Vinnytsia National Technical University Ukraine
E-mail: tanapolg93@gmail.com

ABSTRACT

The paper is devoted to the development of an intelligent system for detecting mobile app install fraud. The developed system consists of the following blocks: block of user data characteristics identifying; data heterogeneity resolution block; classification model training block; classification block; knowledge base (for detecting fraud) formation block; block of fraudsters database creation; data mining and user patterns formation block; general fraudster portrait prediction block. Software for the system was developed based on the proposed methods and models during the mobile app install fraud detection intelligent system design. The values of sensitivity (99%) and specificity (80%) of the development system are high, and it means that the developed intelligent system for detecting mobile app install fraud has adequate results.

Keywords - Fraud Detection, Mobile Application, Data Mining, Intelligent System.

I. INTRODUCTION

The presence of fraud during mobile applications installation has recently become a widespread and significant (expensive) problem. For example, the AppLift study "Fighting Mobile Fraud in the Programmatic Era" [1] showed that, overall, 34% of mobile traffic that was observed was fraudulent. It costs more than $ 4.5 billion in losses. The objective of fraud during mobile applications installation is fraudulent data entry in the application. Since mobile application-developing companies invest in marketing campaigns, which is intended to result in engaging new organic users, the purpose of such fraud is to withdraw funds from the developer companies without engaging organic users in return. At this stage, let's consider the known ways of fraud during mobile applications installation: mobile hijacking [2, 3], click spamming [2 – 5], action farms [2, 3]. With this in consideration, it should be noted that an automatic intelligent system development for detecting mobile app install fraud is of great significance, namely the intellectual component of the system, that will allow finding fraudsters with new characteristics.

At the moment, there are several systems designed to detect fraud. Let's consider these systems by dividing them into groups. The first group consists of Fraudlogix [6] and Kraken [7], as these systems detect fraudsters by only one feature and use the existing fraud databases, so there is the possibility of uncertainty among many new fraudsters. The second group consists of Adjust [8], Kochava [9], and

TCM Attribution Analytics [10]. Systems of this group also do not use all available user data, that reduces the effectiveness of fraud detection by these systems and make fraud detection impossible for events with new features. The Protect360 [11] by Appsflyer is in the third group – it also does not use all criterions and makes fraud detection based on the existing database of fraudsters and has the same disadvantage as the Forensiq system – the inability to find out the reason why a user was labeled as a fraudster, which is often an important task.

The fourth group consists of FraudScore [12] system and the AppMetrica [13] system that just integrated with FraudScore, unlike most previous systems, have their own updated algorithms and a self-learning system based on the neural network, but the systems do all of their assessments only based on some data, so fraudsters with other data may be omitted. The above systems do not use all available data, which reduces the effectiveness of fraud detection. Therefore, there was a need to develop a fraud detection system that, based on all available incoming data about mobile application users, will provide not only certain fraud detection, but also determine fraudsters patterns. Thus, it would be a possibility to solve the prediction and the fraudsters portraits creation problems using data mining based on determined patterns. This article shows the proposed approach to the development of an intelligent system for detecting mobile app install fraud.

II. INTELLIGENT SYSTEM FOR DETECTING MOBILE APP INSTALL FRAUD

The authors have proposed a fraud detection system based on data mining [14, 15], which consists of the following blocks (Fig. 1): block of user data characteristics identifying; data heterogeneity resolution block; classification model training block; classification block; knowledge base (for detecting fraud) formation block; block of fraudsters database creation; data mining and user patterns formation block; general fraudster portrait prediction block.

The presence of fraudsters is considered as an anomaly inlarge arrays of heterogeneous data about users during the design of the system.

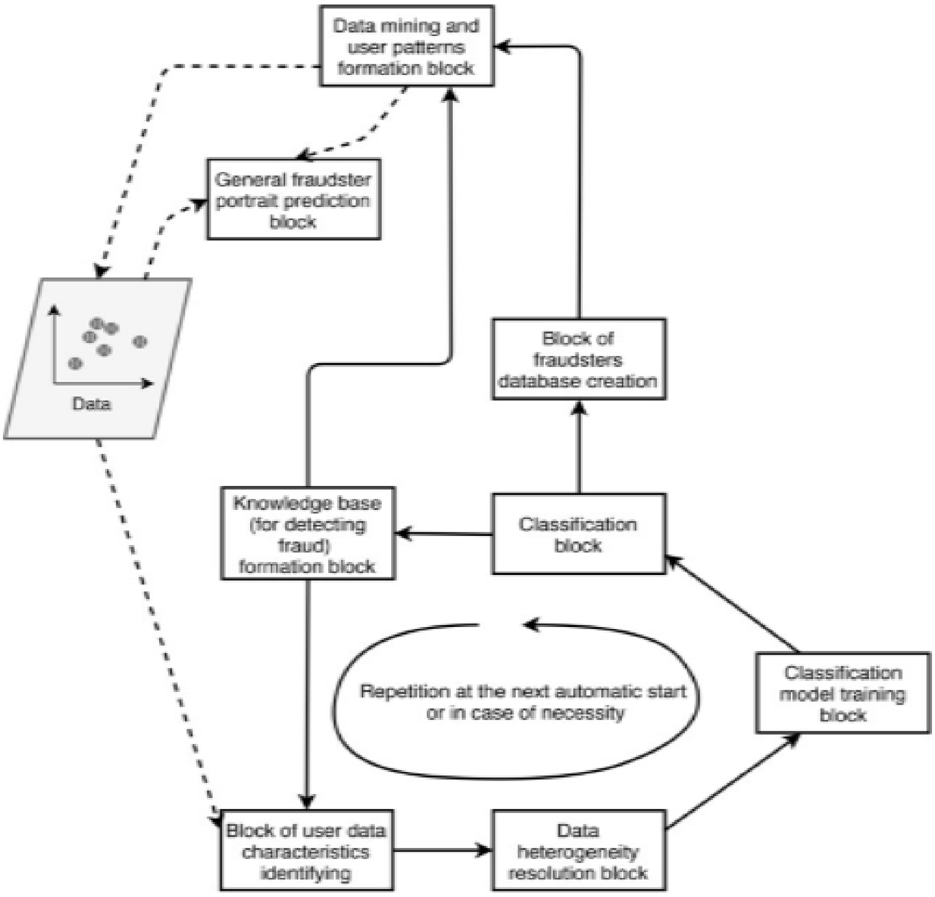

Data mining and user patterns formation block
General fraudster portrait prediction block
Data
Block of fraudsters database creation
Knowledge base (for detecting fraud) formation block
Classification block
Repetition at the next automatic start or in case of necessity
Classification model training block
Block of user data characteristics identifying
Data heterogeneity resolution block

Fig.1. Fraud detection system based on data mining.

There was a problem of processing heterogeneous data in the process of developing the system. For this purpose, a generalized model of heterogeneous data and a method for data heterogeneity resolution were proposed in this paper using a scaling procedure based on the proposed four scales, which formed the basis for the block of data heterogeneity resolution. The transition from large data sets to two groups of coefficients was performed in the process of data heterogeneity resolution. Identified fraudsters are entered into the database, and their patterns are formed on the basis of the first groupof coefficientsG1 (Fig. 2). Unknown in the first stage fraudsters are identified and a knowledge base is formed in the process of analysis of the second groupG2 (Fig. 3), which based on the finding of similarity coefficients between users [2].

The second group G2 will include data on which it is impossible to uniquely determine the value of the coefficient. For example, unknown time limits between events that allow you to uniquely identify whether a user is a fraudsters or an organic user. One of the coefficients will be determined based on the types of events that the user makes. This factor can not be included in the first group G1, since such data does not have a clearly defined condition that will not change over time. There is a similarity coefficient between the current user and the fraudsters fingerprints and characteristics to normalize the data of the second group (Figure 3). Note, that the value of the similarity coefficients belongs to the interval [0; 1]. A value of 1 means that users are identical by a given characteristic, 0 in turn means that users do not have anything in common by current characteristic. For example, after considering such data as the time between events, we have a set of times between the events of the current user$Tu = \{t \mid t >= 0\}$and the set of times between the events of each of the uniquely determined fraudsters$Ti = \{t \mid t > 0\}$. With two sets of non-binary, but homogeneous TuandTi,data, let's apply the appropriate user similarity coefficient. For such kinds of sets, the Tanimoto coefficientKT Tu , Ti [2, 4, 16], is chosen, which is defined as$KT(Tu, Ti) = \frac{N_C}{N_a + N_b - N_c}$, where N_C – is the number of common Tu and Ti elements, Na – is the number of elements in the set Tu, Nb – is the number of elements in theTi set. In turn, for sets of binary data, such as a set with binary values for each of the existing types of events in a mobile application, where 0 means that the user did not use such an event, and 1 means that he used, the similarity coefficients of users are determined using and Ti elements, Na – is the number of elements in the set Tu, Nb – is the number of elements in theTi set. In turn, for sets of binary data, such as a set with binary values for each of the existing types of events in a mobile application, where 0 means that the user did not use such an event, and 1 means that he used, the similarity coefficients of users are determined using the cosine similarity coefficientKcos A1, A2 [2, 4, 16], which most effectively works with binary data. In turn, $K_{cos}\ A_1, A_2 = \cos\ A_1, A_2 = \frac{A_1 \cdot A_2}{A_1 \cdot A_2}$, vhereA1, A2 – is the set with the above binary data of the current user and fraudulent user respectively.

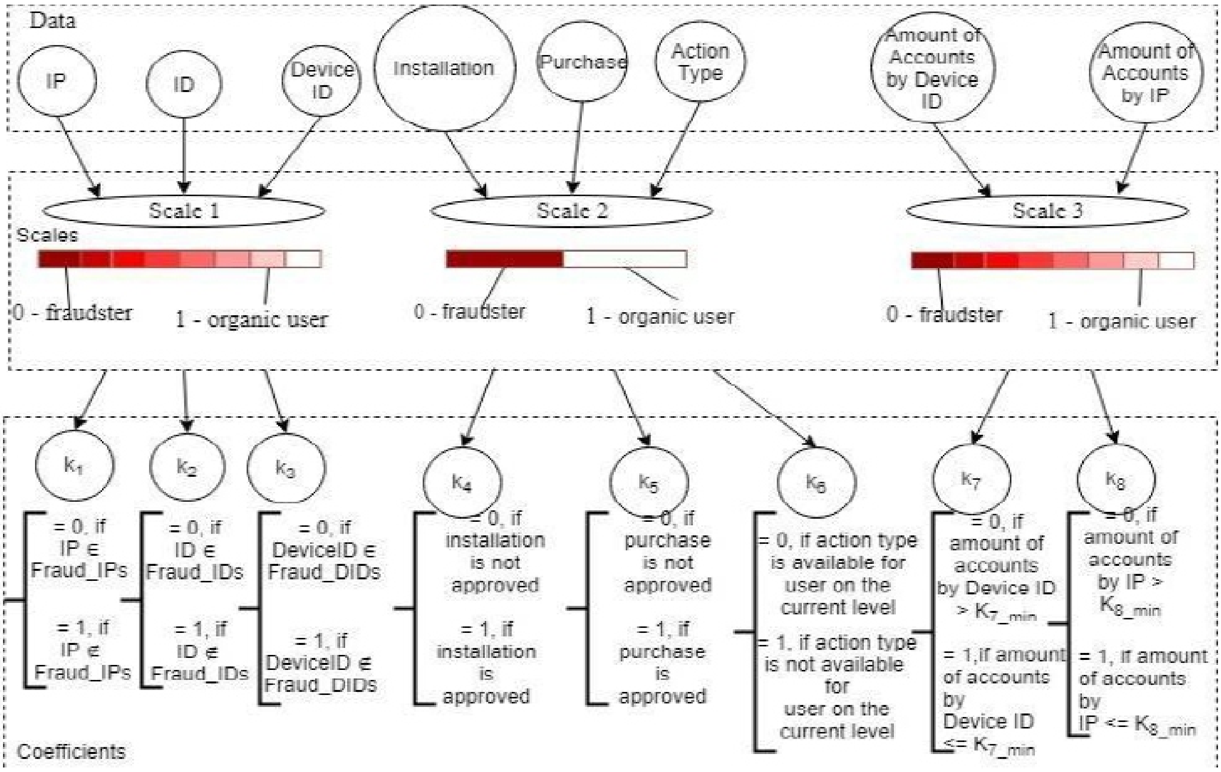

Data
IP
ID
Device ID
Installation
Purchase
Action Type
Amount of Accounts by Device ID
Amount of Accounts by IP
Scale 1
Scale 2
Scale 3
Scales
0 - fraudster
1 - organic user
0 - fraudster
1 - organic user
0 - fraudster
1 - organic user
k1
k2
k3
k4
k5
k6
k7
k8
= 0, if IP ∈ Fraud_IPs
= 1, if IP ∉ Fraud_IPs
= 0, if ID ∈ Fraud_IDs
= 1, if ID ∉ Fraud_IDs
= 0, if DeviceID ∈ Fraud_DIDs
= 1, if DeviceID ∉ Fraud_DIDs
= 0, if installation is not approved
= 1, if installation is approved
= 0, if purchase is not approved
= 1, if purchase is approved
= 0, if action type is available for user on the current level
= 1, if action type is not available for user on the current level
= 0, if amount of accounts by Device ID > K7_min
= 1,if amount of accounts by Device ID <= K7_min
= 0, if amount of accounts by IP > K8_min
= 1, if amount of accounts by IP <= K8_min
Coefficients

Fig.2. Scales of the first group of coefficients.

Then, fraudsters characteristics and patterns are formed and recorded in the knowledge base according to the fraudsters characteristics, detected based on the coefficients of the second group. A generalized fraudster template (portrait) is formed, based on the formed knowledge base and database. This template allows to identify fraudsters, as the experimental studies research showed, with an accuracy of 99.14%. It should be noted that the availability of a knowledge base will help speed up the detection of fraudsters in new data sets.

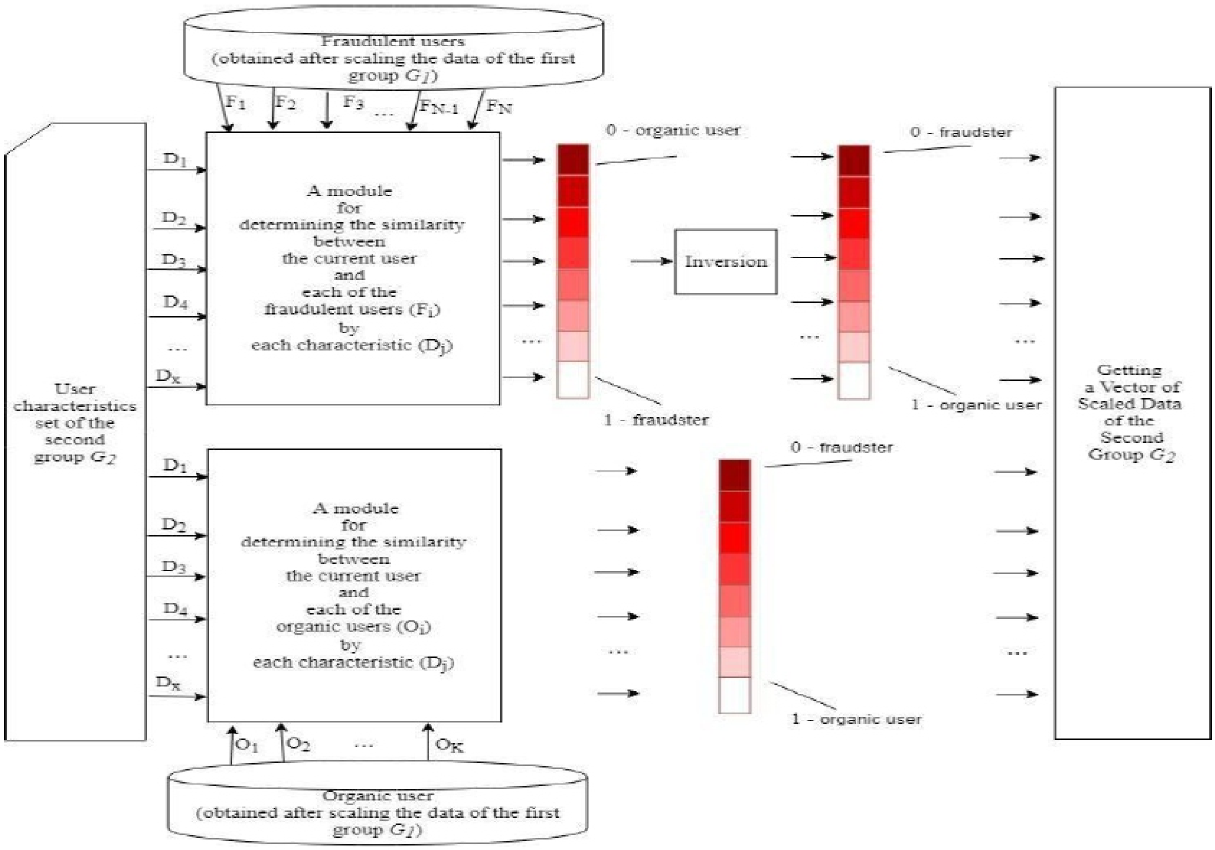

Fraudulent users
(obtained after scaling the data of the first group G_1)
F_1
F_2
F_3
...
F_{N-1}
F_N
User characteristics set of the second group G_2
D_1
D_2
D_3
D_4
...
D_X
A module for determining the similarity between the current user and each of the fraudulent users (F_i) by each characteristic (D_j)
0 - organic user
1 - fraudster
Inversion
0 - fraudster
1 - organic user
A module for determining the similarity between the current user and each of the organic users (O_i) by each characteristic (D_j)
0 - fraudster
1 - organic user
Getting a Vector of Scaled Data of the Second Group G_2
O_1
O_2
...
O_K
Organic user
(obtained after scaling the data of the first group G_1)

Fig.3. Scales of the second group of coefficients.

III. DETAILS EXPERIMENTAL

Let us build the system which would give the adequate results. It should be noted that an adequate system must be endowed with two properties: it is endowed with predictive ability and well-generalized for data that it has not yet met. To assess these properties, the error metric (how much the model is seriously mistaken) and the strategy for validating the adequacy are determined. Consequently, let's implement the proposed intelligent system for detecting mobile app install fraud, based on the developed method, mathematical model and algorithms, using the Python programming language. However, to begin with, let us note that to obtain adequate results, one can not test and teach the model on the same dataset. Therefore, to adequately evaluate the system and avoid retraining, we divide the available dataset of data into three sets: training, validation, and test (control). The system will be trained on training data and will be evaluated on verifying, and after creating the final version of the system, it will be tested on the control data. So, after obtaining the intelligent system with the adequate results, let's construct a confusion matrix (Fig. 4) to compare the predicted results with the expected results. To test the system, a control sample of 56962 records was taken, 56870 of which correspond to organic users, and 92 – to fraudsters.

The confusion matrix which is given in Fig.4shows that 56639 objects of the first class ("Organic"), representing 99.59% of all 56870 objects classified in the first class. 231 objects of the first class are mistakenly classified as objects of the second class ("Fraud"). 99.59% of the firstclass objects are correctly classified and 0.41% are classified incorrectly. As for objects of the second class – fraudsters – 74 objects from 92, classified into the second class, are classified correctly, and 18 – mistakenly attributed to the first class. Thus, 80.43% of fraudsters were correctly classified and 19.57% were classified incorrectly. In general, 99.56% of objects (56639 + 74 = 56713 objects out of 56962 objects) are correctly considered, 0.44% of objects are incorrectly classified.

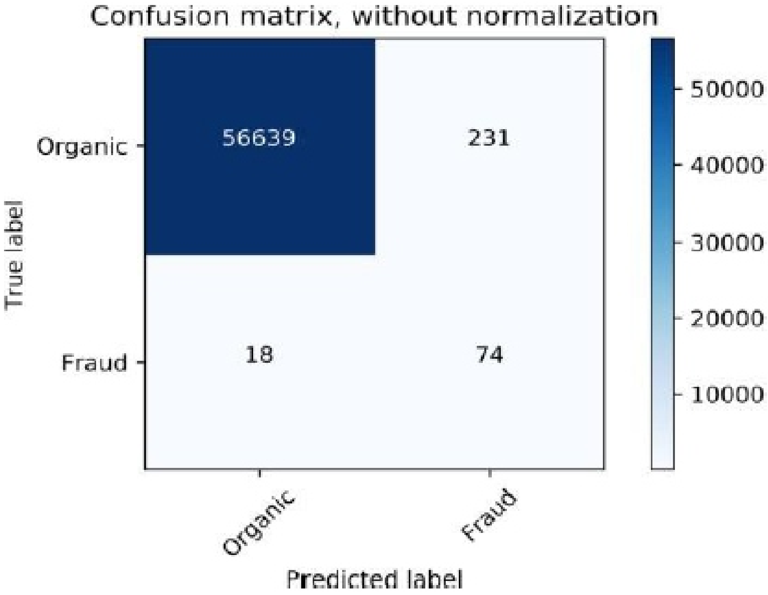


Fig.4. Confusion matrix of the control set without normalization

Let us estimate the sensitivity (1) also called the true positive rate (TPR) or recall, which measures the proportion of actual positives that are correctly identified as such, and specificity (2) of the system also called as true negative rate (TNR), which measures the proportion of actual negatives that are correctly identified as such.

$$TPR = \frac{TP}{P} = \frac{TP}{TP+FN} = \frac{56639}{56870} = 99\%, \ (1)$$

where TP – true positive,
TN – true negative,
FP – false positive,
FN – false negative.

$$TNR = \frac{TN}{N} = \frac{TN}{TN+FP} = \frac{74}{92} = 80\%. \ (2)$$

The values of TPR = 99% and TNR = 80% are high, and it means that the developed intelligent system for detecting mobile app install fraud has adequate results.

IV. RESULTS AND DISCUSSION

Thereby, the approach to the development of an intelligent system for detecting mobile app install fraud was proposed in the paper. The advantage of proposed system is in ability to work with all existing arrays of heterogeneous input data when installing mobile applications with application of algorithms of data scaling (Fig.2, 3).

The results of the experiments have shown effectiveness of proposed approach for fraud detection and the possibility of expanding formats and characteristics of fraudulent users based on intellectual analysis and knowledge bases. The study was carried out using software developed on the basis of the system proposed in this paper with the help of the Python programming language and TensorFlow, Pandas and Numpy libraries in the PyCharm development environment. To implement the classification block, a fully-connected deep neural network with 3 hidden layers was used.

Using of the proposed approach to data heterogeneity resolution helps to find fraudsters with high accuracy, which is achieved thanks to the fact that the system makes it possible to determine fraudsters even by implicit fingerprints and characteristics.

V. CONCLUSIONS

The approach to the development of an intelligent system for detecting mobile app install fraud was proposed in this paper. Let's notice major conclusions are as follows:

1. The advantage of the proposed system is that it works with all available heterogeneous input data of different templates, dimensions, metrics, which allows defining a user class more precisely.
2. The structure of the system and the list of the main blocks of the system were noted (Fig. 1)
3. Software was developed based on the proposed methods and models during the mobile app install fraud detection intelligent system design. A sample of 300567 records from a real mobile application is taken for experimental research. All data of a sample are heterogeneous and each

user of it is labeled with the class (fraudulent or organic). The system developed by the authors does not know about the user's labels. The labeled set was chosen for further verification of the accuracy of the developed system. The fraudsters characteristics and the formed rules in the knowledge base were more precisely defined due to the developed method of data heterogeneity resolution. Python programming language and TensorFlow library are selected for implementation of the system.

4. The experiments were done using confusion matrix. The results showed that the values of sensitivity (99%) and specificity (80%) of the development system are high. It means that the developed intelligent system for detecting mobile app install fraud has adequate results.

ACKNOWLEDGMENTS

I would like to show my gratitude to the professors of Vinnytsia National Technical University for sharing their pearls of wisdom with me during the course of this research, and I thank 2 reviewers for their so- called insights. I am also immensely grateful to YarovyiAndrii, Tsmots Ivan, and Timchenko Leonid for their comments on an earlier version of the paper, although any errors are my own and should not tarnish the reputations of these esteemed persons.

REFERENCES

[1] S. Benndorf, G. Kakulapati, A. Pham and others, "Fighting Mobile Fraud in the Programmatic era: AppLift". AppLift GmbH, 14 p., 2015.

[2] Yarovyi A. A., Romanyuk O. N., Arsenyuk I. R., Polhul T. D., "Program applications install fraud detection using data mining", NaukovipratsiDonetskohonatsionalnohotekhnichnohounivers ytetu. Seriya: "Informatyka, kibernetyka ta obchysliuvalnatekhnika". 2017. Issue 2 (25). P. 126–131, available at: http://science.donntu.edu.ua/wp- content/uploads/2018/03/ikvt_2017_2_site-1.pdf

[3] "Our take on mobile fraud detection",available at: http://geeks.jampp.com/data-science/mobile-fraud/

[4] Vacha Dave, Saikat Guha, Yin Zhang, "ViceROI: Catching Click-Spam in Search Ad Networks",available at: http://www.sysnet.ucsd.edu/~vacha/ccs13.pdf

[5] Dave, V., Guha, S., Zhang Y., "Measuring and Fingerprinting Click-Spam in Ad Networks. In Proceedings of the Annual Conference of the ACM Special Interest Group on Data Communication (SIGCOMM)". Helsinki, Finland, vol. 175, 186 pp., Aug. 2012.

[6] "Fraudlogix: Ad Fraud Solutions for Exchanges, Networks, SSPs & DSPs", available at: https://www.fraudlogix.com/

[7] "Kraken Antibot", available at: http://kraken.run/

[8] "Adjust", available at: https://www.adjust.com/

[9] "Kochava Uncovers Global Ad Fraud Scam",available at: https://www.kochava.com/

[10] "TMC Attribution Analytics", available at: https://help.tune.com/marketing-console/attribution-analytics/

[11] "Appsflyer: Protect your data from mobile fraud: Protect360",available at: https://www.appsflyer.com/product/protect360/

[12] "FraudScore: FraudScore fights ad fraud using Machine Learning", available at: https://fraudscore.mobi/

[13] "AppMetrica", available at: https://appmetrica.yandex.ru/

[14] T. Polhul, A. Yarovyy, L. Krylyk, "Developing of the method of mobile application installations fraud detection using data mining" in Scientific Works of Conference "XLVII Scientific and technical conference of subdivisions of Vinnytsia National Technical University (2018)", Vinnytsia, 2018. [Electronic resource],available at::http://ir.lib.vntu.edu.ua/bitstream/handle/123456789/2272 2/079.pdf?sequence=1 (Ukr.).

[15] Polhul, T., Yarovyi, A., "Development of a method for fraud detection in heterogeneous data during installation of mobile applications", Eastern-European Journal of Enterprise Technologies, №1/2 (97), 2019. – doi: 10.15587/1729- 4061.2019.155060

[16] Kiulian A. H., Polhul T. D., Khazin M. B. "Matematychna model rekomendatsiynohoservisunaosnovi metodukolaboratyvnoifilt ratsiyi",Kompiuternitekhnolohiyi ta Internet v informatsiynomususpilstvi. 2012. p. 226–227, available at:: http://ir.lib.vntu.edu.ua/bitstream/handle/123456789/7911/22 6-227.pdf?sequence=1&isAllowed=y

A Data Mining Infrastructure for Cheminformatics

Jungkeekim
Assoc. Prof., Sungkonghoe University, 320, Yeondong-ro, Guro-gu, Seoul,Korea
E-mail: jake@skhu.ac.kr

ABSTRACT

An enormous increase of data sources for chemical information and biological science requires a new development methodology for mining useful information. Such data sources give us an opportunity to utilize computational tools to mine useful information and to find new patterns in data sets that explain scientific phenomena not yet known. It is also important that non-expert users can access the latest cheminformatics methodology and models to spread the new discoveries. We present our previous developments in cheminformatics procedures and infrastructure that provide an appropriate approach to mining large chemical datasets. We also discuss the limitation of previous challenge and propose a new infrastructure with the state-of-the-art techniques expected to improve the performance.

Keywords - Cheminformatics, Work flow, Web service, Big Data.

I. INTRODUCTION

Recent progress in chemistry and life sciences have led to a large volume of new data sources called Big Data. Big Data is a data set that cannot be proficiently handled by conventional data processing technics. When we classify big data, different principles should be considered in data mining [1]. Cheminformatics belongs to a multidisciplinary filed that integrateslife science, chemistry and computer science. The in silico illustration of chemical structures employs particular formats such as XML-based Chemical Markup Lange (CML), SMILES, SDF, and so on. The data in those formats are frequently used in large chemical databases such as PubChem[2], ChEMBL[3], and BindingDB[4]. Therefore, we can access a large volume of chemical compounds and biological activities in a diversity of biological assays.

We need to connect chemical structures to the life science information. For example, systems biologists study the complex biological systems that integrate microarray datasets to biological pathways, using a large number of other data sets to provide evidence for the links [5].

A typical method to access data is a traditional query to the database manage systems by a human. A software agent can access and process the data in a uniform manner without human intervention. Web services are techniques of aggregating and integrating data sources and software. They allocate software applications and data source to be published on the network, therefore making tools and data broadly available with a standardized interface and enabling the construction of application that use distributed resources and data to resolve complex tasks. There are three standards to create Web services. Web Service Description Language (WSDL) is an XML- based standard for presenting Web services and their parameters. Simple Object Access Protocol (SOAP) provides the envelope existing applications to match abstract interfaces in WSDL to their actual executions. Universal Description, Discovery and Integration (UDDI) provokes the publishing and browsing of Web services by user groups. Representational State Transfer (REST) architectural style replaces WSDL since the REST-based design methodology [6] was emerged. In RESTful style, there is no standard such as SOAP and any other payload formatted in HTML, XML, JavaScript Object Notation (JSON), or another format. The aspect of connections between distributed resources is important because it is easy to collect information from a diversity of high throughput screening and vendor catalogues.

The MapReduce framework [7] provides a programming model for parallel and distributed handling of batch jobs on a large number ofcomputing nodes. Each job in the MapReduce divided into two phases –map and reduce. The map phase divides the input data by relating each element with a key. The reduce phase handles each split independently, and all data is processed based on key-value pairs. The map function processes a certain key-value pair and produces a certain number of new key-value pairs. The reduce processes all intermediate values grouped by the same key into another set of key value pairs as output.

A scientific workflow is a specialized form of the general workflow, which designed particularly to compose and implement a set of tasks in an order depending on their relations in a scientific application [8]. The technic of scientific workflow has been successfully applied to the scientific fields including cheminformatics and life science. Scientific Workflow Management System (SWfMS) is a tool to implement workflows and handle data sets. Several Grids workflow projects are developed. Triana [9], Kepler [10], and Taverna [11] are typical examples. Triana is started from a single platform but supports distributed services with Grid awareness. Kepler is also started from a single platform and it fully supports Grid environment. It is widely used in many scientific domains and provides graphical user interface to organize workflows intuitively. Taverna is part of myGrid project and focuses on applications of life science.It recognizes the importance of provenance and semanticsby a textual language.

A workflow scheduler is critical for the efficient workflow management system. Many scientific workflow management systems hire their own scheduling algorithms [12, 13]. We need to find a proper algorithm for a good performance.

The rest of this paper is organized as follows. Section 2 describes our previous work. Section 3 presents MapReduce framework and scheduler. We illustrate a new architecture for data mining of large data sets in Section 4. We summarize and conclude in the last Section.

II. PREVIOUS WORK

We developed an infrastructure of Web service for cheminformatics that simplifies the access to drug discovery information and the computational techniques that can be applied to it [14]. At that time, the Web services were mostly based on Java. Using Java allows us to deploy our Web services in a Tomcat application module, which allows us to easily support a variety of services and provide an integration with our execution environments. The services themselves are hosted on a diversity of servers and are generally separated from database and functionality. Therefore Web services that provide database functionality will connect to a remote database server to retrieve results. We implemented Web service wrappers for several free and commercial cheminformatics tools. The commercial tools that we were permitted to use tools such as Digital Chemistry Divisive K-Means forclustering. We had a working relationship with the Murray-Rust group at Cambridge University[15] that was one of sites that had semantic Web approaches to cheminformatics. We implemented several of their Web services such as InChIGoogle, InChIServer, CMLRSSServer, and OSCAR for automatic mining of chemical structure information from documents. We also implemented a large amount of the functionality of the Chemistry Development Kit (CDK) as Web services such as descriptor calculation, 2D similarity and fingerprint calculations, and 2D structure depiction. We experimented a special modified Web service implementation of ToxTree [16] for toxicity flagging. Web services can be used in workflow tools such as Taverna workbench. The tools allow the creation of new functionality by linking together services and other application and data resource into workflows. Figure 1 illustrates an example of Taverna workbench in a graphical user interface. The interface encloses a list of processes that the user enables invoking on that service. After selecting an operation, the user isaccessible with an interface for the operation, which enables the user to specify all the input parameters to the operation. And the user can invoke the operation on the service and obtain the output results.

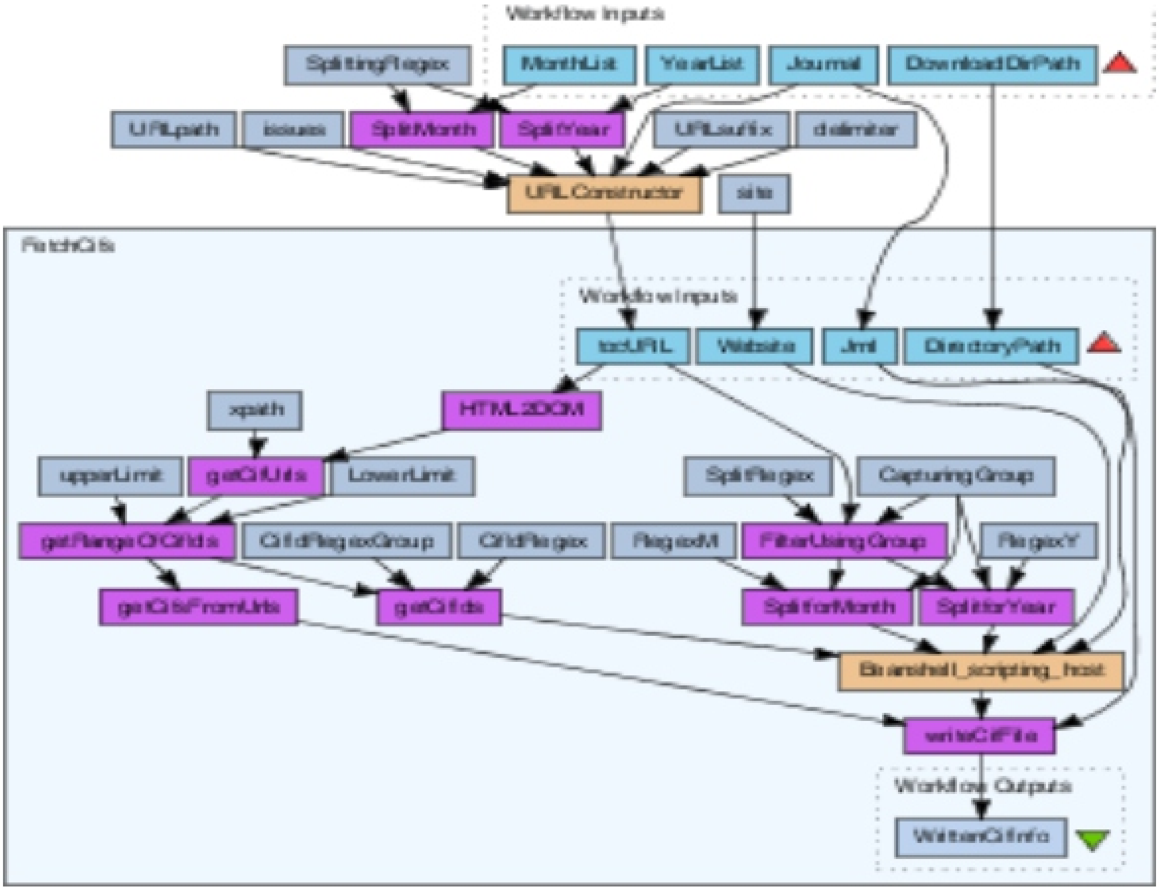


Figure 1: CDK Workflow in Taverna workbench.

III. MAPREDUCE FRAMEWORK AND SCHEDULER

MapReduce frameworks execute much better in tough environments than other tightly coupled distributed programming frameworks such as Message Passing Interface (MPI) because of their fault tolerance capabilities [17].They are suitable to support many scientific use cases, and many scientists can build large data-oriented applications easily under could computing environment.

Apache Hadoop [18] is a framework that provides distributed processing of large data sets and the implementation is based on Google MapReduce [19]. The Hadoop Distributed File System (HDFS) follows write-one-read-many pattern and does not support functions to change an existing file. The HDFS is designed for deployment on unreliable clusters and succeeds in reliability by the replication of data files. The Hadoop minimizes data communication by processing computations near the place it is stored. The architecture consists of a master node with many worker nodes and uses a queue for task scheduling and succeeds in load balance naturally among computing tasks.

A classical workflow for collecting related data and inserted into a local database management system (DBMS) before processing data. The HDFS replaces the local database for a temporary storage. Apache Hadoop framework is a promising system to store the extracted huge datasets from databases.

A scheduler of scientific workflows allocate tasks mapping on heterogeneous and distributed resources. A good algorithm can make tasks allocated to the proper resources and arrange the best sequence of parallel tasks. We need to consider two groups – users and service providers. Users are concerned with reliability and the service quality. So they want the result within the proper time. However, the servers aim at their efficiency to capture maximum revenues. We can consider several strategies such as execution time-based strategy, just-in-time strategy, linear scheduling, policy base strategy, virtual machine strategy, gossip based strategy, reservation based strategy, and heuristic based strategy. We need to optimize our scheduler among those strategies in the future work.

IV. ARCHITECTURE FOR DATA MINING OF LARGE CHEMICAL DATASETS

In our previous work [14, 20], we introduced a chemical mining process to collect chemical structures. Figure 2 presents the architecture of the process implemented on a supercomputer with Message Passing Interface (MPI). Using PHP script queries and PubMed ID, we collect abstracts of research papers in the first step. A group of the abstract text files are assigned to each node in a supercomputer. In a node, a series of batch processes extracts chemical compounds and their three dimensional structures.

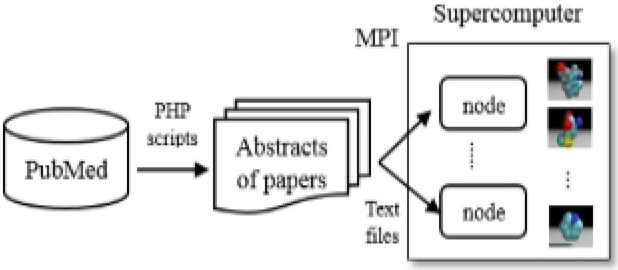


Figure 2: Architecture of a chemical compound mining process

In the experiment, even the super computer system took a lot of time to process about 500,000 abstracts. We suggest an architecture in which the Hadoop MapReduce Framework replaces the super computer system with a simple MPI. Figure 3 illustrates a new architecture replacing the super computer system in the Figure 2. The input text files are stored in the database (HDFS). The server provides graphical user interface as a part of workflow bench such as Tarverna. Workers are instances of the server and are only accessed by a scheduler that assigns execution tasks for mapping or reducing.We can employ a SOAP library to allow consumption of Web services.

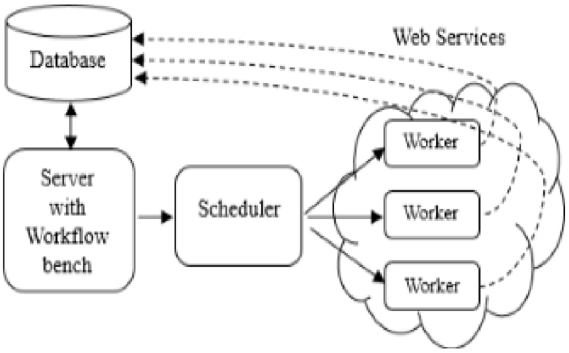


Figure 3: New Architecture for Mining Process

V. CONCLUSION

With recent progress in chemistry and life science generating a large datasets forces many requirements on a storage and an analysis framework. We describe a review of distributed systems designed to process chemical information. We also present our Web service and workflow development for cheminformatics in the previous work. However, the case study for mining chemical compound demonstrates a need for more efficient architecture for processing large datasets in chemistry and life science field. Thus we propose a new architecture with MapReduce framework to expect to address the performance problem.

REFERENCES

[1] Y. Hu and J. Bajorath. "Entering the 'big data' era in medicinal chemistry: molecular promiscuity analysis revisited," Future science OA, vol.3, 2017.

[2] E. Bolton, et al., "PubChem: integrated platform of small molecules and biological activities," Annual reports in computational chemistry, Vol. 4. Elsevier, pp.217-241, 2008.

[3] T. Liu, et al.. "BindingDB: a web-accessible database of experimentally determined protein–ligand binding affinities," Nucleic acids research, vol.35, pp. 198-201, 2006.

[4] A. Gaulton, et al., "ChEMBL: a large-scale bioactivity database for drug discovery," Nucleic acids research, vol. 40, D1100-D1107, 2011.

[5] Y. Saeys, I. Inza, and P. Larranaga. "A review of feature selection techniques in bioinformatics," bioinformat ics vol. 23, pp.2507-2517, 2007.

[6] R. Fielding and R. Taylor, "Principled design of the modern Web architecture,"ACM Transactions on Internet Technology (TOIT), vol. 2, pp. 115-150, 2002.

[7] J. Dean and S. Ghemawat. "MapReduce: simplified data processing on large clusters," Communications of the ACM vol. 51.1, pp. 107-113, 2008.

[8] J. Liu et al. "A survey of data-intensive scientific workflow management," Journal of Grid Computing vol. 13.4, pp.457- 493, 2015.

[9] I. Taylor, M. Shields, I. Wang, and A. Harrison, "The Triana Workflow Environment: Architecture and Applications," Workflows for e-Science, Springer, pp. 320-339,2007.

[10] D. Pennington, D. Higgins, A. Peterson, M. Jones, B. Ludascher, S. Bowers, "Ecological Niche Modeling Using the Kepler Workflow System," Workflows for e-Science, Springer, pp. 91-108, 2007.

[11] T. Oinn, P. Li, D. Kell, C. Goble, A. Goderis, M. Greenwood, D. Hull, R. Stevens, D. Turi, and J. Zhao, "Taverna / myGrid: aligning a workflow system with the life sciences community," Workflows for e-Science, Springer, pp. 300-319, 2007.

[12] G. Gharooni-farda, F. Moein-darbari, H. Deldari, and A. Morvaridi, Scheduling of scientific workflows using a chaos- genetic algorithm. Procedia Computer Science, vol. 1, pp144501454, 2010.

[13] M. A. Rodriguez and R. Buyya, "Deadline based resource provisioningand scheduling algorithm for scientific workflows on clouds," IEEE transactions on Cloud Computing, vol. 2, pp. 222-235, 2014.

[14] X. Dong, et al. "Web service infrastructure for chemoinformatics." Journal of chemical information and modeling, vol. 47, pp. 1303-1307, 2007.

[15] Murray-Rust Research Group, http://www-pmr.ch.cam.ac.uk.

[16] ToxTree,http://sourceforge.net/projects/toxtree.

[17] T. Gunarathne, et al. "Cloud computing paradigms for pleasingly parallel biomedical applications," Proc. of ACM Int. Symp. on HPDC, ACM, pp. 460-469, 2010.

[18] Apache Hadoop, http://hadoop.apache.org/.

[19] J. Dean, and S. Ghemawat, "MapReduce: simplified data processing on large clusters," Communications of the ACM, vol. 51, pp. 107-113, 2008.

[20] J. Kim, "Chemical Compound Mining for Big Data,"Proc. of Intl. Conf. on Future Generation Information Tech., 2019.

Implementation of Artificial Intelligence to Detect Images in the Manufacturing Process

[1] Marek Kocisko, [2] Juraj Kundrik, [3] Martin Pollak, [4] Monika Teliskova

[1,2,3,4]Technical university of Košice, Faculty of manufacturing technologies with a seat in Prešov Slovakia

E-mail: [1]marek.kocisko@tuke.sk, [2]juraj.kundrik@tuke.sk, [3]martin.pollak@tuke.sk, [4]monika.teliskova@tuke.sk

ABSTRACT

Automation of the production of ever more complex products brings along an increase in product control requirements. In addition, industry, especially automotive, requires a further increase in production accuracy, minimization of downtime and delivery by just-in-time. Quality control by accidental selection and by mechanical devices is not sufficient to meet the production requirements. In most cases, visual quality control is ensured by people. This article describes the basic paradigms of using artificial intelligence systems for quality control detection. To describe the deployment techniques of artificial intelligence elements, the RetinaNet Convoluntary Neural Network (CNN) was used.

Keywords - Neural Networks, Quality Control, Error Detection.

I. INTRODUCTION

At present, the world is talking about the fourth industrial revolution, which brings many attractive opportunities for industrial companies. The development of automation, robotics and digital technologies is determined by the global development of the economy and the individual industries, as well as the emerging technological and social trends.

Traditional production hierarchy with centralized management is increasingly moving towards decentralized automatic control, where the resulting product independently communicates with production facilities and actively interferes with the production process.

An integral part of automated active production process management is to automate and correctly detect possible errors occurring in the production process. At present, these errors, such as mechanical damage, mistaken production or scratches, are evaluated only manually, based on acquired human abilities. To automate this activity, you need to correctly detect and evaluate these errors. The solution to this problem can be the use of artificial intelligence techniques.

Artificial intelligence techniques are therefore generally used for problems that involve a certain amount of uncertainty. Therefore, these techniques may be suitable for controlling machining.

Artificial Intelligence, which deals with decision-making and solving complex issues in its sub-areas, looks very promising for the description and management of complex systems in the future.

The basic elements of artificial intelligence are: expert systems, fuzzy logic, genetic algorithms, neural networks, neuro-fuzzy systems, hybrid systems. However, several intelligent computing tools have been applied to solve manufacturing problems. In the literature, it is possible to find about three dozens of methods, such as artificial life, associational memory, automation theory, biological production systems, chaos theory, computer vision, conceptual dependence charts, restriction based search, expert systems, fuzzy logic, genetic algorithms, , heuristic search, holon production systems, immune networks, knowledge base systems, knowledge representation, machine learning, multiagent systems, natural language processing, neural networks, Petri nets, quantitative justification, justification techniques, learning consolidation, similarity theory, syntactic and statistical recognition samples and etc.

For image recognition, it is best to apply neural networks that are rapidly expanding area of artificial intelligence.

With advancements in this area, predefined image recognition solutions are available. However, most existing solutions focus on recognizing common subjects. The introduction of neural networks in industrial production is still unexplored area. Pre-trained convolutional neural networks (CNNs) are available to transfer abstraction capabilities to the digital environment.

The advantages of this solution are:

- automation of control,
- linking control and production process,
- improving control accuracy,
- categorization of failures, interconnection of the type of failure with the manufacturing operation/procedure.Implementation of the RetinaNet neural network to detect image recognition errors and incorporate it into the product quality control process was carried out in collaboration with Optisolutions Ltd.

II. POSSIBILITIES OF IMAGE DETECTION BY IMPLEMENTING ARTIFICIAL INTELLIGENCE TECHNIQUES

The neural network is a massive parallel processor that tends to retain knowledge for their further use. It imitates the human brain in two aspects:

- Knowledge is collected during learning,
- Neural connections (weights) are used to store knowledge. [1]

Among the basic types of tasks that can be solved using neural networks can also be classified into classes. For visual data, this type of task is called image detection. Image detection in neural networks uses convolutional and pooling layers, so these neural networks are also called convolutional neural networks (CNNs). Recently, there are predefined implementations of convolutional neural networks that can be used to look for visual defects by detecting images in an industrial environment.

For use in the experiment, the Focal Loss for Dense Object Detection model was selected for very accurate image recognition in objects.

„Focal Loss for Dense Object Detection is the highest accuracy object detectors to date are based on a two-stage approach popularized by R-CNN, where a classifier is applied to a sparse set of candidate object locations. In contrast, one-stage detectors that are applied over a regular, dense sampling of possible object locations have the potential to be faster and simpler, but have trailed the accuracy of two-stage detectors thus far.“ [2]

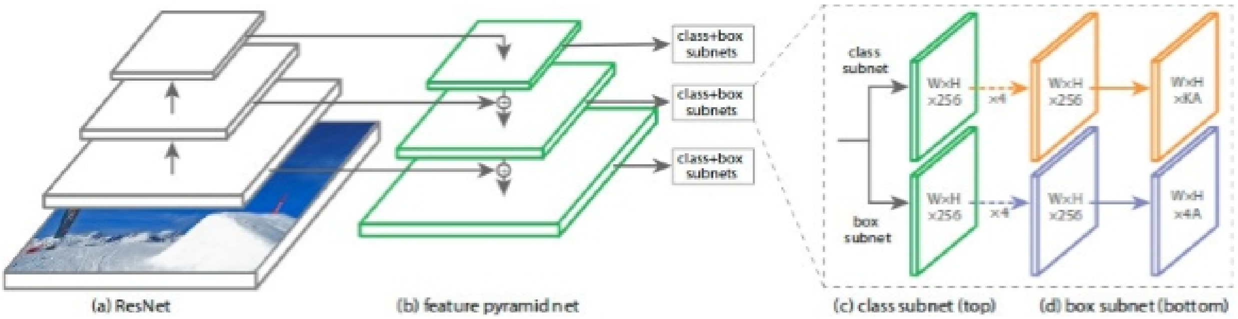
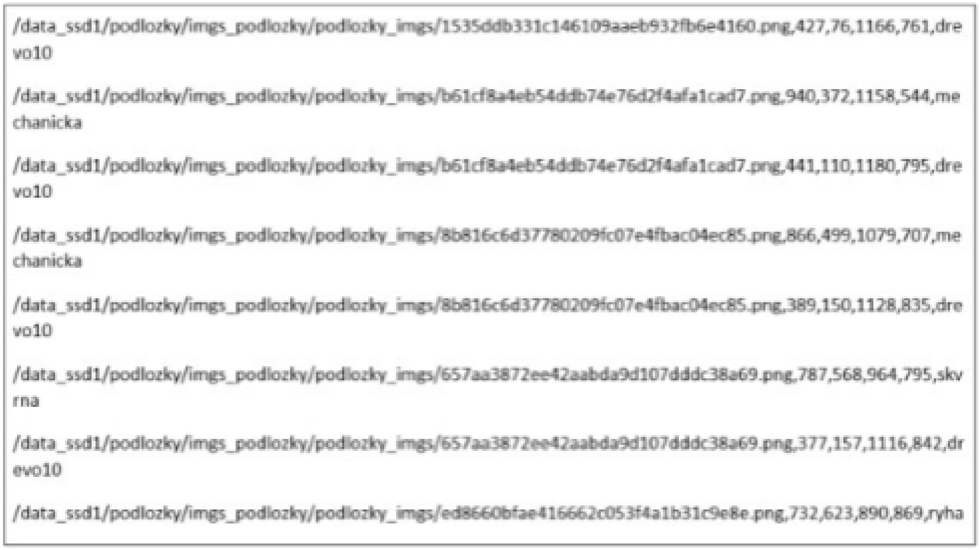


Fig. 1 Structure of the convolutional neural network RetinaNet

The predefined RetinaNet neural network is a software implementation in Keras and Tensoflow environment. The package includes a predefined and pre-trained neural network model that can be used as the basis for a specific task. Selecting this model allows you to deal only with knowledge transfer problems and the interpretation of neural network recognition results in an industrial environment. [3], [4]

III. IMPLEMENTING THE NEURON NETWORK IMAGE DETECTION IN INDUSTRIAL PRACTICE

The first step is to select and implement neural network learning. Selection of the neural network was performed on the basis of output requirements and the RetinaNet network was selected. The input source for learning neural network RetinaNet is an image of the investigated product. The output is a matrix describing their images and their location on the detected image. Thus described images represent the knowledge that seeds to be transferred to the neural network. For the learning of the neural network, approximately 320 images of the investigated products were used. Various types of washers represented different product types. On some of the washers (about 20 pieces) was simulated damage, that was artificially created in the form of scratches, stains and mechanical damage.



```
/data_ssd1/podlozky/imgs_podlozky/podlozky_imgs/1535ddb331c146109aaeb932fb6e4160.png,427,76,1166,761,dre
vo10

/data_ssd1/podlozky/imgs_podlozky/podlozky_imgs/b61cf8a4eb54ddb74e76d2f4afa1cad7.png,940,372,1158,544,me
chanicka

/data_ssd1/podlozky/imgs_podlozky/podlozky_imgs/b61cf8a4eb54ddb74e76d2f4afa1cad7.png,441,110,1180,795,dre
vo10

/data_ssd1/podlozky/imgs_podlozky/podlozky_imgs/8b816c6d37780209fc07e4fbac04ec85.png,866,499,1079,707,me
chanicka

/data_ssd1/podlozky/imgs_podlozky/podlozky_imgs/8b816c6d37780209fc07e4fbac04ec85.png,389,150,1128,835,dre
vo10

/data_ssd1/podlozky/imgs_podlozky/podlozky_imgs/657aa3872ee42aabda9d107dddc38a69.png,787,568,964,795,skv
rna

/data_ssd1/podlozky/imgs_podlozky/podlozky_imgs/657aa3872ee42aabda9d107dddc38a69.png,377,157,1116,842,dr
evo10

/data_ssd1/podlozky/imgs_podlozky/podlozky_imgs/ed8660bfae416662c053f4a1b31c9e8e.png,732,623,890,869,ryha
```

Fig 2. File describing knowledge about errors and definition of washers.

In the set of created images were six types of washers and three types of visual damage. Each type of washer and each type of damage creates a class that categorizes that knowledge. For each image a description was created that defines the location and class of knowledge which the image contains. Example of record is in Fig. 2.

In order to obtain the error type location, an application has been developed to identify and record their location on the investigated image. An example of the visual identification of the error is shown in Fig. 3. Using this procedure, we will get a set of images and a set of descriptions that make up the necessary knowledge for learning the neural network itself. The RetinaNet neural network software package also includes utility programs that provide learning, training and also evaluation of the neural network.

After training the neural network followed its testing process. For the test, the washers that were used to teach it and their other types were used. Again some of the washers were visually damaged.

The resulting image recognition using the RetinaNet neural network is a triple matrix that describes the class of product classes, namely:

- location, exact coordinates of occurrence of a given category of errors on the image being examined,
- the probability of detection of individual product categories,
- designation of the product category.

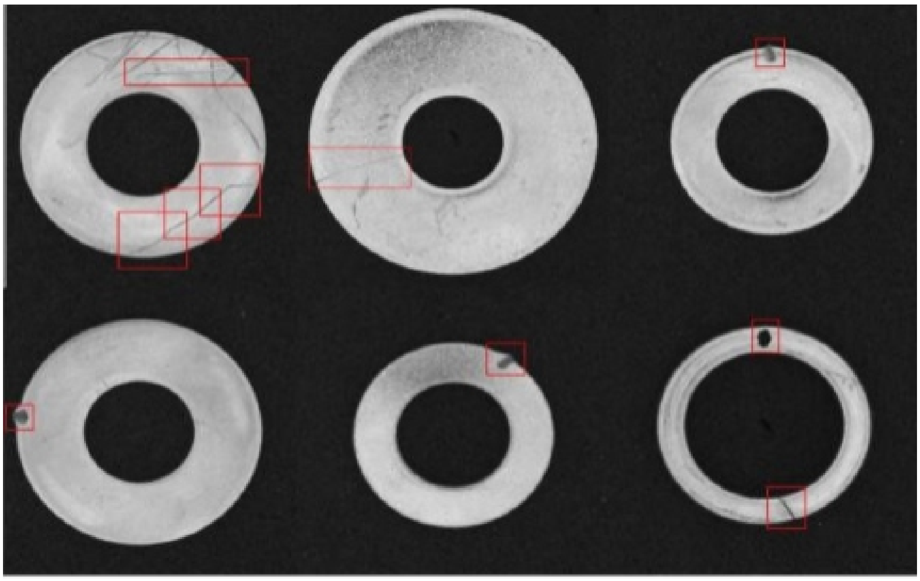


Fig. 3 An example of products with individual errors designed to train a neural network

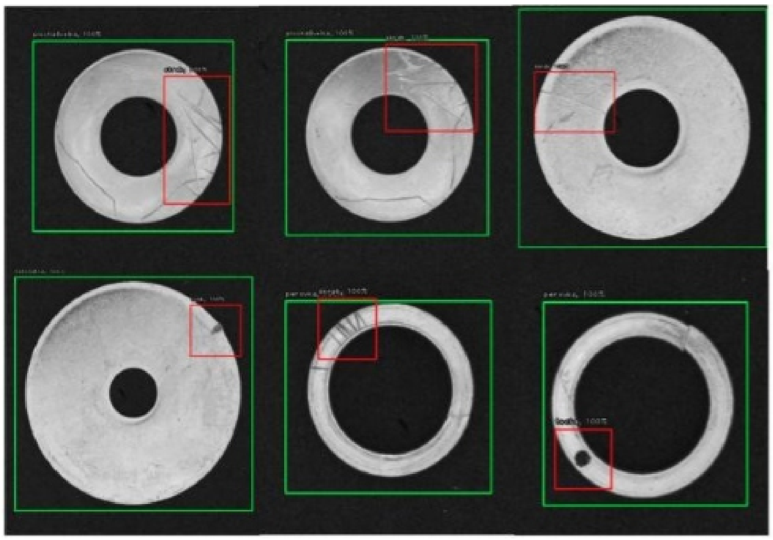


Fig. 4 Examples of error handling using a neural network

These matrices do not allow visual inspection of neural network functionality. To verify the functionality was used a Python script which visualizing the errors directly in the input images. The visualized results are shown in Fig. 4.

This method is inappropriate for usage in real practice. The method of automated data processing from neural networks is called interpretation. To interpret the results, it is necessary to create a program that processes the results delivered by the neural network. Based on the interpretations, it is possible to directly influence the production process using the required parameters. The specific solution to the program is based on the process of managing the production process and the problem solved.

A necessary condition is to determine the probability values for which the object and its category will be recognized and considered as valid product identification.

The neural network can recognize multiple types of products in the picture with different probabilities. Each recognized object has a probability that corresponds to the mathematical expression of the object's recognition. Appropriate selection of boundary probability can prevent unambiguous interpretation of the results.

Then the required parameters control the manufacturing process. The specific solution of the program depends on the method of managing the production process and solving the problem.

After this decision, interpreting may occur depending on the specific problem. For example, if we know the location and method of occurrence of any error category, it is possible to automatically correct the respective manufacturing operation or use the position of the product in the image for positioning the robot gripper in order to grab it in real operation.

Subsequent visual control of outputs confirmed the assumption that the trained neural network RetinaNet can identify individual product categories and possibly recognize their visual impairments. The neural network has recognized all of the available washers, including those that were not used for training. The system has been able to repeatedly divide the washers by type and look for visual damage. In the second phase of verification of neural network functionality, the network was interfaced with the UR5 robot control program, which tested the washers according to the type and the damage. The type of damage information was stored in the database for further processing.

IV. CONCLUSION

The experiment has shown that designed and transformed neural networks can be used to solve visual control problems. The availability of predefined neural networks such as RetinaNet allows focusing on solving the quality problem itself instead of designing and programming the neural network.

The use of artificial intelligence is no longer a matter only for IT professionals. CNN becomes a tool that can be used directly in production. In such use, the selection of suitable methods for categorizing knowledge, interpreting results correctly and linking with production, becomes important. This should be the subject of further research on neural networks to manage production processes.

ACKNOWLEDGEMENT

This article was supported by the Agency for Research and Development under contract no. APVV-16-0355 and project KEGA 001TUKE4/2018 „Implementation of concurrent engineering philosophy to educational tool in the field of computer aided technological preparation“.

REFERENCES

[1] P. Sinčák, and G. Andrejková, "Neural Networks: Engineering approach (1. part). Elfa: Košice, 107 p., 1996. ISBN 80-88786-38.

[2] T. Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," In: Proceedings of the IEEE international conference on computer vision, p. 2980-2988, 2017. ISBN 978-1-5386-1032-9.

[3] P. Poór, T. Kamaryt, and M. Šimon, "Introducing autonomous maintenance by implementing OTH hybrid positions and TPM methods in metallurgical company," In: International Journal of Engineering and Technology, Vol. 7(3), p. 817-824, 2015.

[4] P. Poór, and J. Basl, "Czech Republic and Processes of Industry 4.0 Implementation," In: Proceedings of the 29Th International DAAAM Symposium 2018, p. 0454-0459, 2018.

Bayesian Artificial Neural Networks for Survival Modeling

[1] Hansapani Rodrigo, [2] Chris Tsokos
[1]University of Texas Rio Grande Valley, Edinburg, TX 78439, USA
[2]University of South Florida, Tampa, FL, 33620, USA
E-mail: [1]hansapani.rodrigo@utrgv.edu, [2]ctsokos@usf.edu

ABSTRACT

Artificial neural networks have been widely used in the field of pattern recognition within the past two decades. Bayesian learning of artificial neural networks is very useful in overcoming many of the inherited problems in neural networks, including the network over fitting, which is critical in obtaining generalized predictions with higher prediction accuracies. Despite its importance, very few studies have used a Bayesian neural network for survival predictions. Accurate prediction of patients survival is key for identifying the relevant treatment protocols in the various onset of cancers. In this study, we demonstrate the use of Bayesian artificial neural networks for accurate survival predictions. In fact, we discuss how to develop a piecewise constant hazard model using Bayesian neural networks. The uncertainties of the predictions are captured using the error bars.

Keywords - Artificial Neural Networks, Bayesian Learning, Piecewise Constant Hazard Model, Survival Prediction,

I. INTRODUCTION

An Artificial neural network (ANN) is an information processing archetype that is inspired by the biological neural networks systems, such as the brain. They have been successfully applied in almost every field including in engineering, computer science, and medicine [1]-[5]. An ANN has the strength of making predictions based on both individual attributable variables and possible complex interactions. Moreover, they serve as a powerful tool for modeling nonlinear functions and non-additive effects[6]. These are the main reasons ANN have become popular in different applications. However, they also bring their challenges. The main concern is that their final results are less interpretable[7].This can be overcome by building hybrid models using both neural networks and other statistical models like multiple regression, logistic regression, and multinomial logistic regression. An ANN consists of as several interconnected layers. In this work, we have focused on feed-forward neural networks, where each layer has a collection of artificial neurons (nodes) with connections made among the layers without any feedbackloops. ANN use a supervised learning technique where both inputs and outputs need to be fed into the network for the initial training process. Fig.1 represents the architecture of a feed- forward ANN with three layers, namely the input, hidden and output. Here we have assumed that it has d inputs, M hidden and K output nodes.

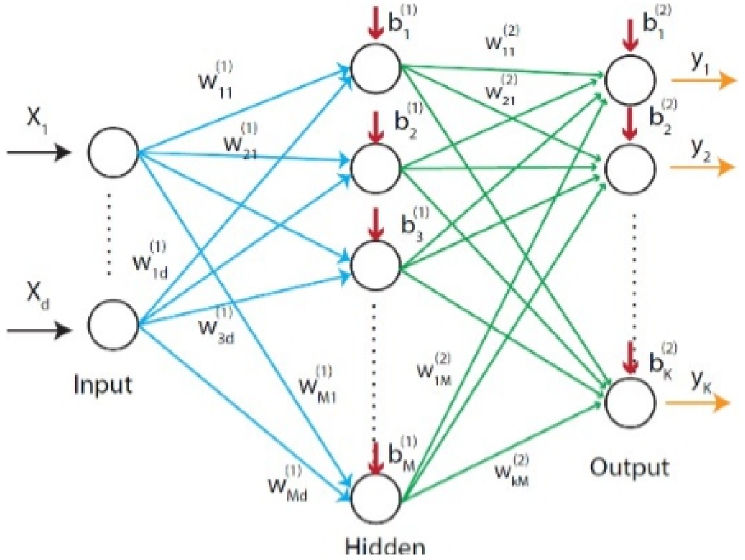


Fig.1. A Feed-forward Artificial Neural Network with 3 Layers: Input, Hidden, and Output

The outcome of a feed-forward ANN can be expressed by (1),

$$y_k(\boldsymbol{x}, \boldsymbol{w}) = g\left(\sum_{j=1}^{M} w_{kj}^{(2)} h\left(\sum_{i=1}^{d} w_{ji}^{(1)} x_i + b_j^{(1)}\right) + b_k^{(2)}\right) \qquad (1)$$

and this is simply a nonlinear function from a set of input variables xi to a set of output variables yk linked with adjustable weight parameters [8],

$$w = w_{11}^{1}, w_{12}^{1}, \ldots . , w_{21}^{2}, w_{22}^{2}, \ldots . w_{KM}^{2} .$$

A. Network Training and Error Function Network training plays a major role when using a neural network to find solutions to a given problem. By training, we refer to finding the optimal set of weight parameters w, using the training data which can be found by maximizing the relevant likelihood function of the data.

For example, if we consider a set of independent training data xn , tn with a joint probability density function p xn , tn ,then, we can write the likelihood function as in (2),

$$\begin{aligned} p\ D\ w, x &= \underset{n}{p\ x^{n}, t^{n}} \\ &= \underset{n}{p}\ t^{n} | x^{n}\ p\ x^{n} , \end{aligned} \qquad (2)$$

where p xn , tn is the conditional density t given x, and p xn is the unconditional density of x. Since it is more convenient to minimize the negative log likelihood function than maximizing the likelihood function, we introduce an error function of the form(3). Note that, we have removed the term, p xn as it does not effect on the network weight parameters,

$$E\ w = -\ln p\ D\ w,\ x$$
$$= -\ln p\ t^{n}\ |x^{n}\ . \qquad (3)$$
$$n$$

Hence, the choice of the error function entirelydepends on the conditional distribution. For our example of survival modeling, we will be using the Poisson distribution as the conditional distribution function.

B. Bayesian Artificial Neural Networks

Bayesian neural networks provide a more intuitive approach for network training. A significant amount of research in this area was conducted by David Mackay in 1992 [9]-[10]. In the maximum likelihood (ML) method, we find a single set of weight parameters by minimizing the error function. In contrast to the ML method, in the Bayesian approach, a probability distribution is used to capture the uncertainties associated with the weight parameters [11].

Use of Bayesian learning in ANNs provides several advantages. In fact, with this approach, the use of regularization parameters can be given a natural interpretation. Moreover, it allows to use a relatively large number of regularization parameters while optimizing them during the training process. The automatic relevance determination prior [12]-[14] helps to identify the relative importance of each input variable. Additionally, the variation in predictions in regression problemscan be captured and can be visualized using error bars. Moreover, prediction accuracies can be increased by creating network committees, combining different networks.

In the Bayesian setting, we first introduce a prior distribution p w for the weights, representing our knowledge on weight parameters. In our analysis, we have considered a zero meanGaussian prior of the form (4),

$$p\ \ wx\ = \frac{1}{\alpha}\ exp - \frac{\alpha}{2} w^{T} w\ Z_w$$
$$= \frac{1}{\alpha}\ exp\ -\alpha E_p\ w\ Z_w \qquad (4)$$

where $Z_w = \frac{2\pi}{\alpha}^{\frac{w}{2}}$ and α is the inverse variance of the distribution, also known as the hyper parameter of the prior distribution. As a part of Bayesian learning, we optimize this hyper parameter. The error term Ep (w)is chosen to be½ wTw, as it penalizes the weights of large magnitudes for a better generalization.

Once we observe the data, the Bayes' theorem is used to update our beliefs and the posterior probability density p w D, x of the weight parameters can be obtained as,

$$p\ w\ D,\ x = \frac{p\ D\ w,\ x\ p\ w}{p\ D\ x} \qquad (5)$$

Here, p D w, x is the likelihood function, and p D x is the normalization factor which is given by,

$$p\ D\ x = p\ D\ w,\ x\ p\ w\ x\ dw. \qquad (6)$$

We then use this posterior distribution to make inferences. That is to make new predictions based on,

$$p\ t\ x^{*}\ ,\ D = p\ t|x^{*}\ ,\ w\ p\ w\ x\ dw. \qquad (7)$$

II. SURVIVAL MODELING WITH BAYESIANANN

Accurate prediction of the survival is a challenging, yet substantial task which depends on the underlying the hazard function. These hazard functions can of tenbe complex and might not follow a particular distribution. Moreover, its behavior can significantly be affected by the risk factors which drives the function. Even with decades of research dedicated to survival analysis (and hence in hazardmodeling), medical practitioners still search for exclusive predictive models which can handle the modern biomedical data [15].

An efficient solution is to use flexible modeling of survival analysisutilizing techniques like kernel density[16], ANNs [17]-[19] and cubic splines [20].

Among those, ANN-based survivalanalysis models have widely used mainly due to the capability of handling complexnonlinear relationships among the predictor variables and due to the fewerassumptions involved with the modeling. Faraggi and Simon have used ANN as abasis for a non-linear proportional hazard model [17]. Another method based onpopular multi-layer perceptron: partial logistic regression has been developed by Biganzoliet al.[18].ANN has been used to predict the patientoutcome with censored survival data, including time as a covariate [19].

The development of the piecewise exponential model using ANN has first been proposed by Forniliet al. [21]. Their method accommodates greater flexibility in modeling complex hazard functions. In one of our work, we have extended their study by in corporating the Bayesian learning of network parameters[22].Additionally, for censored subjects, we used Kaplan-Meier[23] hazard probabilities in their ANN output nodes. In this study, we proposed a different approach to survival modeling without the Kaplan-Meier estimates. Our model providesaccurate survival predictions compared to [21] and other conventional methods like linear Poisson regression and generalized estimating equations (GEE). This has been demonstrated with lung cancer patient data taken from Surveillance, Epidemiology and End Results (SEER) program.

The modelperformances have been evaluated using root mean square error (RMSE), meanabsolute error (MAE), mean percentage error (MPE), and relative squared error(RSE). These error measurements help in assessing different aspects of prediction accuracies.

A. METHODOLOGY

Let T be the survival or the follow-up time for subjects i = 1,2, … . , N where T =min{Survival Time, Censoring Time}, and xbe the covariates. Let's assume that there areRnumber of competing risks, which causes the subject to observe the same eventof interest [24]. Then, (8) defines the hazard function for the rthrisk,

$$\lambda\ r, t, x = \lim_{\Delta t \to 0^+} \frac{P\ t < T \leq t + \Delta t, R = r\ T \geq t, x}{\Delta t} \qquad (8)$$

The corresponding survival and the probability density functions are given by(9) and (10),

$$S\ t, x = \exp - \lambda\ .\ , \overset{t}{\underset{0}{u}}, x\ du\ , \qquad (9)$$

and

$$f\ t\ x = \lambda\ .\ , u, x\ , S\ t, x\ , \qquad (10)$$

where $\lambda\ .\ , u, \overset{1}{x} = {}^{Ri}\ \overset{r=1}{\lambda}\ r, u, x$ for each individual with R possible competing risks. Thus, for independent observations, assuming non-informative censoring, the likelihood function L can be written as in (11),

$$L = \prod_{i=1}^{N} f(t_i|\boldsymbol{x_i})^{\delta_i} S(t_i|\boldsymbol{x_i})^{1-\delta_i} = \prod_{i=1}^{N} \frac{\lambda(.\,, t_i, \boldsymbol{x_i})}{\exp\left(\int_0^{t_i} \lambda(.\,, u,\ \boldsymbol{x_i}) du\ \right)}, \qquad (11)$$

where δ_i is equal to 0 if the subject i is censored and 1 otherwise. Under the piecewise constant hazard model, the follow-up timeT is divided into several disjoint timeintervalsa0, a_1 , … . , a_J where $a_0 = 0$ and $a_J = \infty$and the hazard function for rth risk is assumed to be constant during the jth time period a_{j-1} , aj . Hence, we have, $\lambda\ .\ , t, x_i = \lambda\ .\ , j, x_i$ where $\lambda\ .\ , t, x_i = \ {}^{R_i}_{r=1}\ \lambda\ r, j, x_i$ for each subject. Then, the modified likelihood function can bewritten as in (12),

$$L = \prod_{i=1}^{N} \frac{\prod_{j=1}^{J_i} (\lambda(.\,, j, \boldsymbol{x_i})^{\delta_{ij}})}{\exp\left(\sum_{j=1}^{J_i} \lambda(.\,, j, \boldsymbol{x_i}) \tau_{ij}\right)} = \frac{1}{\prod_{i=1}^{N} \prod_{i=1}^{J_i} \tau_{ii}^{\delta_{ij}}} \prod_{i=1}^{N} \prod_{j=1}^{J_i} \frac{\left(\lambda(.\,, j, \boldsymbol{x_i}) \tau_{ij}\right)^{\delta_{ij}}}{\delta_{ij}!\, exp\left(\lambda(.\,, j, \boldsymbol{x_i}) \tau_{ij}\right)}, \qquad (12)$$

where,

δ_{ij} = 1, if the i^{th} subject is deceased during the j^{th}interval 0, otherwise

J_i is the last interval that the subject i is observed and

τ_{ij} is the corresponding exposure time which is defined by,

$$\delta_{ij} = \begin{cases} a_j - a_{j-1}, \text{if} t_i \geq a_j \\ t_i - a_{j-1}, \text{if} a_{j-1} < t_i \leq a_j \\ 0, \text{if} t_i \leq a_{j-1} \end{cases}$$

The kernel given in (12) corresponds to the likelihood of a Poisson random variable δ_{ij} with a mean $\mu_{ij} = \lambda\,.\,, j, x_i\, \tau_{ij}$. By applying the logarithm on both sides of this, we get,

$$\log(\mu_{ij}) = \log(\lambda(.\,, j, \boldsymbol{x}_i)) + \log(\tau_{ij}) \qquad (13)$$

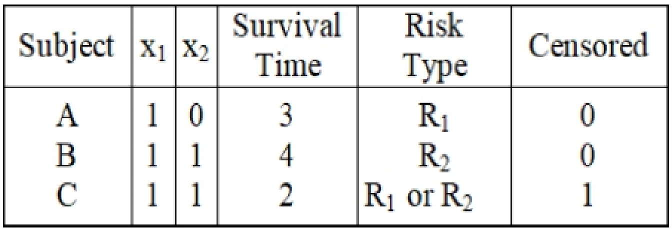
We can model, $\lambda\,.\,, j, x_i$ in (13) with a Poisson log- linear model of the form $\log \lambda\,.\,, j, x_i = \alpha_j + x_i\,\beta$ as given in [25]-[26]. Nevertheless, this approach has several difficulties in handling large number of δij observations which mightarise with substantial amount of subject data or/and longer follow up.

B. The Proposed Bayesian ANN Model

In this study, we introduce a Bayesian artificial neural network model to predict λ r, j, xi in (8). This new ANN model has several output nodes, each of which corresponds to a different time interval. This structure is similar to the ANN model used by Mani et al. [27].

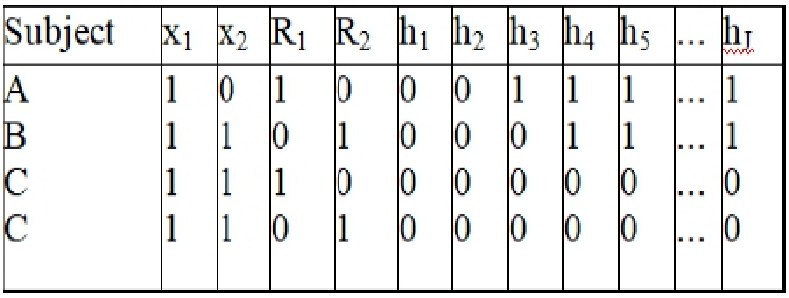
C. Data Preprocessing

We begin with the data preprocessing procedure. Let's assume that there are three subjects, namely A, B and C and they have been observed for J number of years. Information on their risk factors x_1 and x_2 , survival time and whether they are deceased or not during the given period are known. In particular, we have considered two competing risk types, R_1 or R_2, for each subject, where they can decease due to one of that reason. The "censor" variable indicates whether a subject has lost follow up somewhere during the study period or has been alive until the end of a study. Hence, for all deceased subjects during the study period, it is set to be zero. As can be seen, subject A and B have decreased due to risk types, R_1 and R_2 after 3 and 4 years, respectively. According to Table 1, subject C has lost follow-up after 2years.



Subject	x_1	x_2	Survival Time	Risk Type	Censored
A	1	0	3	R_1	0
B	1	1	4	R_2	0
C	1	1	2	R_1 or R_2	1

Table 1.Sample data



Subject	x_1	x_2	R_1	R_2	h_1	h_2	h_3	h_4	h_5	…	h_J
A	1	0	1	0	0	0	1	1	1	…	1
B	1	1	0	1	0	0	0	1	1	…	1
C	1	1	1	0	0	0	0	0	0	…	0
C	1	1	0	1	0	0	0	0	0	…	0

Table 2.Preprocessed data

The new ANN model requires data to be represented as in Table 2. Since, there are four inputs, covariates x_1 and x_2 and two indicator variables $R_1$and R_2, we need to create an ANN with 4 inputs. Censored subjects like C, can be exposed to any of the competing risks and hence, his/her information is presented twice into the model as given in Table 2. If we assume a constant hazard for each year, then there are J number of output nodes in the ANN. i.e., if a subject is alive or censored, then hj = 0, if a subject is deceased, then h_j = 1, 0, subject is alive or censord

$$h_j = 1, \text{ subject is decreased in thej}^{\text{th}} \text{ time interval}$$

D. Network Training

The hyperbolic tangent and exponential activation functions are used in the new ANN model in its hidden and output layers. The process of developing an exponential activation function is related to the nonlinear Poisson regression model [28]. The proposed ANN structure is represented in Fig. 2. The network output, y j r, x , gives the hazard for eachtime interval j, as in (14),

$$y(j|\boldsymbol{r},\boldsymbol{x}) = \lambda(\boldsymbol{r},j,\boldsymbol{x}) = \exp\left(b_{jh}^{(2)} + \sum_{h=1}^{K} w_{jh}^{(2)}\, tanh\left(b_{hl}^{(1)} + \sum_{l=1}^{d} b_{hl}^{(1)} x_l\right)\right). \quad (14)$$

where j = 1,2 … . . , J. Moreover, x_1 , … x_d are tne inputs, and w_{hl} and w_{jh} are the hidden and output layer weights.

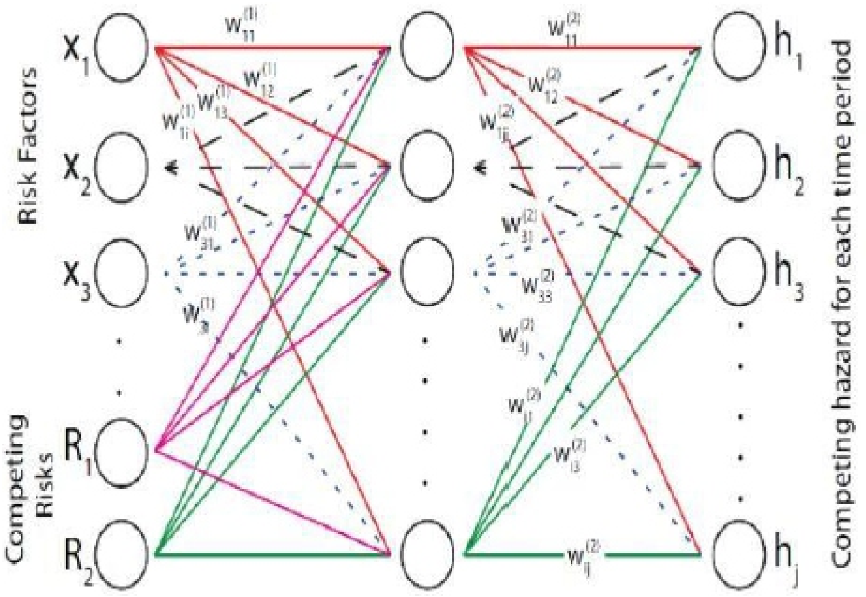


Fig.2.The proposed ANN model

We have used both ML and Bayesian approaches when training ANN models. A zero mean Gaussian prior has been used for the weight distribution with the Bayesian approach. During the training process, we minimized the regularized canonical error function given by (15), where α is the non-negative weight decay parameter. As per [29], we trained several ANN models with weight decay values with $\{0.01, 0.025, 0.5, 0.075, 0.1\}$.

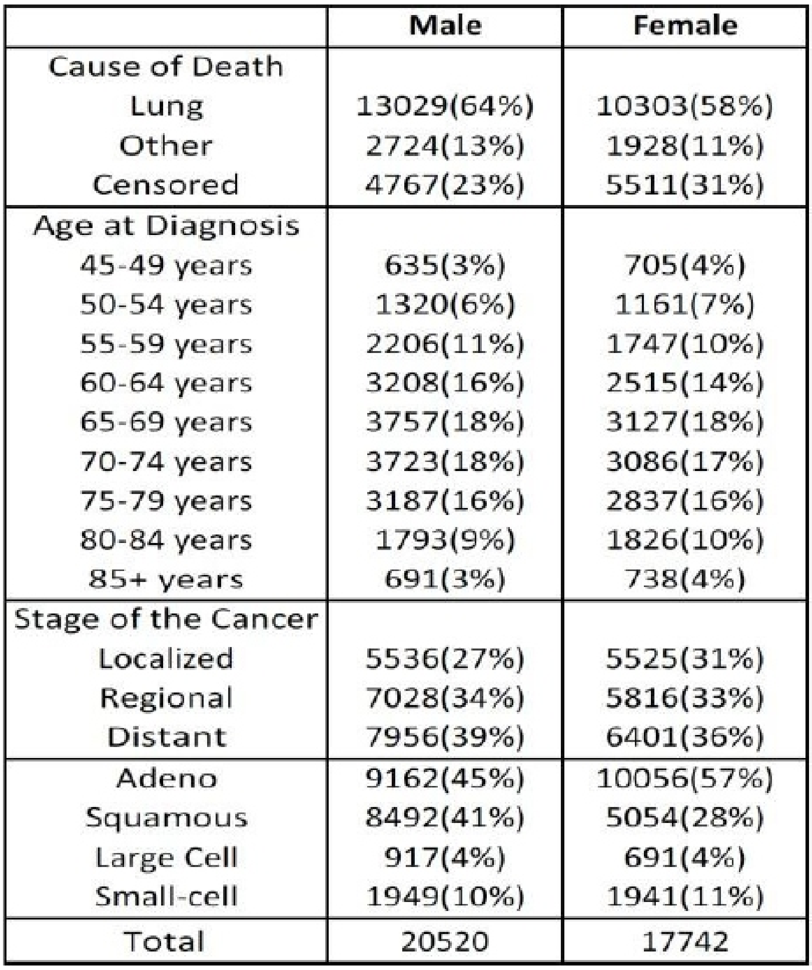
$$E = -\sum_{i=1}^{N}\sum_{j=1}^{J}\Big(h_j \log\big(y(j|r,x_i)\big) - y(j,r,x_i)\Big)\tau_{ij} + \frac{\alpha}{2}\left(\sum_{h=1}^{K}\sum_{l=1}^{d}\left(w_{hl}^{(l)}\right)^2 + \left(b_h^{(2)}\right)^2 + \sum_{j=1}^{J}\sum_{h=1}^{K}\left(w_{jh}^{(2)}\right)^2 + \left(b_j^{(2)}\right)^2\right) \qquad (15)$$

In the ML approach, we used a 5-fold cross- validation technique to find the optimal number of hidden nodes for each network. The optimal network for each decay value is chosen based on the minimum average validation error, and that is used for hazard predictions using the testing data. Automatic relevance determination prior is used to determine the relative importance of the risk factors. When using the Hybrid Monte Carlo (HMC) [14] and Hybrid Bayesian methods[28], we used a 5000 burn-in period, prior to sampling.

$$S_l(j) = exp\left(-\left(\sum_{j=1}^{J} y(j|\boldsymbol{r},\boldsymbol{x}_l)\right)\right) \qquad (16)$$

The corresponding survival probabilities are obtained using (16). We evaluated the models using several error measurements calculated based on the predicted median survival time and the actual survival time of the non-censored subjects in the testing data.

E. The Lung Cancer Data



	Male	Female
Cause of Death		
Lung	13029(64%)	10303(58%)
Other	2724(13%)	1928(11%)
Censored	4767(23%)	5511(31%)
Age at Diagnosis		
45-49 years	635(3%)	705(4%)
50-54 years	1320(6%)	1161(7%)
55-59 years	2206(11%)	1747(10%)
60-64 years	3208(16%)	2515(14%)
65-69 years	3757(18%)	3127(18%)
70-74 years	3723(18%)	3086(17%)
75-79 years	3187(16%)	2837(16%)
80-84 years	1793(9%)	1826(10%)
85+ years	691(3%)	738(4%)
Stage of the Cancer		
Localized	5536(27%)	5525(31%)
Regional	7028(34%)	5816(33%)
Distant	7956(39%)	6401(36%)
Adeno	9162(45%)	10056(57%)
Squamous	8492(41%)	5054(28%)
Large Cell	917(4%)	691(4%)
Small-cell	1949(10%)	1941(11%)
Total	20520	17742

Table 3.Lung Cancer patient Information

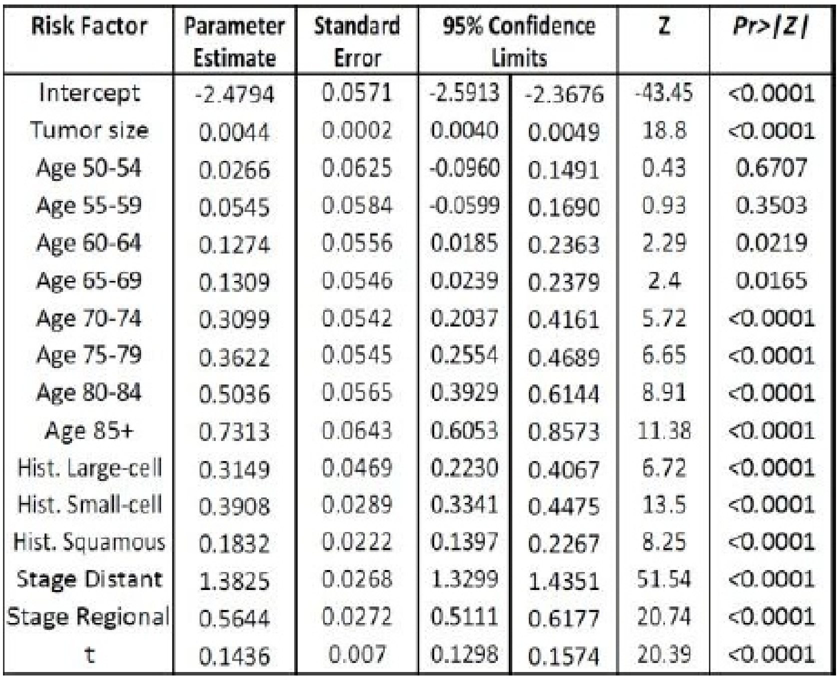
In our study, 38262 white lung cancer patients diagnosed from 2004 to 2009were selected from the Surveillance, Epidemiology and End Results (SEER) program [30]. Out of those patients, 23332 subjects were deceased due to lung cancer and 4652 were deceased due to other causes. The remaining were considered as censored due to their missing information or lost in the follow-up.

Four risk factors were taken into consideration: age at diagnosis, tumor size, histology, and the stage of cancer. The majority of patients were between the ages of 65-75 and many of them had distant metastasis (refer Table 3). Most of these patients were diagnosed with a deno or squamous cell carcinoma. The median follow-up time for males and females were 1.33 and two years respectively. The median tumor sizesfor those two groups were about 38 mm and 32 mm.

Our preliminary analysis confirmed the fact that the survival times between males and females to be significantly different from each other, as similar to [31].Therefore, we conducted two separate analyses for the two groups. To develop the piecewise constant hazard model, we partitioned the total follow-up time into six disjoint intervals, each with a 12-month period. For our analysis with GEE and ANN models, we have used SAS and MATLAB.

III. RESULTS

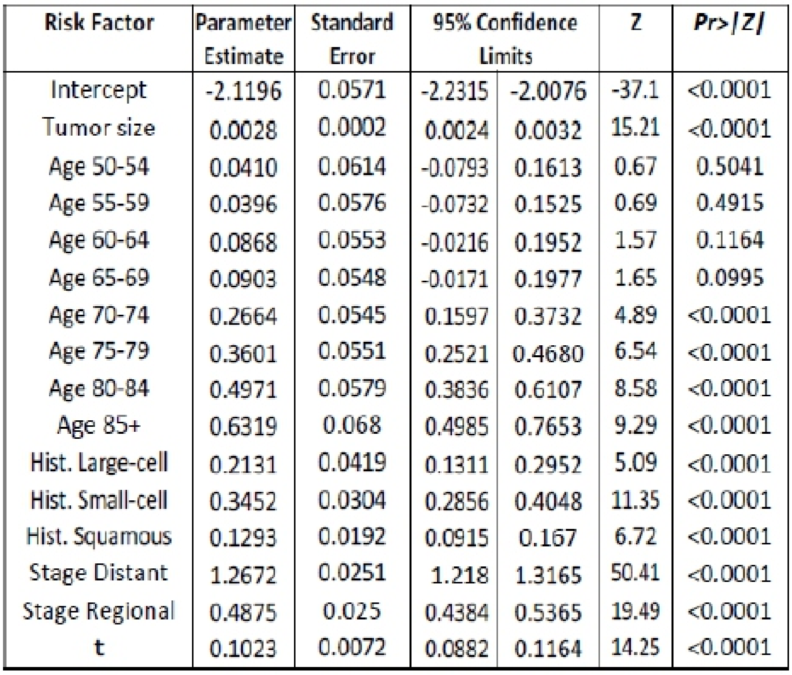
For both males and females, we created a training data set (70%) and a testing data set (30%). The training set was used to train the models while the testing dataset was used to evaluate the prediction accuracies of the proposed models.



Risk Factor	Parameter Estimate	Standard Error	95% Confidence Limits		Z	*Pr>|Z|*
Intercept	-2.4794	0.0571	-2.5913	-2.3676	-43.45	<0.0001
Tumor size	0.0044	0.0002	0.0040	0.0049	18.8	<0.0001
Age 50-54	0.0266	0.0625	-0.0960	0.1491	0.43	0.6707
Age 55-59	0.0545	0.0584	-0.0599	0.1690	0.93	0.3503
Age 60-64	0.1274	0.0556	0.0185	0.2363	2.29	0.0219
Age 65-69	0.1309	0.0546	0.0239	0.2379	2.4	0.0165
Age 70-74	0.3099	0.0542	0.2037	0.4161	5.72	<0.0001
Age 75-79	0.3622	0.0545	0.2554	0.4689	6.65	<0.0001
Age 80-84	0.5036	0.0565	0.3929	0.6144	8.91	<0.0001
Age 85+	0.7313	0.0643	0.6053	0.8573	11.38	<0.0001
Hist. Large-cell	0.3149	0.0469	0.2230	0.4067	6.72	<0.0001
Hist. Small-cell	0.3908	0.0289	0.3341	0.4475	13.5	<0.0001
Hist. Squamous	0.1832	0.0222	0.1397	0.2267	8.25	<0.0001
Stage Distant	1.3825	0.0268	1.3299	1.4351	51.54	<0.0001
Stage Regional	0.5644	0.0272	0.5111	0.6177	20.74	<0.0001
t	0.1436	0.007	0.1298	0.1574	20.39	<0.0001

Table 4.Analysis of GEE parameter estimates: males

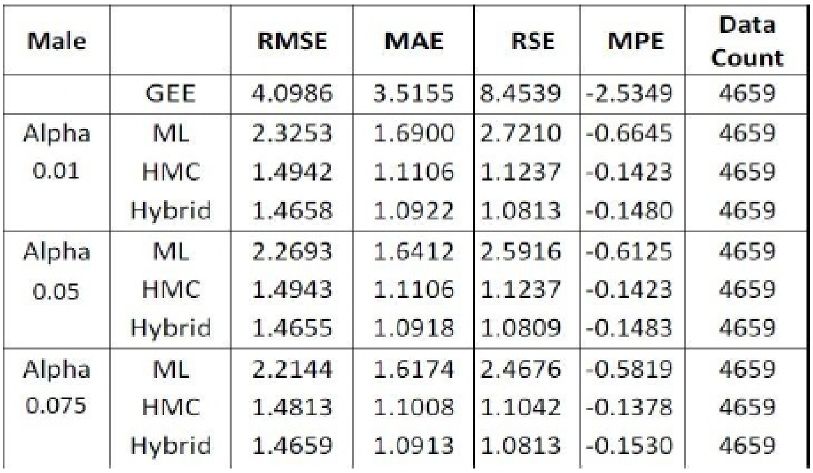
Poisson regression models were not able to capture the true variance of the data as revealed by its deviance and the Pearson chi-square statistics [32],due to being susceptible to correlated observations. Poisson models with over dispersion parameters and negative binomial models resulted in the same conclusion. Hence, an alternative method, generalized estimating equations (GEE) model has been considered.



Risk Factor	Parameter Estimate	Standard Error	95% Confidence Limits		Z	Pr>|Z|
Intercept	-2.1196	0.0571	-2.2315	-2.0076	-37.1	<0.0001
Tumor size	0.0028	0.0002	0.0024	0.0032	15.21	<0.0001
Age 50-54	0.0410	0.0614	-0.0793	0.1613	0.67	0.5041
Age 55-59	0.0396	0.0576	-0.0732	0.1525	0.69	0.4915
Age 60-64	0.0868	0.0553	-0.0216	0.1952	1.57	0.1164
Age 65-69	0.0903	0.0548	-0.0171	0.1977	1.65	0.0995
Age 70-74	0.2664	0.0545	0.1597	0.3732	4.89	<0.0001
Age 75-79	0.3601	0.0551	0.2521	0.4680	6.54	<0.0001
Age 80-84	0.4971	0.0579	0.3836	0.6107	8.58	<0.0001
Age 85+	0.6319	0.068	0.4985	0.7653	9.29	<0.0001
Hist. Large-cell	0.2131	0.0419	0.1311	0.2952	5.09	<0.0001
Hist. Small-cell	0.3452	0.0304	0.2856	0.4048	11.35	<0.0001
Hist. Squamous	0.1293	0.0192	0.0915	0.167	6.72	<0.0001
Stage Distant	1.2672	0.0251	1.218	1.3165	50.41	<0.0001
Stage Regional	0.4875	0.025	0.4384	0.5365	19.49	<0.0001
t	0.1023	0.0072	0.0882	0.1164	14.25	<0.0001

Table 5.Analysis of GEE parameter estimates: females

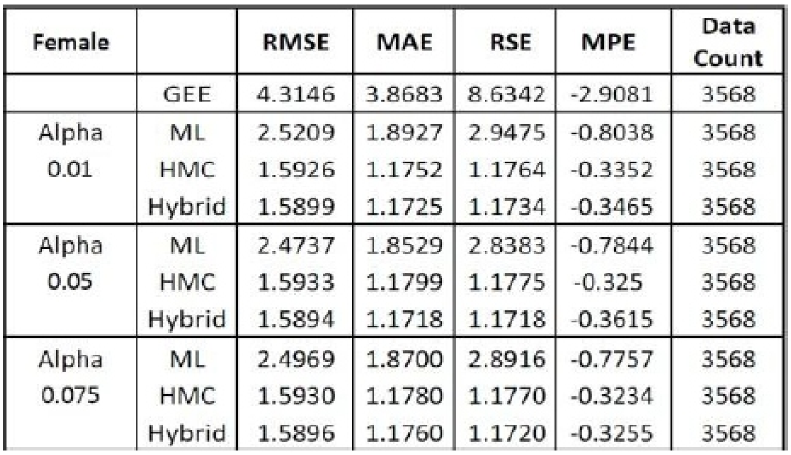
Using GEE, we created two different statistical models for males and females, which are given in Tables 4, and 5, respectively. We can see that, for each 10 mm increase in the tumor size, the hazard rate for males increases by 4% and by 3% for females. In general, as patients get older, their lung cancer hazard rates get increased. Furthermore, we can see that the patients diagnosed with small cell carcinoma have the highest hazard compared to other histology types. For males, their hazard is 48% higher than the patients with adeno cell carcinoma. For females, it is about 41%. Over time, the hazard rates seem to increase rapidly for males than females. Applying these two models, we were able to predict the hazard and to obtain the corresponding survival probabilities for our lung cancer testing data.



Male		RMSE	MAE	RSE	MPE	Data Count
	GEE	4.0986	3.5155	8.4539	-2.5349	4659
Alpha 0.01	ML	2.3253	1.6900	2.7210	-0.6645	4659
	HMC	1.4942	1.1106	1.1237	-0.1423	4659
	Hybrid	1.4658	1.0922	1.0813	-0.1480	4659
Alpha 0.05	ML	2.2693	1.6412	2.5916	-0.6125	4659
	HMC	1.4943	1.1106	1.1237	-0.1423	4659
	Hybrid	1.4655	1.0918	1.0809	-0.1483	4659
Alpha 0.075	ML	2.2144	1.6174	2.4676	-0.5819	4659
	HMC	1.4813	1.1008	1.1042	-0.1378	4659
	Hybrid	1.4659	1.0913	1.0813	-0.1530	4659

Table 6.Model evaluation for males

Next, we created the ANN models with different learning techniques, ML and Bayesian. In each situation, we created several ANN models by varying the number of hiddennodes from 3 to 13and also used different weight decayvalues. As mentioned earlier, the optimal networks in the ML method are selectedbased on the minimum average validation error. In the Bayesian approach, weused the minimum of regularized cost function to find the best set of models.By using each optimal network, we predicted the hazard and correspondingsurvival probabilities for the testing data. In order to evaluate the predictionaccuracies of different ANNs and GEE, we used the actual survival times andtheir predicted median survival times of non-censored subjects in the same dataset. For a better comparison, we calculate several prediction errors, including the root mean square error (RMSE), mean absolute error (MAE), mean percentage error (MPE), and relative squared error (RSE) as given in Tables 6 and 7.



Female		RMSE	MAE	RSE	MPE	Data Count
	GEE	4.3146	3.8683	8.6342	-2.9081	3568
Alpha 0.01	ML	2.5209	1.8927	2.9475	-0.8038	3568
	HMC	1.5926	1.1752	1.1764	-0.3352	3568
	Hybrid	1.5899	1.1725	1.1734	-0.3465	3568
Alpha 0.05	ML	2.4737	1.8529	2.8383	-0.7844	3568
	HMC	1.5933	1.1799	1.1775	-0.325	3568
	Hybrid	1.5894	1.1718	1.1718	-0.3615	3568
Alpha 0.075	ML	2.4969	1.8700	2.8916	-0.7757	3568
	HMC	1.5930	1.1780	1.1770	-0.3234	3568
	Hybrid	1.5896	1.1760	1.1720	-0.3255	3568

Table 7.Model evaluation for females

As per Tables 6 and 7, we can see that the Bayesian approach tends to provide better predictions compared to both GEE and ML methods, for both genders. Although the predictions from Bayesian ANN show negative MPE values, which indicates underestimations of the survival, that is significantly less than of other models. The smallest error values were found with the Hybrid Bayesian approach which was trained using a weight decay value of 0.05, for both genders. Further analysis on patients' hazard survival was carried out using those two models. Fig. 3 depicts the variation in the survival probabilities among males and females patients according to different tumor sizes while keeping the other categorical risk factors in their mode categories. We can see that, as the tumor sizeincreases men tend to have a lesser survival probability compared to females.

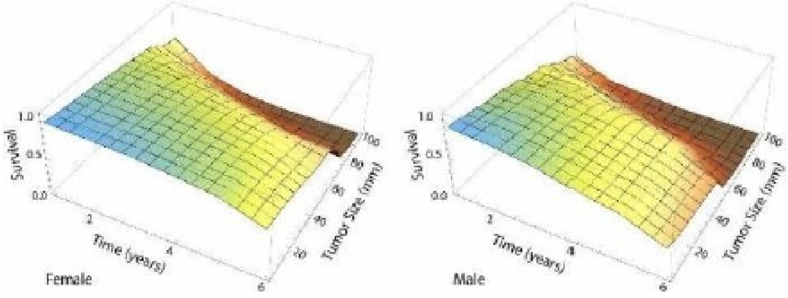


Fig. 3.Tumor size vs survival probabilities for females and males

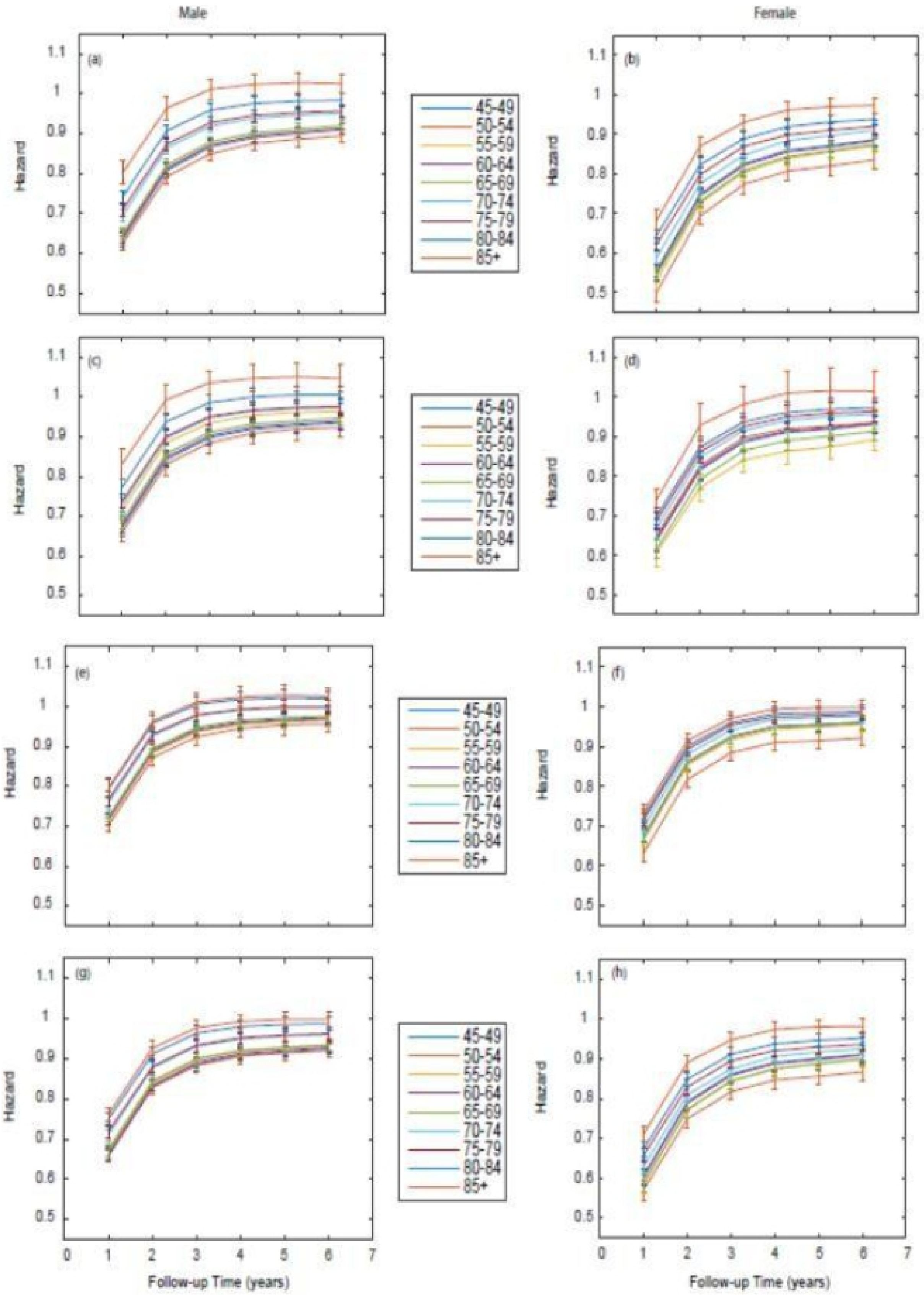


Fig. 4.Hazard variation for males and females for different histology types (a) Male-Adeno (b) Female-Adeno (c) Male- Large cell (d) Feale- Large cell (e) Male-Small cell (f) Female- Small cell (g) Male- Squamous cell (h) Female-Squamous cell

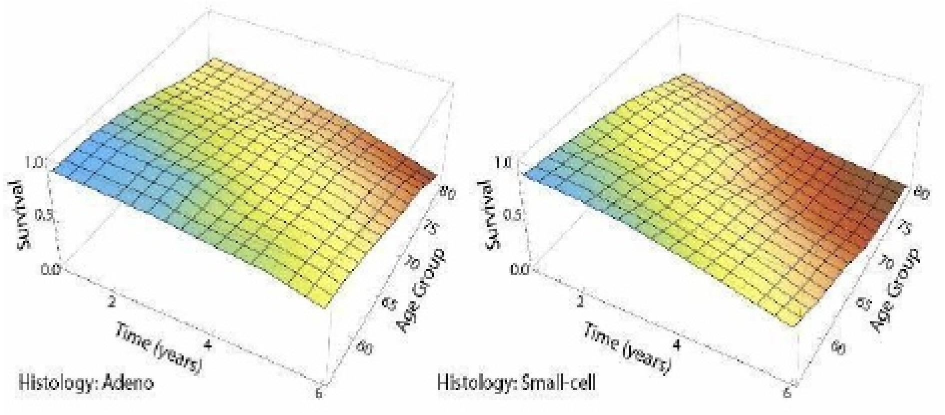


Fig. 5.Survival probabilities of females with different histology types

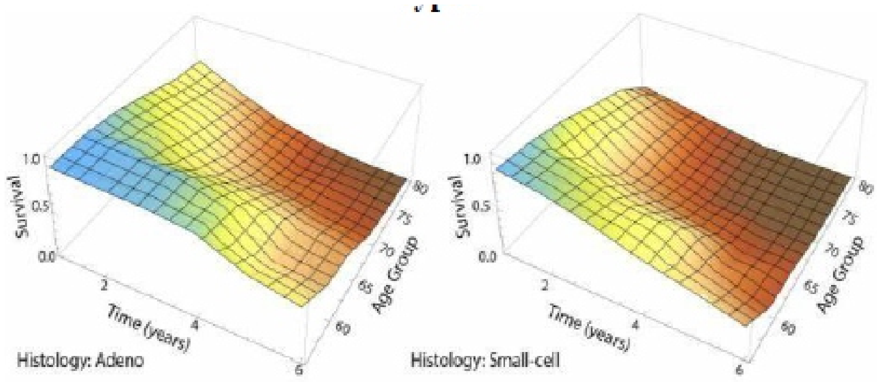
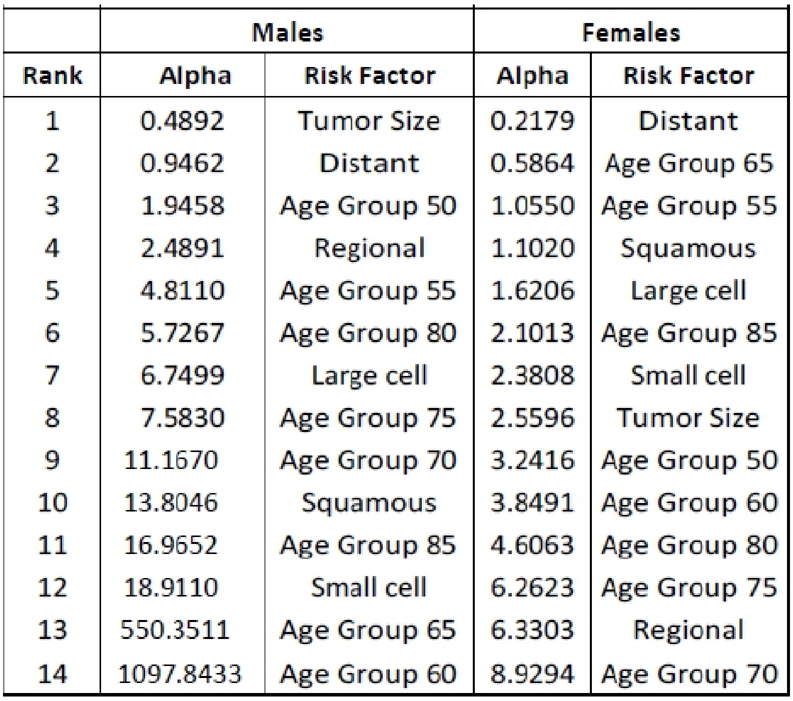


Fig. 6.Survival probabilities of males with different histology types

It is a known fact there is a significant variation between the hazard rates among the different histology types for different genders. Fig. 4 represents the variation in the hazard rates we obtained from our models according to the patients' age group and histology types, for both males and females. The left panel of Fig. 4 shows the hazard for males with their histology types, adeno, large cell, small cell and squamous cell carcinoma, respectively. The right panel shows the hazard for females for the same histology types. Error bars were created to represent one standard deviation from our mean hazard predictions. Unlike in the GEE approach, these error bars have not been affected by the underestimated standard errors of parameter estimates. For all the histology types, males have a higher hazard than females. Moreover, a higher hazard can be seen for the patients who diagnosed with small cell carcinoma. In fact, this is the most dangerous lung cancer out of the four we have considered in our analysis. Older patients show a relatively higher hazard in both genders.



	Males		Females	
Rank	Alpha	Risk Factor	Alpha	Risk Factor
1	0.4892	Tumor Size	0.2179	Distant
2	0.9462	Distant	0.5864	Age Group 65
3	1.9458	Age Group 50	1.0550	Age Group 55
4	2.4891	Regional	1.1020	Squamous
5	4.8110	Age Group 55	1.6206	Large cell
6	5.7267	Age Group 80	2.1013	Age Group 85
7	6.7499	Large cell	2.3808	Small cell
8	7.5830	Age Group 75	2.5596	Tumor Size
9	11.1670	Age Group 70	3.2416	Age Group 50
10	13.8046	Squamous	3.8491	Age Group 60
11	16.9652	Age Group 85	4.6063	Age Group 80
12	18.9110	Small cell	6.2623	Age Group 75
13	550.3511	Age Group 65	6.3303	Regional
14	1097.8433	Age Group 60	8.9294	Age Group 70

Fig. 5 manifests the variation in the survival for different age groups, for the patients who diagnosed with small cell carcinoma and adeno cell carcinoma. It further confirms the fact that patients with small cell carcinoma have a significantly lower survival probability compared to the other group. This pattern remains the same for all the age groups.

Fig. 6 represents the variation in the survival probabilities of males according to the age group and the same histology types, adeno,and small cell carcinoma. Similarly, to females, we can see the same survival patterns as for males. However, in overall, men tend to have lower survival probabilities compared to females [33].

We used ARD prior toidentifying the relevant importance of the risk factors into the network. Table 8 summarizes the rankings of those risk factors based on these hyperparameter values. Risk factors with smaller hyperparameters are highly contributing to the model outcome. Tumor size and distant metastasis are the top two key factors which highly contribute to the Male ANN model.

For females, the most contributing key factors include the distant metastasis and being in the age group of 65. These rankings confirm the fact that our findings have a faithful agreement between the true nature of lung cancer survival.

REFERENCES

[1] T. Ayer, O. Alagoz, J. Chhatwal, J.W. Shavlik, C. E. Kahn, E.S. Burnside, "Breast cancer risk estimation with artificial neural networks revisited: Discrimination as calibration,"Cancer, vol. 116, pp. 3310-3321, July 2010.

[2] C. E. Floyd, J. Y. Lo, A. J. Yun, D. C. Sullivan, P. J. Kornguth, "Prediction of breast cancer malignancy using an artificial neural network,"Cancer, vol. 74, pp. 2944-2948, Dec.1994.

[3] R. K. Orr, "Use of an artificial neural network to quantitate risk of malignancy forabnormal mammograms", Surgery, vol. 129, pp. 459-466, Apr. 2001.

[4] Y. Wu, M. L. Giger, K. Doi, C. J. Vyborny, R. A. Schmidt, C. E. Metz, "Artificialneural networks in mammography: application to decision making in the diagnosisof breast cancer,"Radiology, vol. 187, pp. 81-87, Apr. (1993)

[5] H. Rodrigo, C. P. Tsokos, T. Sharaf,"Regularized Neural Network to Identify Potential Breast Cancer: A Bayesian Approach," A Bayesian Approach. Journal of Modern Applied StatisticalMethods, vol. 15, pp. 563-579, Nov. 2016.

[6] M.J.Somers, J.C. Casal, "Using Artificial Neural Networksto ModelNonlinearity,"OrganizationalResearch Methods,vol. 12, pp. 403-417, July 2009.

[7] J.V. Tu, "Advantages and disadvantages of using artificial neural networks versus logistic regression for predicting medical outcomes,"Journal of Clinical Epidemiology, vol. 49, pp. 1225-31, Nov. 1996.

[8] C. M. Bishop, "Pattern Recognition and Machine Learning", Springer, 2006.

[9] D. J. C. Mackay, "The Evidence Framework Applied to Classification Networks," Neural Computation,vol. 4, pp.720-736,Sept.1992.

[10] D. J. C. Mackay, "Information-based objective functions for active data selection," Neural Computation, vol. 4, pp. 590- 604, July 1992.

[11] C. M. Bishop, "Neural networks for pattern recognition", Oxford university press," 1995.

[12] D. J. C. Mackay, "Bayesian methods for backpropagation networks," In: E. Domany, J. L. van Hemmen, K. Schulten (eds.), Models of Neural Networks III, Chapter 6, New-York: Springer-Verla (1994).

[13] D. J. C. Mackay, "Bayesian non-linear modelling for the 1993 energy predictioncompetition," In: G. Heidbreder,(ed) Maximum Entropy and Bayesian Methods,Santa Barbera, 1994.

[14] R. M. Neal, "Bayesian Learning for Neural Networks", Ph.D. thesis, University of Toronto, Canada, 1994.

[15] Z. Ma, A.W. Krings,"Survival analysis approach to reliability, survivability and prognostics and health management", IEEE Aerospace Conference. pp. 1-20, 2008

[16] S. Diehl, W. Stute, "Kernel density and hazard function estimation in the presenceof censoring" Journal of Multivariate Analysis, vol. 25, pp. 299-310, May 1988.

[17] D. Faraggi, R. A. Simon, "A neural network model for survival data", Statistics inMedicine, vol. 14, 73-82, Jan. 1995.

[18] E. Biganzoli, P. Boracchi, L. Mariani, E. Marubini, "Feed forward neural networks for the analysis of censored survival data: a partial logistic regression approach,"Statistics in Medicine, vol. 17, pp. 1169-1186, May 1998.

[19] P. M. Ravdin, G. M. Clark, "A practical application of neural network analysis forpredicting outcome of individual breast cancer patients," Breast Cancer Researchand Treatment, vol. 22, 285-293, Oct. 1992.

[20] P. Borachi, E. Biganzoli, E. Marubini, "Joint modelling of cause-specific hazardfunctions with cubic splines: An application to a large series of breast cancer patients,"Computational Statistics and Data Analysis, vol. 42, pp. 243-262,Feb. 2003.

[21] M. Fornili, F. Ambrogi, P. Boracchi, E. Biganzoli, "Piecewise Exponential Artificial Neural Networks (PEANN) for Modeling Hazard Function with Right Censored Data," In: E. Formenti, R. Tagliaferri, E. Wit, (eds.) Computational IntelligenceMethods for Bioinformatics and Biostatistics, vol. 8452, Springer,2014.

[22] H. Rodrigo, C.P. Tsokos, "Artificial Neural Network Model for Predicting Lung Cancer Survival," Journal of Data Analysis and Information Processing, vol. 5, pp.33-47, Feb. 2017.

[23] A.E.L.Kaplan, P. Meier, "Nonparametric Estimation from Incomplete Observations," Journal of the American Statistical Association, vol.53, pp.457-481, June 1958.

[24] J. M. Satagopan, L. Ben-Porat, M. Berwick, M. Robson, D. Kutler, D. Auerbach, "A note on competing risks in survival data analysis," British Journal of Cancer, vol. 91, pp. 1229- 1235Oct. 2004.

[25] T. R. Holford, "The Analysis of Rates and of Survivorship Using Log-Linear Models,"Biometrics, vol. 36, pp. 299-305, Jun 1980

[26] N. Laird,D. Olivier, "Covariance analysis of censored survival data using log-linearanalysis techniques",Journal of the American Statistical Association," vol.76, pp. 231-240, Jun. 1981.

[27] D. Mani, J. Drew, A. Betz, P. Datta, "Statistics and data mining techniques forlifetime value modeling," In Proceedings of the fifth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 94-102, 1999.

[28] H. Rodrigo, C. P. Tsokos, "Bayesian Artificial Neural Networks in Health and Cybersecurity," Ph.D. thesis, University of South Florida, USA, 2017.

[29] B. D. Ripley, "Pattern Recognition and Neural Networks," Cambridge UniversityPress, 1996.

[30] Surveillance Epidemiology and End Results (SEER) Program, Research Data (1973-2009), Division of Cancer Control and Population Sciences, National Cancer Institute, Surveillance Research Program, Surveillance Systems Branch, 2012.

[31] J. P. Wisnivesky E. A. Halm, "Sex differences in lung cancer survival: Do tumorsbehave differently in elderly women?" Journal of Clinical Oncology,vol. 25, pp. 1705-1712,May 2007.

[32] A. Pedan, "Analysis of Count Data Using the SAS System," In Proceedings of the Twenty-Sixth Annual SAS Users Group International Conference, Long Beach, 247, 2001.

[33] K. E. Osann, J. T. Lowery,M. J., Schell, "Small cell lung cancer in women: risk associated with smoking, prior respiratory disease, and occupation," Lung Cancer,vol, 28, pp. 1-10, Apr 2000.

Instructions for Authors

Essentials for Publishing in this Journal

1 Submitted articles should not have been previously published or be currently under consideration for publication elsewhere.

2 Conference papers may only be submitted if the paper has been completely re-written (taken to mean more than 50%) and the author has cleared any necessary permission with the copyright owner if it has been previously copyrighted.

3 All our articles are refereed through a double-blind process.

4 All authors must declare they have read and agreed to the content of the submitted article and must sign a declaration correspond to the originality of the article.

Submission Process

All articles for this journal must be submitted using our online submissions system. http://enrichedpub.com/ . Please use the Submit Your Article link in the Author Service area.

-

Manuscript Guidelines

The instructions to authors about the article preparation for publication in the Manuscripts are submitted online, through the e-Ur (Electronic editing) system, developed by **Enriched Publications Pvt. Ltd**. The article should contain the abstract with keywords, introduction, body, conclusion, references and the summary in English language (without heading and subheading enumeration). The article length should not exceed 16 pages of A4 paper format.

Title

The title should be informative. It is in both Journal's and author's best interest to use terms suitable. For indexing and word search. If there are no such terms in the title, the author is strongly advised to add a subtitle. The title should be given in English as well. The titles precede the abstract and the summary in an appropriate language.

Letterhead Title

The letterhead title is given at a top of each page for easier identification of article copies in an Electronic form in particular. It contains the author's surname and first name initial .article title, journal title and collation (year, volume, and issue, first and last page). The journal and article titles can be given in a shortened form.

Author's Name

Full name(s) of author(s) should be used. It is advisable to give the middle initial. Names are given in their original form.

Contact Details

The postal address or the e-mail address of the author (usually of the first one if there are more Authors) is given in the footnote at the bottom of the first page.

Type of Articles

Classification of articles is a duty of the editorial staff and is of special importance. Referees and the members of the editorial staff, or section editors, can propose a category, but the editor-in-chief has the sole responsibility for their classification. Journal articles are classified as follows:

Scientific articles:

1. Original scientific paper (giving the previously unpublished results of the author's own research based on management methods).

2. Survey paper (giving an original, detailed and critical view of a research problem or an area to which the author has made a contribution visible through his self-citation);

3. Short or preliminary communication (original management paper of full format but of a smaller extent or of a preliminary character);

4. Scientific critique or forum (discussion on a particular scientific topic, based exclusively on management argumentation) and commentaries. Exceptionally, in particular areas, a scientific paper in the Journal can be in a form of a monograph or a critical edition of scientific data (historical, archival, lexicographic, bibliographic, data survey, etc.) which were unknown or hardly accessible for scientific research.

Professional articles:

1. Professional paper (contribution offering experience useful for improvement of professional practice but not necessarily based on scientific methods);
2. Informative contribution (editorial, commentary, etc.);
3. Review (of a book, software, case study, scientific event, etc.)

Language

The article should be in English. The grammar and style of the article should be of good quality. The systematized text should be without abbreviations (except standard ones). All measurements must be in SI units. The sequence of formulae is denoted in Arabic numerals in parentheses on the right-hand side.

Abstract and Summary

An abstract is a concise informative presentation of the article content for fast and accurate Evaluation of its relevance. It is both in the Editorial Office's and the author's best interest for an abstract to contain terms often used for indexing and article search. The abstract describes the purpose of the study and the methods, outlines the findings and state the conclusions. A 100- to 250-Word abstract should be placed between the title and the keywords with the body text to follow. Besides an abstract are advised to have a summary in English, at the end of the article, after the Reference list. The summary should be structured and long up to 1/10 of the article length (it is more extensive than the abstract).

Keywords

Keywords are terms or phrases showing adequately the article content for indexing and search purposes. They should be allocated heaving in mind widely accepted international sources (index, dictionary or thesaurus), such as the Web of Science keyword list for science in general. The higher their usage frequency is the better. Up to 10 keywords immediately follow the abstract and the summary, in respective languages.

Acknowledgements

The name and the number of the project or programmed within which the article was realized is given in a separate note at the bottom of the first page together with the name of the institution which financially supported the project or programmed.

Tables and Illustrations

All the captions should be in the original language as well as in English, together with the texts in illustrations if possible. Tables are typed in the same style as the text and are denoted by numerals at the top. Photographs and drawings, placed appropriately in the text, should be clear, precise and suitable for reproduction. Drawings should be created in Word or Corel.

Citation in the Text

Citation in the text must be uniform. When citing references in the text, use the reference number set in square brackets from the Reference list at the end of the article.

Footnotes

Footnotes are given at the bottom of the page with the text they refer to. They can contain less relevant details, additional explanations or used sources (e.g. scientific material, manuals). They cannot replace the cited literature.
The article should be accompanied with a cover letter with the information about the author(s): surname, middle initial, first name, and citizen personal number, rank, title, e-mail address, and affiliation address, home address including municipality, phone number in the office and at home (or a mobile phone number). The cover letter should state the type of the article and tell which illustrations are original and which are not.

Address of the Editorial Office:

Enriched Publications Pvt. Ltd.
S-9,IInd FLOOR, MLU POCKET,
MANISH ABHINAV PLAZA-II, ABOVE FEDERAL BANK,
PLOT NO-5, SECTOR -5, DWARKA, NEW DELHI, INDIA-110075,
PHONE: - + (91)-(11)-45525005