SAINTGITS

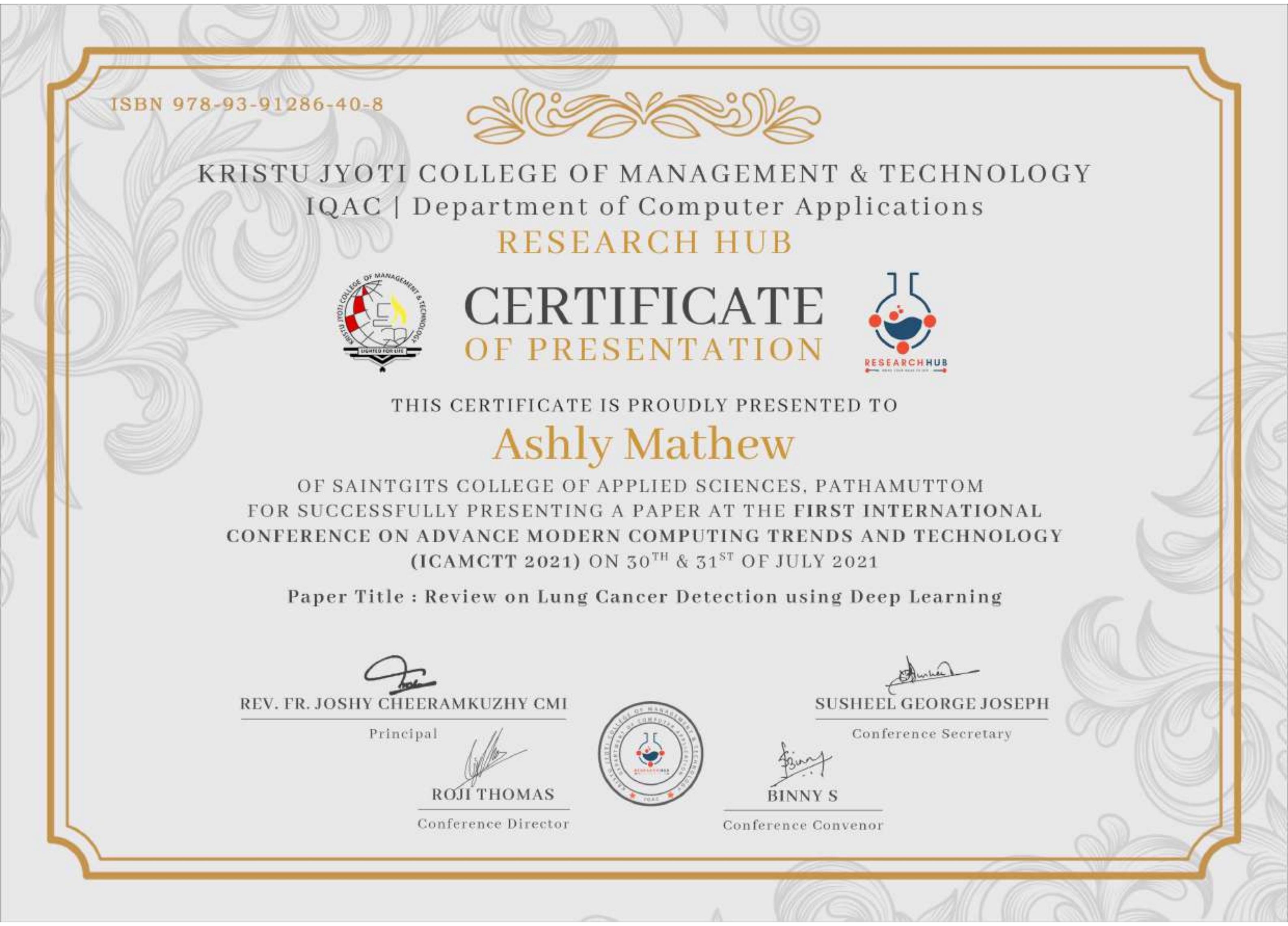
COLLEGE OF APPLIED SCIENCES

SAINTGITS
LEARN.GROW.EXCEL

Criterion 3: Research, Innovations and Extension

3.3.2 Number of books and chapters in edited volumes/books published and papers published in national/international conference proceedings per teacher

ASHLY MATHEW



ISBN 978-93-91286-40-8

KRISTU JYOTI COLLEGE OF MANAGEMENT & TECHNOLOGY
IQAC | Department of Computer Applications
RESEARCH HUB

CERTIFICATE
OF PRESENTATION

THIS CERTIFICATE IS PROUDLY PRESENTED TO

Ashly Mathew

OF SAINTGITS COLLEGE OF APPLIED SCIENCES, PATHAMUTTOM FOR SUCCESSFULLY PRESENTING A PAPER AT THE **FIRST INTERNATIONAL CONFERENCE ON ADVANCE MODERN COMPUTING TRENDS AND TECHNOLOGY (ICAMCTT 2021)** ON 30TH & 31ST OF JULY 2021

Paper Title : Review on Lung Cancer Detection using Deep Learning

REV. FR. JOSHY CHEERAMKUZHY CMI
Principal

SUSHEEL GEORGE JOSEPH
Conference Secretary

KRISTU JYOTI COLLEGE OF MANAGEMENT & TECHNOLOGY
DEPARTMENT OF COMPUTER APPLICATION
IQAC

ROJI THOMAS
Conference Director

Binny
BINNY S
Conference Convenor

ISSN 2394-3777 (Print)
ISSN 2394-3785 (Online)
Available online at www.ijartet.com

International Journal of Advanced Research Trends in Engineering and Technology (IJARTET)
Vol. 8, Special Issue 1, August 2021

REVIEW ON LUNG CANCER DETECION USING DEEP LEARNING

JITHU VARGHESE
BCA Department
Saintgits College of Applied Sciences
Pathamuttomm, Kottayam-686532
jithuvj.bca1922@ saintgits.org

SHIJU KOSHY VARGHESE
BCA Department
Saintgits College of Applied Sciences
Pathamuttomm, Kottayam-686532
shijukv.bca1922@saintgits.org

ASHLY MATHEW
Asst. Professor
BCA Department
Saintgits College of Applied Sciences
Pathamuttomm, Kottayam-686532
ashly.mathew@saintgits.org

ABSTRACT

This review provides a detailed approach to detecting lung cancer from CT (computed tomography) scans and other methods using deep residual learning. Scans provide valuable data in diagnosing lung diseases. The primary aim of this work is to compare the provided lung data input for the identification of cancerous lung-nodules . The effectiveness of the cancer prediction system helps people know their cancer risk at a low cost, and also helps them make the right decision based on their cancer risk status.

KEYWORDS

Deep learning, neural networks, Radiographs, CT scans, Residual Units, lung nodules, Artificial Neural Network, Gradient Tree Boosting.

INTRODUCTION

The worldwide leading cause for cancer-related deaths is due to lung cancer. The significant steps in detecting early-stage cancer are figuring out whether any pulmonary nodules inside the lungs could develop into a tumor. This review aims to determine the likelihood that a given CT scan data of the lungs will be cancerous.

The Computer-Aided Diagnostic system (CAD) is an efficient medical diagnostic system and is essential for the practicality of medical imaging today. The doctor creates an additional second opinion using the Computer-Aided Diagnostic system for obtaining an accurate diagnosis for the treatment to be effective.

Deep Learning consists of multiple layers of processing to achieve a high level of abstraction when learning data representations in different domains such as speech recognition and visual object recognition. Deep learning techniques for segmenting medical images received a lot of interest because they can process and learn large amounts of data quickly and accurately.

EXISTING METHODS

Different architectures are proposed and compared in various studies, mainly the Convolutional-Neural-Network (CNN) and its variations. CNN can be used in both 2D (known as 2D CNN / ConvNet) and 3D data (known as 3D CNN / C3D / 3D ConvNet). CNN is very similar to normal neural networks, neural network's consist of neurons with learnable weights and distortions. Every neuron in the neural network, receives some input, performs a dot product on it, and is conditionally tracked with a nonlinearity modified for different applications and data sets. For image segmentation, CNN is modified and various architectures formed such as UNet, SegNet, FCN, Enet, DenseNet, DilatedNet, PixelNet, ICNet, ERFNet, DeconvNet and many more in FPSO in order to reduce the complexity of computation in the CNN . FPSOCNN improves CNN's efficiency. This review also explores the possibility of using an artificial neural network (ANN) model for the detection the existing cancerous lung nodules. The ANN plays a significant role in better data set analysis, classification and feature extraction of these cancerous nodules. Another method, GCMS is a method used to identify a specific chemical in the human body that aids in diagnosing lung cancer.

LITERATURE SURVEY

Hiroki Yamagishi, et.al [1] conveyed in their research "Lung Cancer Diagnosis Deep Learning Application Test for Medical Sensor Systems". The main goal is to create simple health monitoring system that can predict some diseases by analyzing and detecting data such as breath, saliva, and urine, which can be collected without harming the human body. Biomarkers in human urine can be identified by gas chromatography, mass spectrometer (GCMS) which are converted into numerical data such as properties, retention time, mass-to-charge ratio, and ionic strength, images in Deep Neural Network are classified into some classes by extracting features from image pixel data, Deep Neural Networks are also effective in identifying lung cancer patients from GC-MS data of a patient's urine.

The patient's urine is converted into three-dimensional data by GCMS, this data is input into the neural network. The data are normalized before the calculation and is used in the primary layer of the neural network. The neural network's output is expressed in one of two values: either the patient has lung cancer or the person does not. The neural network is repeatedly trained with the backpropagation method and results in better accuracy. We evaluate the results and optimize the normalization path. Learning parameters and network structure. The approach is evaluated by comparing precision, sensitivity and specificity. Sensitivity is the proportion of estimated cancer patients to actual cancer patients, specificity is the proportion of estimated healthy patients to healthy patients, and precision is the proportion of correctly estimated patients to all patients, this shows the condition and results of the experiment, the proposed method works well and achieved an accuracy of 90 percent, this precision is sufficient for the pre-diagnosis and means that this method has shown the possibility of detecting lung cancer without any medical knowledge or experience, but with just Deep Neural Network.

The system uses GCMS data from human urine to successfully identify lung cancer. However, the instrument of the GCMS system is too large to be used in daily life. This research teams goal is to mount this system on a small device without a GCMS. The team is developing a flexible sensor node that can combine several small sensors on-site and process detection data. The final challenge was to find a smaller, more effective set of chemical sensors and a mapping of the function of the deep neural network and analyze the Deep Neural Network they had generated to make it more compact for the purpose.

Bohdan Chapaliuk, et.al [2]. A common tool for diagnosing lung cancer is a computed tomography (CT) scan. CT scans consist of a several x-rays shows a 3D visual of the specific tissue being scanned. When stacking, all serial x-ray images can be incorporated into a 3D image of the scanned body. The CT scan is treated as 3D medical data and this data is utilized in the progression of the automated diagnostic model. Specific spots in the lung that may show features of cancer is known as the lung nodule, and have a diameter of 7-30mm Related work methods used to resolve memory problem several strategies might be used: -2D CNN, 3D CNN (vnet or RCNN based approach) RNN (recurrent neural network) (2D CNN combined by RNN). This dataset contains CT scans of over a 1000 patients with labels on which patient has lung cancer. The data of one patient consist of the set of x-ray images, where each image looks like a slice of a human lung and lung cancer label which was determined after an year of scanning. DSB dataset contains 1397 patients scan images in the training dataset, 198 patient scan images in validation set and 506 patient scans in the test set. Available training set is highly unbalanced and contains 1035 samples which do not contain lung cancer and 362 samples which are confirmed with cancer. The experiments had checked several types of approaches to determine lung cancer in the computed tomography images. First, trained C3D and 3D DenseNet network for whole image classification. They show quite similar results, however, DenseNet show a bit higher accuracy. Results on neural networks which are trained for identifying Lung Cancer on the entire lung's 3D image show worse accuracy in comparison to the two-stage approach, when two different neural networks are trained for segmentation and classification. Recurrent neural networks show competitive accuracy and performance, primary goal of the research is to increase the network ability to learn with less quantities of data and an ability to use high-resolution patient data.

Diksha Mhaske, et.al [3] conveys in their paper, Machine-learning-algorithms as well as deep-learning algorithms are the two emerging techniques that have recently attracted many researchers. Deep-learning methods have also achieved great success in computer and technological vision. The technique used here perform a uniform feature extraction, classification framework for users and also free them from handcrafted feature extraction which are troublesome. Deep-learning techniques offer the opportunity to increase the efficiency of the early detection of diseases. This work aims to develop an advanced computer-aided diagnostic (CAD) system with the help of deep learning algorithms that extracts data from images of CT scans and gives accurate information. Here, deep learning techniques, namely the Convolutional Neural Network and the Recurrent Neural Network, are being used to propose a model for computed tomography diagnosis of lung cancer and to achieve high precision. The traditional method of entering segmented CT scans directly into CNN 3D for classification proved inadequate. In [2], therefore, a modified UNet which was trained on LUNA16 data was used in order to recognize nodules in the lungs. The most likely nodal candidates were located by the UNet output and entered into a Convolutional-Neural-Networks (CNN), which classified the CT scan as either positive or negative for having lung cancer. The planned CAD system performed better than currently in use CAD systems, allowing for more efficient training, better accuracy, and greater generalization to other cancers. An image-based CAD algorithm had been created that uses regions with CNN characteristics (RCNN) to identify lung abnormalities. RCNN was used to detect different categories of lung abnormalities such as pulmonary nodules and diffuse lung disease unique for lung cancer detection using image analysis problems, a new deep learning algorithm was proposed to learn high-level image representation to achieve high classification precision in binary medical image classification tasks. They evaluated the model on Kaggle Data Science Bowl 2017 (KDSB17) data set, and compared it with some related works proposed in the Kaggle competition. Was to accurately model the form of the recognized lung nodules with the help of a new seventh-order MGRF model. The two groups of traits fed to a deep autoencoder (AE) classifier to differentiate between malignant (cancerous) and benign (non-cancerous) nodules with a detection accuracy of 91.20%. In it a Unet architecture was suggested for segmenting lung CT images. This proposed architecture consisted of a shrinking path that extracted high-level information and a symmetrical expansion path that recovered the required information. Cube coefficient Method used is CNN-LSTM system, The Lung Image Database Consortium (dataset) is used in this work, the dataset (LIDC-IDRI) consists of lung cancer screening CT scans (thoracic). In conclusion the proposed system is a hybrid CNN-LSTM model used for Lung cancer detection. The process starts with accepting CT Images. These CT images are further pre-processed and segmented. Finally the classification is done using proposed CNN-LSTM algorithm. In this, CNN model performs the feature extraction and LSTM model performs prediction and classification. The proposed CNN-LSTM system is compared with other existing detection models based on the accuracy measure. This work aims to improve the accuracy of the prediction systems. This aim is achieved by the proposed system as it provides a precision of 97%, which currently is the highest accuracy achieved so far.

Jong Hyuk Lee, et.al [4] conveys in their paper, performance of a deep learning algorithm on chest radiographs for detection of lung cancer in a health screening population is unknown. Tests performed on samples of deep learning algorithms performed with chest x-rays from individuals who underwent a full medical examination (validation test) between July and December 2008, Detection of visible lung cancer to evaluate the area under the operative characteristic of the receptor, curve (AUC) and diagnostic measures, including sensitivity and false positive rate (FPR); The performance of the algorithm was compared to the performance a of radiologists using the McNemar test along with the Moskowitz method; In addition, the deep learning algorithm was applied to a screening cohort that underwent chest radiography between 2008 January and 2012 December and their performance was calculated. The results in a validation test consisting of 10285 radiographs belonging to 10202 individuals of which 5857 were men with 10 radiographs of confirmed lung cancers, the algorithm showed comparable sensitivity (90%) to that of a radiologists (60%), In short, a deep learning algorithm detected lung cancer on chest x-rays with comparable accuracy to that of a radiologists and helps radiologists predict lung cancer in healthy populations with better accuracy and prediction rates.

Ruchita Tekade et.al[5], Hongyang Jiang offers a different approach to preprocessing lung CT photographs before delivering them to the CNN architecture. This results in better results as there
There are so many unmapped areas that the correctness by feature extraction can be reduced. Objects can overlap in 2D images, which mean that detection can have a high false-positive rate. Therefore, Xiaojie Huang et al. makes use of 3D cnn images to find pulmonary nodules with the us of 3D cubes from lung CT scans. Since 3D images provide a clearer picture of objects, 3DCNN compares well with 2DCNN BotongWu etal. 3D UNet architecture. In this thesis, nodule identification & malignancy prediction are performed simultaneously by learning high-level attributes from the lower part of UNet and 3DCNN, where segmentation was done. On the basis of this literature review it is concluded that 3DCNN is always better to achieve good results from the application. The combination of different approaches enables a different handling the information and also delivers better results.

Dataset used within the paper is from TCIA repository named as, LIDC-IDRI(LungImage DatabaseConsortiumandImageDatabaseResource Initiative). This data contains 1010 patient cases and 1018 scans acquired from them in DICOM format. This dataset also contains the labels of malignancy level by the lung nodule. There are 4 levels of malignancy mentioned in the dataset as 0 = Unknown, 1 = Benign or non- malignant disease, 2 = Malignant, Primary lung cancer, 3 = malignant metastatic. Benign are the lung tissues which grow gradually and this growth stop at certain point. These tissues are commonly non- cancerous and does not affect seriously to health. And malignant tissues are cancerous and grow very fast. These tissues can affect to other body parts also.

Kaggle Data Science Bowl was a competition held in 2017 to increase the efficacy of algorithms for categorizing the cancer and to detect if nodules of CT scans are malignant i.e. cancerous. The dataset have two stages but only first stage is used in the paper. This stage 1 data contains 1595 patient cases with 285380 computerized tomography picutres in DICOM format. If nodule is benign then value is 0 and if malignant then it 1. These labels are made use to classify the cancer. Some information are from LIDC-IDRI and some from LUNA16 and Kaggle Data Science Bowl 2017 are combined to identify nodules location in CT scans and categorize the cancer types respectively.

The ultimate aim of the technique is to improve the efficiency of detecting nodules in lungs and degree of malignancy prediction using pulmonary CT images. This experiment is performed with LIDC IDRI, LUNA16 and Data Science Bowl 2017 data sets on Tesla K20 GPUs with CUDA. The UNet architecture is opted for segmenting the nodules in lungs from lung computed tomography images and the proposed VGG-like 3D multi-path architecture is intended to classify nodules in lungs and predict the extend of malignancy. This is useful in predicting whether or not the patient will have cancer in the next two years.

Siddharth Bhatia et.al[6], conveys in this article that the by Hua et. al simplifies the image analysis process for conventional computer aid lung cancer diagnostics. Sun etal, experimented with Convolutional Neural Networks (CNN), Deep Belief Networks (DBN) and Automatic Noise Encoder (SDAE) in the collection of the LIDCIDRI (LungImageDatabaseConsortium), with accuracies 79%, 81% and 79%, respectively. LIDCIDRI image collection contains CT (computedtomography) scans of chest for the diagnosis and detection of annotated lung cancer with lesions. It consists of a thousand or more high-risk patient scans in DICOM image format. Each scan contains a bunch of images with multiple axial slices of the

thoracic cavity. Each scan has a variable number of 2D, slices, which can vary by different device performing the scan and the patient.

have a header that contains the details about the patient id, also other scan parameters such as the slice thickness.

Deepresidualnetworks have emerged as a family of extremely deep architectures that have convincing precision and good convergence behavior. DeepResidualNetworks(ResNets) consist of many stacked "Residual Units". Each subsequent layer in a deepneuralnetwork is basically only responsible for fine-tuning the result of the previous layer by simply adding a learned "remainder" to the input. This differs from a more traditional way where every layer had to generate the whole desired output.

Extreme Gradient Boosting builds upon the criteria of “boosting” many vulnerable predictive fashions right into a robust one, with the structure of ensemble of vulnerable fashions which is referred as Gradient Tree Boosting. There is a lot of gradient tree boosting algorithms, however particularly XGBoost makes use of second-order approach with the help of using Friedman etal and employs a greater regularized version formalization to manipulate over-fitting, which offers it higher performance. Random Forest Classifier is a meta-estimator based on subsampling over many decision trees which controls over-fitting well. The basis of random forest is that randomization over many decision trees can increase the accuracy of the general classification by boosting the selection rates of features that contribute more toward the classification among others.

The preprocessing step includes a chain of packages of vicinity developing and morphological operations. It identifies and separates the lung systems and nodules to resource the characteristic extraction. Segmenting lungs from the CT test objectives to become aware of distinguishing functions to resource the classifier and classify the applicants better. This is likewise crucial because of fact the CT test is simply too large to be fed into the classifier directly. It will take quite a few time , for classifier to become aware of differentiating featured from the large DICOM images. Segmentation process of lung systems may be very difficult trouble in general due to the fact that there may be no homogeneity withinside the lung location. There are comparable densities within the pulmonary structures. The process of lung segmentation was observed with the aid of using normalization and 0 centering. We are able to get an efficiency of 84% using the combination of UNet+RandomForest and ResNet+XGBoost which individually have accuracies 74% and 76%, respectively.

Through this paper, we propose an approach to lung cancer detection employing feature extraction using deep residual networks. We examine overall performance of tree-primarily based totally classifiers like Random Forest and XGBoost. The maximum accuracy we get is 84% by using ensemble of Random Forest and XGBoost classifier.

A. Asuntha et.al[7], makes use of quality function extraction strategies which include Histogram of orientated Gradients (HoG), wavelet transform-primarily based totally functions, Local Binary Pattern (LBP), Scale Invariant Feature Transform (SIFT) and Zernike Moment. After extracting texture, geometric, volumetric and depth functions, Fuzzy Particle Swarm Optimization (FPSO) set of rules is implemented for choosing the quality function. Finally, those functions are categorized the use of Deep learning. A novel FPSOCNN decreases computational difficulty of CNN. A greater valuation is finished on every other dataset that came from Arthi Scan Hospital which is a real-time data set. From the experiments conducted, it's far proven that FPSOCNN plays higher than different strategies.

During the work, first the entered photo is more suitable by the help of using histogram equalization for photo evaluation and de-noised by means of the usage of Adaptive Bilateral Filter (ABF). After pre-processing, the following step is to discover the lung location extraction. To find the lung location, ArtificialBeeColony (ABC) segmentation method is executed. After locating the region of the cancerous lung nodules the next procedure is to categorize the lung disease name and its severity primarily based totally at the function extraction. An advanced CNN approach primarily based totally upon FPSO to lower the computational difficulty by CNN is put forward. FPSOCNN enhances the capabilities of CNN.

A FPSO consists of well known base, that contains data given by the means of the professionals, by linguistic manipulate fuzzy rules, a fuzzification interface, which has the effect of reworking crisp statistics into fuzzy sets, an inference system, that makes use both of them together with the understanding base to make inference by the help of reasoning approach, and a defuzzification interface, that interprets the fuzzy manipulate motion as the result acquired to a real manage motion the usage of a defuzzification approach. A CNN includes one or greater convolutional layers & pooling layers. Pooling layers also known as sub sampling layers. Normally CNN are used for category purpose. Here, CNN is made use to categorize the lung most cancers disease. Pooling-layer is made use to carry out down sampling. It is made use to lessen the quantity of computation time with the aid of using lowering the extracted functions in convolution layer. There are styles of pooling layers, max and average pooling. In max pooling, most important pixels' value is taken into consideration with the receptive discipline of the filter. In common pooling, the common of every values is taken into consideration with the receptive area. Pooling layer result is passed as source to the convolution layer. CNN has very excessive computational price for big characteristic maps. CNN is gradual to educate big characteristic maps. To conquer the disadvantage of CNN, FPSOCNN (FuzzyParticleSwarm OptimizationConvolutionNeuralNetwork) is proposed. This reduces excessive computation price and improves speed. The size reduction of photograph area is found out by the help of using vector of capabilities which is created using FPSO from multidimensional photograph area to low dimensional characteristic space. This technique appreciably reduces the variety of capabilities for the lung cancer ailment classification. Instead of using pooling ideas max and average in CNN, PSO and GA are carried out. Lung cancer photographs are gathered from Aarthi Scan Hospital, Tirunelveli, Tamilnadu,

India. Aarthi Scan Hospital dataset includes almost a thousand lung photographs. The dataset was originally taken from sufferers in Digital Imaging Communication Medicine (DICOM) photographs.

A FPSOCNN is put forth which reduces computational difficulty of CNN. This paper makes use of excellent characteristic extraction strategies which includes HistogramofOrientated Gradients(HoG), wavelet transform-primarily based capabilities, LocalBinaryPattern (LBP), ScaleInvariantFeatureTransform (SIFT) & the zernike Moment. Following the extraction, FuzzyParticleSwarmOptimization (FPSO) set of rules is carried out for choosing the excellent characteristic. An extra valuation is executed on other dataset coming from Arthi Scan Hospital that is the real-time statistics record. From experimental outcomes, it is proven that FPSOCNN plays higher than different techniques.

In future, further improvization will be conducted in the classification of pulmonary nodules & optimize the proposed version. In fact, the similarly work might be grading the photos primarily based totally at the degree of nodules, which is of valuable importance for the analysis and remedy of lung most cancers in scientific applications.

Ola Kweik et.al[8], on this paper explores the opportunity of utilizing ArtificialNeuralNetwork (ANN) version to come across the presence of the lung carcinoma in someone’s body. The functions of this have a look at are:

- To apprehend a few suitable elements that leads in lung carcinoma.
- To version an ArtificialNeuralNetwork that may be used to come across the existence of lung carcinoma ArtificialNeuralNetworks (ANNs) are similar to the neural networks and provide a pretty suitable method, which resolves the trouble of categorization and prediction. ANN, a mathematical version which is endorsed with organization and useful characteristic of natural neural networks, Neural networks contain source and output layers, also (in maximum cases) hidden layers that remodel the source into something that output layer can use.

When a neural network is made use for cancer detections, the ANN Model undergo 2 stages, training and validation. First, the network is skilled on a dataset. Then weights of connections among neurons are constant so a test is conducted on network to decide the classifications of a brand new dataset. Throughout the session, we have used approximately 67% of the entire pattern records for network training, and 33% for network validation.

The dataset were downloaded which represents whether or not the sufferers have lung most cancers or not. This dataset can be located in DataWorldWebsite.

We did a few preprocessing at the records, after which we skilled our ANN version and tested it. We did a few preprocessing and transformation so the records were extra appropriate for predictive analysis. We used the primary 15 attributes as inputs to our version and the lung cancer characteristic because the anticipated output primarily based upon the source attributes. We normalized the attributes: gender, age, lung cancers. Gender scope turns into 1 for male, zero for female, lung cancer scope turns into 1 (yes), zero (No). However, age characteristic normalized to turns to be actual on account of the fact this is higher for ANN.

An ANN for diagnose the existence of lung cancer in sufferers were developed. The version was tested and gave a precision of 99.01%. This observation confirmed that neural community is capable of diagnose lung cancer, so it is made use as a diagnose device through physician.

Brahim AIT SKOURT et.al[9], conveys on this paper that a specific form of machine learning which is composed of a couple of processing layers to attain excessive stages of abstraction with regards to studying representations of records is known as deep learning. In distinctive domain names including speech recognition and visual object recognition. ConvolutionalNeuralNetwork (CNN), also a branch of machine learning approach and also a category of deep learning, lately supersedes many image segmentation approaches. It is based on multiple layer processing, to high model level and complicated extraction in data.

As with photo classification, CNN has had big achievement on photo segmentation problems. In 2015, FullyConvolutionalNetwork (FCN) was delivered through Long etal. and made CNN structure famous for dense prediction with no absolutely related layers, this novel method allowed to generate segmentation maps for any photo and changed into much quicker compared to classical strategies of photo segmentation.

Fully related layers have now been no longer the challenge, however also pooling the layers that lessen the item information, thus, the up-sampling layers had been followed to address this issue. Hence, this method changed into using with the encoder-decoder architecture, wherein the encoder reduces the spatial measurement of items with pooling layers and decoder recovers the item info with up-sampling layers. The U-net, the structure followed on within the work, is among the famous architectures in the category of the encoder-decoders.

In the trial phase, the supply photos and their corresponding masks are made use to coach U-net, and on the test phase, we offer a picturegraph as supply to generate the corresponding mask as output. And then, we comply with the mask of corresponding picturegraph to phase the place of interest, i.e. the lungs in our case. Lung segmentation performed by the U-net network does not contain substance of trachea and the bronchus regions & doesn’t eliminate lesions such as nodules and portions of blood vessels, which implies the followed approach is accurate.

Through the paper, we offered a lung parenchyma segmentation using U-net structure and we received a perfect segmentation with 0.9502 Dice-coefficient index. The advantage by the method offered on the paper is that, it is uniform and may be used by a huge area of various clinical picture segmentation tasks. Our goal within the subsequent stage is to perform a lung-nodule segmentation primarily based at the effects of the proposed work.

CONCLUSION

In end of this review paper, deep learning algorithm detected lung cancer in chest-radiographs with an overall performance similar to radiologists, with a purpose to be beneficial for the radiologists in healthful populations to analyze lung cancer. Accurate detection of size and area of the lung cancer performs a crucial function in diagnosis of lung cancer. Various algorithms based on deep learning are made use to find the lung carcinoma in chest-radiographs, MRI scans with a high rate of nodule detection which helps in recognizing lung cancer by early stages.

REFERENCE

[1]Hiroki Yamagishi, Mototsugu Hamada, Ryota Shimizu, Shusuke Yanagawa, Tadahiro Kuroda, Toru Shimizu and Yasutaka Monde, (2016). "Deep Learning Application Trial to Lung Cancer Diagnosis for Medical Sensor Systems" 2016 IntwenationalSoC Design Conference.

[2] Bohdan Chapaliuk, Yuriy Zaychenko, (2018)."Deep learning approachh in computer-aided detection system for lung cancer". 2018 IEEE First International Conference on System Analysis & Intelligent Computing (SAIC).

[3]Diksha Mhaske , Kannan Rajeswari and Ruchita Tekade, (2019)."Deep Learning Algorithm for Classification and Prediction of Lung Cancer using CT Scan Images" 5th International Conference On Computing, Communication, Control And Automation (ICCUBEA).

[4]Chang Min Park, Eui Jin Hwang, Hye Young Sun, Hyungjin Kim, Jin Mo Goo, Jong Hyuk Lee and Sunggyun Park, (2020). "Performance of a deep learning algorithm compared with radiologic interpretation for lung cancer detection on chest radiographs in a health screening population" Epub 2020 Sep 22.

[5]Ruchita Tekade, Prof. Dr. K. Rajeswari (2018) “Lung Cancer Detection and Classification using Deep Learning” Fourth International Conference on Computing Communication Control and Automation (ICCUBEA).

[6] Siddharth Bhatia, Yash Sinha and Lavika Goel (2019) “Lung Cancer Detection: A Deep Learning Approach”.

[7] A. Asuntha & Andy Srinivasan (2020) “Deep learning for lung Cancer detection and classification”.

[8] Ola Mohammed Abu Kweik, Mohammed Atta Abu Hamid, Samer Osama Sheqlieh, Bassem S. Abu-Nasser, Samy S. Abu-Naser (2020) “Artificial Neural Network for Lung Cancer Detection” International Journal of Academic Engineering Research (IJAER) Vol. 4 Issue 11.

[9] Brahim AIT SKOURT, Abdelhamid EL HASSANI, Aicha MAJDA (2018) “Lung CT Image Segmentation Using Deep Neural Networks” The First International Conference On Intelligent Computing in Data Sciences.

ISSN 2394-3777 (Print)
ISSN 2394-3785 (Online)
Available online at www.ijartet.com

International Journal of Advanced Research Trends in Engineering and Technology (IJARTET)
Vol. 8, Special Issue 1, August 2021

Big data Analytics in Cyber Security

PRATIBHA ANN JAYESH [1] , MERLIN MATHEW [2], ASHLY MATHEW [3]

SCHOLAR, BCA DEPARTMENT ,SAINTGITS COLLEGE OF APPLIED SCIENCES , PATHAMUTTOM , KOTTAYAM, INDIA [1]

SCHOLAR, BCA DEPARTMENT ,SAINTGITS COLLEGE OF APPLIED SCIENCES , PATHAMUTTOM , KOTTAYAM, INDIA [2]

SCHOLAR, BCA DEPARTMENT ,SAINTGITS COLLEGE OF APPLIED SCIENCES , PATHAMUTTOM , KOTTAYAM, INDIA [3]

***Abstract* — Linkage of personal data of individuals possess a severe threat to privacy and civil rights. To overcome these challenges, we need new strategies and security solutions to improve security operations , detecting and analysis of threats and attacks of security. It is important to improve the techniques with the existing cyber security threats. The amount of data is increasing in the internet at an enormous rate. As the volume of data is high , cyber attacks will also increase in an exponential rate . So it is very necessary to interpret and visualize it, inorder to identify certain threats and attacking patterns. Big data analytics helps in tracking large set of user activities inorder to avoid various threats associated with it. This helps to avoid many data breaches**

In this paper we analyse the importance of big data and highlights how is it used as a tool in cyber security to support some security activities. Here we summarize a detailed review on Big Data Analytics in Cyber security field.

Keywords— Bigdata, Bigdata Analytics, Big data query Data analytics, Privacy, threat, Bigdata security, cyber security, visualisation , anomaly

INTRODUCTION

Over the past few years data is being produced rapidly from various application which had led to the production of enormous amounts of data also known as Big Data. Developments in internet services and other communication system in the past years have introduced the term big data, which involves amounts of data which is generated in different forms at a huge rate. The capability to process these bulk amount of data is through big data analytics. By utilizing data which are collected from networks, cloud, computers and other devices can help to detect system exposure and accordingly.

Big Data refers to a huge volume of data which incorporates both structured and unstructured data.

Big data is generally used to facilitate better customer services and provide increased security on customer data moreover using bigdata helps an organisation to make faster and highly informed business decisions.

Major concepts of Bigdata are generalised into 5.V's they are :

volume, velocity, variety ,veracity, value

- Volume is the amount of data that is been generated .
- Velocity can be defined as the rapidness of data from various orgins.
- Variety refers to the diversity or it can be defined as different types of available data such as structured, unstructured and semi-structured data.

- Veracity defines the accuracy of data. It is not at all about the quality of data but about its reliability.
- Value can be refered as the benefit that bigdata can provide.

Big data is defined as too complex to process and analyse. It enables many people to overcome the problems that are associated with small samples of data. There is a need to find new possibilities for accomadating these data because it is growing day after another in an exponential manner.

Two major modes of Big data analytics are : verification and identification.

In verification, data analyst already has an assumption about certain property of services that he wants to verify by means of data analytics

In identification, data analyst gather a large dataset potentially from multiple sources and tries to identify interesting facts hidden within the dataset

. Two key concepts of aggregation of datasets within the big data content must be defined –

The first is the aggregation of schematically identical datasets . For example joining the service access logs of two different online services that are saved on the same web server implementation. mostly wed to attain more information within an existing content.

Second type of aggregation is about linkage created from joining two datasets from disjunct contents, based on some key information shared in both datasets to be aggregated.

A key challenge of big data analytics consists in identifying linkage – a link can be a user email addresses, postal codes/combinations of IP addresses and timestamps

The identity of service plays a major role.

This linkage via user identify bear some very challenging pitfalls in the field of privacy.

Bigdata Analytics is the art of processing, storing, and gathering large data.

Big data mainly focuses on the detection of anomalies and attacks. It allows analysing structured and unstructured data like documents, images and videos which are used as digital evidence in computer forensic process.

When the count of the data increases , it is very difficult to secure it. Confidentiality is the most important side when we consider big data protection. Big data analytics is used as a tool for any data or all business, organization possibilities.

One of the important tools which improves the method processing is Hadoop . In this method they are managing the characteristics of huge volumes of enterprise data. Hadoop combination and revolution analytics giving gain advantages, to unmet the requirement of business for making of strategic decisions. Hadoop split and stores data in different devices and the copy of each dataset will be saved in each devices or in other words those enormous count of data are distributed into large data sets across hundreds of inexpensive servers with help of scalable storage platform are called Hadoop.

It is operated in parallel.

Cyber security is the process of protecting user's data from unauthorised access , attacks or damages. Cyber security has now gone beyond the traditional way . Big data has unfold new ways for cyber security sector.

Here is an overview of fields in cyber security where big data analytics can contribute :-

Forensic Analysis –

Forensic focuses on the analysis , preservation and interpretation of computer data. This field deals with a large dataset, we use various conceptual models for forensic analysis inorder to remove reduntant data .By applying visualisation technique we can reduce the time and improve the effectiveness to find suspicious files.

Big data solutions provide two essential approaches so that the analyst can make his search in abundant data easier. First one is an integrate information from different sources and second has customised visualisation tools.

Malware Detection -

we use bigdata for malware detection.

These are the methods for classifying , combining Bigdata analysis with machine learning, binary instrumentation and dynamic instruct flow analysis.

Security offence -

Security offence include cyber description threat hunting and attack detection.

Cyber desception-

Nowadays it is motivated to use artificial intelligence, game theory and big data to enhance cyber security strategies against attackers.

The main objective of the cyber description is to detect attacks.

Threat hunting-

It is an active defence searching .It is an iterative activity to check through hardware and detect threats in advance instead of waiting for attack alerts. By using big data solution , processing of large amount of information generated by logs can be handled .

Attack detection -

It is very important to detect attacks in the shortest time if possible. It will reduce the time between detection and attack response.

Even though big data enhances security, on the other hand Big data gives a great chance not only for the development of an organisation but also for cyber criminals because they have much more to achieve when they track such a huge volume of data.

EXISTING METHODS

There are various algorithms and analytics used to find out information. They are also applied based on the nature of the data. Some examples for this kind of algorithms are :

Apriori Algorithm and Naive Bayes Classifier Algorithm.

Aprori algorithm works on the principle of bringing frequent data variables, then extending them to larger as long as they are frequent in nature.

Naive Bayes Classifier Algorithm based on Bayes Theorm. It is a classification algorithm with assumptions of independence among predictors. This model is easy to build and work very well for large datasets.

Data mining is also an important process when it comes to big data analytics. It processes large, pre-existing data. It is used for find misure detection and also anomaly detection.

LITERATURE SURVEY

In the paper entitled ‘big data analytics technique in cyber security’ the authors mentioned what bigdata is and how it is useful for the development of an organisation.

Here the corresponding authors proposes the usage of Big Data Analytics for enterprise data which is the data generally shared by users of an organisation.

Their main objective is to access unstructured data from all extreme, and to convert processed data to structured form so that the process of accessing will be more easier. For the easier protection and storage of Big data many organization use tools like Hadoop which distribute and stores the huge data efficiently by using the method of parallel processing. This method is an efficient and best method for Big Data Analytics because it is less expensive since the datas are distributed to inexpensive servers and it is less time consuming.

Here big data is described in a way that it increases data processing efficiency. Here various authors enumerate the major differences between traditional and Bigdata Analytics. This technique is divided into Batch processing and stream processing. In this paper various authors mentioned the desire to build different platforms to store and analyse data. The process is partially enriched and partially illustrative .

In the paper entitled “Special Issue on Big Data Applications in Cyber Security and Threat Inteligence - Part 2” [by kim- Kwang Raymond Choo, Senior Member, IEEE, Mauro Conti, Senior Member, IEEE and Ali De. Dehghantanha , Senior Member, IEEE] focuses on big data applications and threat intelligence. They also shows various research topics on big data for future research which includes anomaly detection for big data, big forensic data provenance, analysis of big data for cyber intelligence, advanced persistent threats detection, big data analytical technique for cyber defence, big data forensic data management and reduction.

In the paper entitled ‘Special issue on Big data applications in Cyber Security and threat Intelligence part 1’ - [by Kim Kwang Raymond Choo, Senior Member, IEEE, Mauro Conti, Senior Member, IEEE, and Ali Dehghantanhe, Senior Member IEEE]focuses on importance of big data analytical techniques to overcome cyber security threats. They shows various technique to interpret, mine and visualise big data from different sources so that it can be applied in cyber forensic, cyber security and threat intelligence.

In the paper entitled "Challenges of Privacy protection in Big Data Analytics"[by Merko Jensen.] Shows various challenges related to big data analytics on privacy .He proposed that data erosion in terms of privacy and user's rights may due to the upcoming trend in big data analytics. He proposed various fields of research on privacy in big data analytics. The most challenging part of privacy in big data analytics is that to provide transparency of personal data of the individuals with respect to type of processing . It is always necessary to process information bound to an individual. Informed consent means that there are many types of big analytics based on complex data algorithm, so each Individual must be given an explanation of all these algorithms so that they can understand what is happening there ,this is a big challenge to data analysis.

An individual decides to revoke the consent for processing personal data later. This is similar to getting a person used among various data collectors and data analysts that is not easier to stop processing on these datas and to delete it. This has become a highly challenging issue. There are various types of attacks such as targeted identification attacks, correlation attacks and arbitary identification attacks. Most threatening type of attack is targeted identification attack. It is to identify some more details of an individual. Inorder to create more unique database entries we link a dataset of uniform data values to other sources . Correlation attacks consist of this kind of linking form datasets. There datasets contain more information per User ID. This helps in analysing more on individual.

Arbitary identification attacks shows failures of a set of anonymized data. This type of attack link atleast to one entry of the dataset to identify a human individual.

A threat to big data analytics is if the information gathered is valid or not. Various types of results can be formed. It will depend on the type of query used by a big data analyst.

Results from different big data query sometimes become a completely wrong final statement. A lot of threats to privacy can also arise from economic consideration in such data trading economic issues of the big data, paradigm is considered to be the fourth category of threats. So threats can be caused due to intentional attacks. It can also caused due to false data processing methodology or caused by interaction with concerned individuals. So field of privacy in big data faces a lot of challenges.

In the paper entitiled ‘Big data and analytics ‘ the authors enumerate about the rapid growth of data. Contribution of smart devices, such as smartphones hand held computers, wireless networks and social media generating more data over past few years.

In social media domains such as facebook, more than 30 million users are updating posting and sharing their images and video per minute .

Like in instagram , also 300 million instagram users share more than 60-million photos everyday.

More than 100 hours of video are uploaded in every minute. This huge enormous data is Big Data and there is a need to protect and secure these data from & unauthorized access.

This Big data allows new possibilities in technolgy as well as in research field.

In the paper entitiled ‘ Big data analytics for cyber security ‘ explains about the spontaneous growth of the internet has resulted in the exponential increase of the number of cyber attacks. Many organisations tried many popular cyber security to prevent these attacks. Also, the intoduction of Big Data made internet with enormous amount of data . To regale this issue, many researches are now focusing on Security Analytics, which is one of the important application of Big Data Analytics techniques to cybersecurity. This paper provides a survey on the art of Security Analytics which including its states such as its description, , trends, technology and tools.

In the paper entitled "Challenges of Privacy protection Big Data Analytics" presented challenges to privacy of Individuals. The paper discusses about various set of challenges that may threaten privacy of individuals. Another threat with respect to privacy in big data analytics is the ability to perform "re-identification attacks", also validity of the result gathered is also a threat. Another threat covers the economic issues of big data paradigm.

In the paper entitled "Research about New Media Security Technology base on Big Data Era” [by Zheng-wu Lu, Communication University of China, Beying] proposed that high-precision, robust, lightweight and identification and understanding of technology is very important.

It will be the direction of future research. Big data based on cloud computing technology will become a major trend. Difficulty of the new media big is because recognition and understanding of new media content is difficult.

To create a healthy innovative new media environment , we need to research how we can safely provide, consume data and dig information faithfully from these datas.

In the paper entitled “An Insight into Big Data Analytics - Methods and Application

[by Dr. Manjula Sanjay and Alamma 13. H Department of Master of Computer Applications, Dayananda Sagar Academy of Technology of Management, Banglore, India] shows that generation of analytical software like Hadoop or other analytical database can be done through commodity hardware. They shows how traditional data analytics differ from big data analytics now They described about three methods of data analytics and various applications of big data on business, social and scientific applications.

In the paper entitled "Security. Analytics : Big Data Analytics for Cyber security"[by Dr..Tariq Muhammed and Uzma Afzal] proposed that malicious and suspicious patterns can be identified by network managers particularly in the surveillance of real-time network streams. They shows the survey on the art of security Analytics. Also the authors proposed that cyber application of analytics will become an imminent part in cybersecurity in the future. They mentioned different types of big data sources for analytics solution.

In the paper entitled.,”Big Data Aanlytics Techniques A survey" by [Poonam Vashist and Vishal Gupta] proposed that big data consist of structured, semi-structured and unstructured data. They shows the methods. to analyse the audio, video and text. They shows different challenges faced by researches while performing big data analysis They also discussed various big data analytics methods and techniques.

CONCLUSION

This paper contains a detailed review on Big Data Analytics in Cyber Security sector . Big data is a new alternative to improve security operations. It has the ability process voluminous data in different format in short time. It is applied to monitor operations and detection of anomalies. Moreover it is used in protective strategies such as threat hunting on cyber deception. It can also detect attack patterns by processing immense data from heterogeneous source.

Big Analytics is often used in cyber security lots of reasons. It facilitate the working of an organization more easier by increasing security with the use of various algorithms and techniques.

The main objective of Big Data analytics is to generate a safe environment for users to protect their data from unauthorised access attacks.

REFERENCE

1] Kim -Kwang Raymond C+ hoo ,Mauro Conti, Ali Dehghantanha

“special issue on Big Data Application in Cyber Security and threat intelligence – part 1”

IEEE transaction on Big Data , July – September (2019)

2] Kim -Kwang Raymond Choo, Mauro Conti, Ali Dehghantanha

“ Special Issue on Big Data Application in Cyber Security and threat intelligence – part 2”

IEEE Transaction on Big Data , October – December (2019)

3] Fontugne R Mazel I and Fuhada K. Hashdoop "A MapReduce framework for network anomaly detection “IEEE conference on work shops (2014)4] Meiko Jensen "Challenges of Privacy Protection in Big Data Analytics”

IEEE International Congress on Big Data (2013)

5] Aviral Apurva, Pranshu Ranakoti, Saurav Yadav, Shashank Tomer, Nihar Ranjan Roy

“Redefining Cyber Security with Big Data Analytics" (2017) International Conference on Computing and communication technologies for Smart Nation (I c3TSN).

6] Poonam Vashisht , Vishal Gupta

“ Big Data Analytics Techniques: A survey (2015) International conference on Green Computing and Internet of things (ICGI0T)

7] Dr. Tariq Muhammed, Uzma Afzal

" Security Analytics Big Data Analytics for Cyber Security (2013) 2nd National Conference on Information Assurance (NCIA)

8] Zheng - Wu Lu "Research about New Media Security Technology bare on Big Data Era”

(2016) IEEE 14th Inernational Conference on Dependable/ Automatic and Secure Computing, 14th international conference on Pervasive Intelligence and computing, 2nd international conference on Big Data Intelligence and computing cyber Security and Technology Congress

9] Danda B Rawat ”Cyber Security in Big Data era:

From securing “ Big Data to Data Driven Security”

IEEE

10] Neha Srivasta , prof. Umesh Chandra Jaiswal

“ Big Data Analytics Technique in Cyber Security-

A Review” proceedings of third international conference on Computing Methodolgies and Communication (ICCMC 2019)

ISSN 2394-3777 (Print)
ISSN 2394-3785 (Online)
Available online at www.ijartet.com

International Journal of Advanced Research Trends in Engineering and Technology (IJARTET)
Vol. 8, Special Issue 1, August 2021

Big data Analytics in Cyber Security

PRATIBHA ANN JAYESH [1] , MERLIN MATHEW [2], ASHLY MATHEW [3]

SCHOLAR, BCA DEPARTMENT ,SAINTGITS COLLEGE OF APPLIED SCIENCES , PATHAMUTTOM , KOTTAYAM, INDIA [1]

SCHOLAR, BCA DEPARTMENT ,SAINTGITS COLLEGE OF APPLIED SCIENCES , PATHAMUTTOM , KOTTAYAM, INDIA [2]

SCHOLAR, BCA DEPARTMENT ,SAINTGITS COLLEGE OF APPLIED SCIENCES , PATHAMUTTOM , KOTTAYAM, INDIA [3]

***Abstract* — Linkage of personal data of individuals possess a severe threat to privacy and civil rights. To overcome these challenges, we need new strategies and security solutions to improve security operations , detecting and analysis of threats and attacks of security. It is important to improve the techniques with the existing cyber security threats. The amount of data is increasing in the internet at an enormous rate. As the volume of data is high , cyber attacks will also increase in an exponential rate . So it is very necessary to interpret and visualize it, inorder to identify certain threats and attacking patterns. Big data analytics helps in tracking large set of user activities inorder to avoid various threats associated with it. This helps to avoid many data breaches**

In this paper we analyse the importance of big data and highlights how is it used as a tool in cyber security to support some security activities. Here we summarize a detailed review on Big Data Analytics in Cyber security field.

Keywords— Bigdata, Bigdata Analytics, Big data query Data analytics, Privacy, threat, Bigdata security, cyber security, visualisation , anomaly

INTRODUCTION

Over the past few years data is being produced rapidly from various application which had led to the production of enormous amounts of data also known as Big Data. Developments in internet services and other communication system in the past years have introduced the term big data, which involves amounts of data which is generated in different forms at a huge rate. The capability to process these bulk amount of data is through big data analytics. By utilizing data which are collected from networks, cloud, computers and other devices can help to detect system exposure and accordingly.

Big Data refers to a huge volume of data which incorporates both structured and unstructured data.

Big data is generally used to facilitate better customer services and provide increased security on customer data moreover using bigdata helps an organisation to make faster and highly informed business decisions.

Major concepts of Bigdata are generalised into 5.V's they are :

volume, velocity, variety ,veracity, value

- Volume is the amount of data that is been generated .
- Velocity can be defined as the rapidness of data from various orgins.
- Variety refers to the diversity or it can be defined as different types of available data such as structured, unstructured and semi-structured data.

- Veracity defines the accuracy of data. It is not at all about the quality of data but about its reliability.
- Value can be refered as the benefit that bigdata can provide.

Big data is defined as too complex to process and analyse. It enables many people to overcome the problems that are associated with small samples of data. There is a need to find new possibilities for accomadating these data because it is growing day after another in an exponential manner.

Two major modes of Big data analytics are : verification and identification.

In verification, data analyst already has an assumption about certain property of services that he wants to verify by means of data analytics

In identification, data analyst gather a large dataset potentially from multiple sources and tries to identify interesting facts hidden within the dataset

. Two key concepts of aggregation of datasets within the big data content must be defined –

The first is the aggregation of schematically identical datasets . For example joining the service access logs of two different online services that are saved on the same web server implementation. mostly wed to attain more information within an existing content.

Second type of aggregation is about linkage created from joining two datasets from disjunct contents, based on some key information shared in both datasets to be aggregated.

A key challenge of big data analytics consists in identifying linkage – a link can be a user email addresses, postal codes/combinations of IP addresses and timestamps

The identity of service plays a major role.

This linkage via user identify bear some very challenging pitfalls in the field of privacy.

Bigdata Analytics is the art of processing, storing, and gathering large data.

Big data mainly focuses on the detection of anomalies and attacks. It allows analysing structured and unstructured data like documents, images and videos which are used as digital evidence in computer forensic process.

When the count of the data increases , it is very difficult to secure it. Confidentiality is the most important side when we consider big data protection. Big data analytics is used as a tool for any data or all business, organization possibilities.

One of the important tools which improves the method processing is Hadoop . In this method they are managing the characteristics of huge volumes of enterprise data. Hadoop combination and revolution analytics giving gain advantages, to unmet the requirement of business for making of strategic decisions. Hadoop split and stores data in different devices and the copy of each dataset will be saved in each devices or in other words those enormous count of data are distributed into large data sets across hundreds of inexpensive servers with help of scalable storage platform are called Hadoop.

It is operated in parallel.

Cyber security is the process of protecting user's data from unauthorised access , attacks or damages. Cyber security has now gone beyond the traditional way . Big data has unfold new ways for cyber security sector.

Here is an overview of fields in cyber security where big data analytics can contribute :-

Forensic Analysis –

Forensic focuses on the analysis , preservation and interpretation of computer data. This field deals with a large dataset, we use various conceptual models for forensic analysis inorder to remove reduntant data .By applying visualisation technique we can reduce the time and improve the effectiveness to find suspicious files.

Big data solutions provide two essential approaches so that the analyst can make his search in abundant data easier. First one is an integrate information from different sources and second has customised visualisation tools.

Malware Detection -

we use bigdata for malware detection.

These are the methods for classifying , combining Bigdata analysis with machine learning, binary instrumentation and dynamic instruct flow analysis.

Security offence -

Security offence include cyber description threat hunting and attack detection.

Cyber desception-

Nowadays it is motivated to use artificial intelligence, game theory and big data to enhance cyber security strategies against attackers.

The main objective of the cyber description is to detect attacks.

Threat hunting-

It is an active defence searching .It is an iterative activity to check through hardware and detect threats in advance instead of waiting for attack alerts. By using big data solution , processing of large amount of information generated by logs can be handled .

Attack detection -

It is very important to detect attacks in the shortest time if possible. It will reduce the time between detection and attack response.

Even though big data enhances security, on the other hand Big data gives a great chance not only for the development of an organisation but also for cyber criminals because they have much more to achieve when they track such a huge volume of data.

EXISTING METHODS

There are various algorithms and analytics used to find out information. They are also applied based on the nature of the data. Some examples for this kind of algorithms are : Apriori Algorithm and Naive Bayes Classifier Algorithm.

Aprori algorithm works on the principle of bringing frequent data variables, then extending them to larger as long as they are frequent in nature.

Naive Bayes Classifier Algorithm based on Bayes Theorm. It is a classification algorithm with assumptions of independence among predictors. This model is easy to build and work very well for large datasets.

Data mining is also an important process when it comes to big data analytics. It processes large, pre-existing data. It is used for find misure detection and also anomaly detection.

LITERATURE SURVEY

In the paper entitled ‘big data analytics technique in cyber security’ the authors mentioned what bigdata is and how it is useful for the development of an organisation.

Here the corresponding authors proposes the usage of Big Data Analytics for enterprise data which is the data generally shared by users of an organisation.

Their main objective is to access unstructured data from all extreme, and to convert processed data to structured form so that the process of accessing will be more easier. For the easier protection and storage of Big data many organization use tools like Hadoop which distribute and stores the huge data efficiently by using the method of parallel processing. This method is an efficient and best method for Big Data Analytics because it is less expensive since the datas are distributed to inexpensive servers and it is less time consuming.

Here big data is described in a way that it increases data processing efficiency. Here various authors enumerate the major differences between traditional and Bigdata Analytics. This technique is divided into Batch processing and stream processing. In this paper various authors mentioned the desire to build different platforms to store and analyse data. The process is partially enriched and partially illustrative .

In the paper entitled “Special Issue on Big Data Applications in Cyber Security and Threat Inteligence - Part 2” [by kim- Kwang Raymond Choo, Senior Member, IEEE, Mauro Conti, Senior Member, IEEE and Ali De. Dehghantanha , Senior Member, IEEE] focuses on big data applications and threat intelligence. They also shows various research topics on big data for future research which includes anomaly detection for big data, big forensic data provenance, analysis of big data for cyber intelligence, advanced persistent threats detection, big data analytical technique for cyber defence, big data forensic data management and reduction.

In the paper entitled ‘Special issue on Big data applications in Cyber Security and threat Intelligence part 1’ - [by Kim Kwang Raymond Choo, Senior Member, IEEE, Mauro Conti, Senior Member, IEEE, and Ali Dehghantanhe, Senior Member IEEE]focuses on importance of big data analytical techniques to overcome cyber security threats. They shows various technique to interpret, mine and visualise big data from different sources so that it can be applied in cyber forensic, cyber security and threat intelligence.

In the paper entitled "Challenges of Privacy protection in Big Data Analytics"[by Merko Jensen.] Shows various challenges related to big data analytics on privacy .He proposed that data erosion in terms of privacy and user's rights may due to the upcoming trend in big data analytics. He proposed various fields of research on privacy in big data analytics. The most challenging part of privacy in big data analytics is that to provide transparency of personal data of the individuals with respect to type of processing . It is always necessary to process information bound to an individual. Informed consent means that there are many types of big analytics based on complex data algorithm, so each Individual must be given an explanation of all these algorithms so that they can understand what is happening there ,this is a big challenge to data analysis.

An individual decides to revoke the consent for processing personal data later. This is similar to getting a person used among various data collectors and data analysts that is not easier to stop processing on these datas and to delete it. This has become a highly challenging issue. There are various types of attacks such as targeted identification attacks, correlation attacks and arbitary identification attacks. Most threatening type of attack is targeted identification attack. It is to identify some more details of an individual. Inorder to create more unique database entries we link a dataset of uniform data values to other sources . Correlation attacks consist of this kind of linking form datasets. There datasets contain more information per User ID. This helps in analysing more on individual.

Arbitary identification attacks shows failures of a set of anonymized data. This type of attack link atleast to one entry of the dataset to identify a human individual.

A threat to big data analytics is if the information gathered is valid or not. Various types of results can be formed. It will depend on the type of query used by a big data analyst.

Results from different big data query sometimes become a completely wrong final statement. A lot of threats to privacy can also arise from economic consideration in such data trading economic issues of the big data, paradigm is considered to be the fourth category of threats. So threats can be caused due to intentional attacks. It can also caused due to false data processing methodology or caused by interaction with concerned individuals. So field of privacy in big data faces a lot of challenges.

In the paper entitiled ‘Big data and analytics ‘ the authors enumerate about the rapid growth of data. Contribution of smart devices, such as smartphones hand held computers, wireless networks and social media generating more data over past few years.

In social media domains such as facebook, more than 30 million users are updating posting and sharing their images and video per minute .

Like in instagram , also 300 million instagram users share more than 60-million photos everyday.

More than 100 hours of video are uploaded in every minute. This huge enormous data is Big Data and there is a need to protect and secure these data from & unauthorized access.

This Big data allows new possibilities in technolgy as well as in research field.

In the paper entitiled ‘ Big data analytics for cyber security ‘ explains about the spontaneous growth of the internet has resulted in the exponential increase of the number of cyber attacks. Many organisations tried many popular cyber security to prevent these attacks. Also, the intoduction of Big Data made internet with enormous amount of data . To regale this issue, many researches are now focusing on Security Analytics, which is one of the important application of Big Data Analytics techniques to cybersecurity. This paper provides a survey on the art of Security Analytics which including its states such as its description, , trends, technology and tools.

In the paper entitled "Challenges of Privacy protection Big Data Analytics" presented challenges to privacy of Individuals. The paper discusses about various set of challenges that may threaten privacy of individuals. Another threat with respect to privacy in big data analytics is the ability to perform "re-identification attacks", also validity of the result gathered is also a threat. Another threat covers the economic issues of big data paradigm.

In the paper entitled "Research about New Media Security Technology base on Big Data Era” [by Zheng-wu Lu, Communication University of China, Beying] proposed that high-precision, robust, lightweight and identification and understanding of technology is very important.

It will be the direction of future research. Big data based on cloud computing technology will become a major trend. Difficulty of the new media big is because recognition and understanding of new media content is difficult.

To create a healthy innovative new media environment , we need to research how we can safely provide, consume data and dig information faithfully from these datas.

In the paper entitled “An Insight into Big Data Analytics - Methods and Application

[by Dr. Manjula Sanjay and Alamma 13. H Department of Master of Computer Applications, Dayananda Sagar Academy of Technology of Management, Banglore, India] shows that generation of analytical software like Hadoop or other analytical database can be done through commodity hardware. They shows how traditional data analytics differ from big data analytics now They described about three methods of data analytics and various applications of big data on business, social and scientific applications.

In the paper entitled "Security. Analytics : Big Data Analytics for Cyber security"[by Dr..Tariq Muhammed and Uzma Afzal] proposed that malicious and suspicious patterns can be identified by network managers particularly in the surveillance of real-time network streams. They shows the survey on the art of security Analytics. Also the authors proposed that cyber application of analytics will become an imminent part in cybersecurity in the future. They mentioned different types of big data sources for analytics solution.

In the paper entitled.,”'Big Data Aanlytics Techniques A survey" by [Poonam Vashist and Vishal Gupta] proposed that big data consist of structured, semi-structured and unstructured data. They shows the methods. to analyse the audio, video and text. They shows different challenges faced by researches while performing big data analysis They also discussed various big data analytics methods and techniques.

CONCLUSION

This paper contains a detailed review on Big Data Analytics in Cyber Security sector . Big data is a new alternative to improve security operations. It has the ability process voluminous data in different format in short time. It is applied to monitor operations and detection of anomalies. Moreover it is used in protective strategies such as threat hunting on cyber deception. It can also detect attack patterns by processing immense data from heterogeneous source.

Big Analytics is often used in cyber security lots of reasons. It facilitate the working of an organization more easier by increasing security with the use of various algorithms and techniques.

The main objective of Big Data analytics is to generate a safe environment for users to protect their data from unauthorised access attacks.

REFERENCE

1] Kim -Kwang Raymond C+ hoo ,Mauro Conti, Ali Dehghantanha

“special issue on Big Data Application in Cyber Security and threat intelligence – part 1”

IEEE transaction on Big Data , July – September (2019)

2] Kim -Kwang Raymond Choo, Mauro Conti, Ali Dehghantanha

“ Special Issue on Big Data Application in Cyber Security and threat intelligence – part 2”

IEEE Transaction on Big Data , October – December (2019)

3] Fontugne R Mazel I and Fuhada K. Hashdoop "A MapReduce framework for network anomaly detection “IEEE conference on work shops (2014)4] Meiko Jensen "Challenges of Privacy Protection in Big Data Analytics”

IEEE International Congress on Big Data (2013)

5] Aviral Apurva, Pranshu Ranakoti, Saurav Yadav, Shashank Tomer, Nihar Ranjan Roy

“Redefining Cyber Security with Big Data Analytics" (2017) International Conference on Computing and communication technologies for Smart Nation (I c3TSN).

6] Poonam Vashisht , Vishal Gupta

“ Big Data Analytics Techniques: A survey (2015) International conference on Green Computing and Internet of things (ICGI0T)

7] Dr. Tariq Muhammed, Uzma Afzal

" Security Analytics Big Data Analytics for Cyber Security (2013) 2nd National Conference on Information Assurance (NCIA)

8] Zheng - Wu Lu "Research about New Media Security Technology bare on Big Data Era”

(2016) IEEE 14th Inernational Conference on Dependable/ Automatic and Secure Computing, 14th international conference on Pervasive Intelligence and computing, 2nd international conference on Big Data Intelligence and computing cyber Security and Technology Congress

9] Danda B Rawat ”Cyber Security in Big Data era:

From securing “ Big Data to Data Driven Security”

IEEE

10] Neha Srivasta , prof. Umesh Chandra Jaiswal

“ Big Data Analytics Technique in Cyber Security-

A Review” proceedings of third international conference on Computing Methodolgies and Communication (ICCMC 2019)