A Project Report on

AUTO-DETECTION OF BRAIN TUMOURS FROM MRI IMAGES USING DEEP LEARNING ALGORITHMS

Submitted in partial fulfilment for award of degree of

MBA

In **Business Analytics**

Submitted by

Satyajit Pal

R18MBA04

Under the Guidance of

Dr. JB Simha

Chief Technology Officer, ABIBA Systems

REVA Academy for Corporate Excellence

REVA University

Rukmini Knowledge Park, Kattigenahalli,

Yelahanka, Bangalore – 560064

September 2020

REVA UNIVERSITY
Bengaluru, India

Candidate's Declaration

I, Satyajit Pal hereby declare that I have completed the project work towards the Master of Business Administration in Business Analytics at, REVA University on the topic entitled Auto-Detection of Brain Tumour from MRI Images Using Deep Learning under the supervision of Dr. JB Simha, Chief Technology Officer, ABIBA Systems. This paper reflects a partial fulfilment of the graduation criteria for the academic year 2020 from my original thesis.

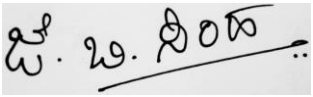
Place: Bengaluru
Date: 05/09/2020

Name of the Student: Satyajit Pal
Signature of Student

REVA UNIVERSITY
Bengaluru, India

Certificate

This is to Certify that the PROJECT work entitled Auto-Detection of Brain Tummour from MRI Images Using Deep Learning carried out by Satyajit Pal with SRN R18MBA04, is a bonafide student of REVA University, is submitting the project report in fulfilment for the award of MBA in Business Analytics during the academic year 2019-2020. The project summary was reviewed for plagiarism for fewer than 15% of the plagiarism examination. The project report has been approved as it satisfies the academic requirements in respect of PROJECT work prescribed for the said Degree.



Dr. J.B. Simha
Guide

<Signature of the Director>
Dr. Shinu
Director

External Viva (Virtual)
Names of the Examiners

1. Ravi Shukla, Sr. Advisor and Data Scientist, Dell
2. Krishna Kumar Tiwari, Senior Data Scientist, CoE, AI/ML, Jio

Place: Bengaluru
Date: 19/09/2020

Acknowledgement

First of all, I would like to thank my project guide, Dr. J.B. Simha., Chief Technology Officer, ABIBA Systems, for providing me his time and knowledge during the building and completion of this project. I would also like to thank Chancellor, Dr. P Shayma Raju, Ex- Vice Chancellor, Dr. S.Y Kulkarni, Dr. K. Mallikharjuna Babu, Vice Chancellor and Dr. M. Dhanamjaya, Registrar, Dr. Shinu, Program Director for giving me this opportunity to implement the idea and learning into shaping this project. Next, I would take the opportunity to pass my gratitude to my educational department, Reva Academy of Corporate Excellence, REVA University campus, for providing ample opportunities and guidance.

Similarity Index Report

Title of the Thesis: AUTO-DETECTION OF BRAIN TUMOURS FROM MRI IMAGES USING DEEP LEARNING ALGORITHMS

Total No. of Pages: 43

Name of the Student: Satyajit Pal

Name of the Guide(s): Dr. J.B. Simha

This is to certify that the above thesis was scanned for similarity detection. Process and outcome are given below.

Software Used: Turnitin

Date of Report Generation: 19/09/2020
Similarity Index in %: 6%
Total word count: 8199

Place: Bengaluru
Date: 06/10/2020

Name of the Student: Satyajit Pal
Signature of Student

List of Abbreviations

Sl. No	Abbreviation	Long Form
1	MRI	Magnetic Resonance Imaging
2	CBTRUS	Central Brain Tummour Registry of the United States
3	GBM	Glioblastoma multiform
4	CSF	Cerebrospinal Fluid
5	WHO	World Health Organisation
6	CT	Computed Tomography
7	DTI	Diffusion Tensor Imaging
8	BOLD	Blood oxygen level based
9	MAP	Mean Average Precision
10	DNN	deep neural networks
11	CNN	Convolutional Neural Network
12	CRISPDM	CrossIndustry ProcessFor Datamining
13	BraTS	Brain Tummour Segmentations
14	BCN	Baseline Convolutional Network
15	FCN	Fully Convolutional Network
16	FIFCN	Fullimage FullyConvolutional Network

List of Figures

List of Tables

Abstract

Cancer is a complex disorder induced by a mixture of the causes of origin, the climate and the lifestyle. The mixture of several genes that cause progression of cancer differs greatly between the cancer types and patients. The diagnosis of cancer is needed early on to treat the patient properly and minimise the risk of cancer death so cancer cells are more likely to survive later and to become more likely to die. About half a million people die per year of brain tumour. The second most frequent cancer diagnosis and second most significant cause of cancer mortality in men and women is brain tumour. Brain Tumour The main tool for diagnosis and staging of brain tumour patients was magnetic resonance imaging (MRI). Brain tumour MRI images are used to measure the region and mean tumour area and tumour size to other regions.

Magnetic resonance imaging (MRI) for radiation therapy preparation is the most preferred medical imaging tool for primary brain tumour diagnostics. The brain tumours regions are typically identified manually by the oncologist or radiologist from the volumetric MRI results. In general, medical imaging is messy in comparison to natural visualisation, since the structure of cancer regions can differ slice by slice. The automated segmentation and measurement of the brain tube is therefore a extremely difficult task, not only because of its scale, but also because of its somewhat irregular structure and strength distribution behaviour. Machine vision and Deep learning resources for creation and study of this project provide adequate care for the patient and reduce the risk of death, and how auto-detection and segmentation of early stage brain tumours / cancerous cells / regions.

A new CNN architecture, different from the current architecture for computer vision. The CNN has powerful qualitative features both local and national. Networks also use the final layer that coevolves a previously attached node, which makes 40 times the bandwidth. safe to the traditional CNN applications. A two-phase training protocol to resolve the mismatch of tumour labelling is identified. A 2-phase training technique is given. Finally, the architecture of the cascade is discussed to use the performance of a key CNN as an alternate knowledge source for the next CNN. Returns from the BRATS data collection demonstrate that the system is up to 30 times faster than the recently reported state of the art.

Keywords: Image Segmentation; Tensroflow; Python; Keras; MRI; Cancer; Brain Tumour; Diffusion-weighted imaging; BRATS Dataset, Deep Learning, CNN

Contents

Chapter 1: Introduction

Brain tumours are a leading cause of cancer-related killing in the US, based on data supplied by the US National Brain-Tummours Registry (CBTRUS). In the case of children under 20 years and in men aged between 20 and 39, leukaemia is the second greatest cause of cancer mortality.(Ain et al., 2010). In the United States, there are about 70,000 new cases of brain tumour in adults per year; 14,000 cases have been identified in malignant (24,000) or nonmalignant (45,000), with 4,300 new diagnoses arising in the case of infants, and over half have been identified in children under 15. The prevalence of occurrence for primary brain and CNS tumours is 221,8 (61,9 malignant and 177,3 non-malignant) for every 100,000 people. The gross average mortality rate in the US between 2007 and 2011 was 69.789 due to primary malignant brain and CNS tumours. (Akselrod-Ballin et al., 2006). For people aged 20 and 44, 58.5 percent decline following a malignant brain tumour diagnosis in the age group of adults aged 45 and 54. These numbers consider only primary tumours in the brain, which begin and reside in the brain.

The most prevalent primary brain tumour is meningioma with 34 percent, but it comprises 80 percent of malignant tumours and is the most common primary tumour in the brain. It involves glioma that arises from adhesive or supporting brain tissue (30 percent of all brain tumour). Glioblastoma multiform (GBM) is the most widespread and violent glioma that accounts for 54% of all gliomas. This kind of tumour has high infiltrated development and exceptionally low prognosis with an estimated one-year post diagnostic survival time [116]. Extensive therapies such as chemotherapy and radiotherapy and surgical resection impact the survival period (Ambrosini et al., 2010). This thesis concentrates in particular on the controlled treatment of quantities in low and large grades of the most recurrent malignant tumour, glioma.

The assessment of images collected is typically conducted manually in standard clinical procedures, based on objective parameters or measurements such as the maximum visible axial-slice diameter. There is also a huge opportunity for testing and treatment preparation of very detailed approaches for systematically interpreting brain tumour scans. Menze et al. however found that even manual annotations of professional raters revealed substantial differences in areas in which strength gradients of bias or volume effects are smooth or blurred between tumour structures and underlying tissue. In addition, only relative strength differences in healthy tissues are known for brain tumour lesions, and for each case, their form, size and positioning make it difficult for standard pattern recognition algorithms to be used. The number of data collected often rises as a result of the growing number of patients(Bauer et al., 2013). As a consequence, algorithms that can process the data automatically become increasingly important, which is the principal inspiration for this work.

The main method of screening and therapy for gliomas is multimodality magnetic resonance imaging (Darmawan, 2019). Accurate tuber segmentation is important for the diagnosis

and the development of a treatment plan to identify features such as length, spread and position. Radiologists are currently segmenting tumour areas manually, but advancement in computer vision has allowed the segmenting process to be automated(Cates et al., 2005). In particular, neural convolution (CNNs) segmentation algorithms were seen to be at least as successful as other automatic tumour segmentation processes. This is a modern approach focused on deep neural networks to the segmentation of gliomas (Capelle et al., 2000). We have two CNN architectures for tumour and non-tummour areas in the patch-specific binary classification as well as a complete CNN architecture. Both BraTS architectures would then be learned and checked and passed to the Rembrandt dataset explored. We often search at many approaches to avoid model duplication and boost robustness due to the comparatively limited scale of the data sets involved.

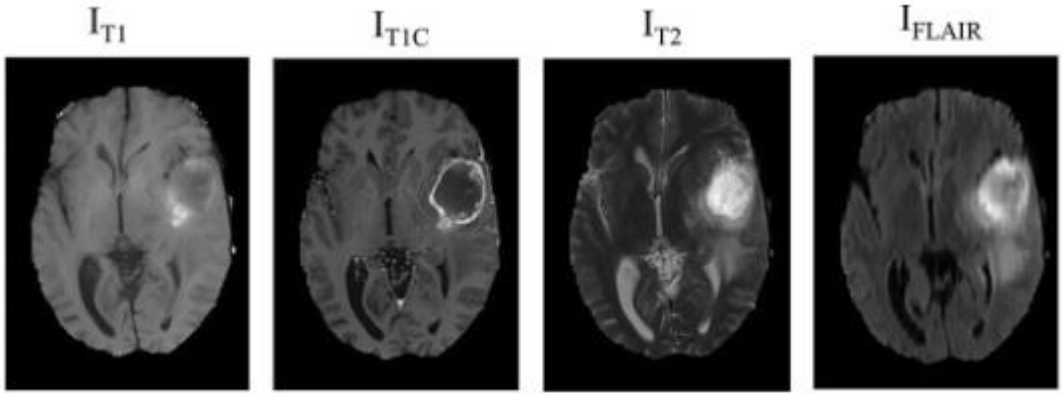


Fig 1: MR Image Types

In this paper, a brief summary of previous studies on the segmentation and conversion of biomedical images is presented and then suggest and analyses the tumour segmentation model architectures (Prastawa et al., 2004). Finally, the findings for learning transfer between data sets for neuroimaging is presented.

Chapter 2: Literature Review

S. Bauer, L. P. Nolte, and M. Reyes,2011, a complete automated analysis of brain tumour photographs by using a vector support computer classification in conjunction with a random field hierarchy.

Umit Ilhan et al., 2017, New threshold technology proposed for better tumour detection accuracy. The fact that this technique is not an automated method is one drawback.

B. Devkota et al., , Suggested to use MRI images to diagnose the brain-tummour for medial filtern pre-processing and C-means Blurry for segmentation in its initial stage. No research or analysis of the suggested approach has been carried out up to assessment.

Amruta Pramod Hebli et al., 2016, This paper shows that machine learning has a crucial role to play in the diagnosis of brain tumours and can be segmented according to the right process. In this paper GA and PSO have been contrasted for segmentation with K-means and fuzzy-c approaches.

Tian Xia et al., , Suggested the automated segmentation of the MRI brain with tumour and classification process. The Ostu threshold and morphological operations are eliminated with 86 percent precision in this tumour area.

Aby Elsa Babu et al., , It has suggested an image-based bilateral processing approach that has been proven to be a successful tool for tumour detection in the brain.

Luxit Kapoor et al., 2017, In this article, he proposed several steps in tumour identification and proved that segmentation is the most essential and conducive.

Miss. Shrutika Santosh et al, 2017, The brain tumour detection system has been suggested and Sobel edge detection operator has identified the borders of the tumour. This paper shows the tumour level and scale that MRI images are ideally suited for diagnosis of the brain tumour.

Geert Litjens et al., 2017, This paper explored the use of deep learning for identification , recognition of objects, segmentation, classification of images and other tasks.

Olfa Limam et al., 2016, The multi-target fuzzy clustering method suggested that produces a range of Pareto solutions dependent on the validity measure of the I and chosen as the final clustering solution. -This approach is used for the 97.5% accuracy of CT images.

H. B. Nandpuru et al., 2013, Provide the method for detection of brain tumours. Using a controlled machine learning technique, SVM brains are graded as normal and cancerous. The first was to strip the colour, grey and symmetric elements. The classifier suggested provides 84% precision.

El-Dahshan et al., 2002, A device entry is the brain MRI; characteristics were derived by the discrete transform of the wavelet, reduced by the main component analysis and graded with artificial neural network (FF- BPNN) and KNN feed back propagation. The accuracy of these classifiers for teaching and analysing data sets is 99%.

Hong Men, et al., 2010, Present two machine learning neural network algorithms as well as SVM for brain MRI classification. They used two types of support vector machine for various parametric values based on the polynomial kernel and radial base function. The result is that the vector support approach is equivalent to the neural network algorithm.

Nilesh Bhaskarrao Bahadure et al. 2017, This paper suggests a novel method that can help diagnose tumours in the brain easily, accurately and in time and provide an appropriate position for the tumour to be formed.

G Rajesh Chandra et al.2016 The method developed uses the naturally inspired GA algorithm which helps overcome problems with optimization through a broad search field.

P.Shanthakumar et al.2015, Offer a way to equate brain anomalies transparently with normal brain tissue. Results of tumour decomposition was estimated on the basis of indexed phenotype, variance fraction and optimistic prediction values in which 0.817% was 0.817%, 0.812%.

Munmun Saha et al.,, This paper analysed and outlined several emerging segmentation approaches for brain tumour diagnosis using MRI images.

Alexander Zotin et al., , The paper introduces the identification of the brain tumour edge using FCM clusters based MRI images. Biomedical image.

Iván Cabria et al. 2017, A segmentation technique named PFS, demonstrating that the output is as strong as other segmentation techniques, was suggested.

Amin et al., The design of DNN has been clarified and evaluated on eight challenges of the data sets and 5 MRI modalities, including flair, DWI, T1,T2 and T1, which together have an average time of 5,502 seconds for CNN models.

Chapter 3: Problem Statement

In medicine, radiologists perform most of the classifications. They face various kinds of challenges. These technological forms are not feasible due to the fact that vast quantities of data cause uncertainty. Photos can involve multiple noise types that can result in incorrect performance. In the case of the medical picture description, consistency is very critical. A computer-based technology must be built to support radiologists. Even if the radiologist is the final one, though, a radiologist will support the final decision.

- For patients with brain tummour, physical symptoms vary from patient to patient
- Some patients don’t even show general symptoms
- MRI scans of patients are more reliable than physical symptoms
- MRI scans sometimes involve multiple noise types that can result in incorrect performance.

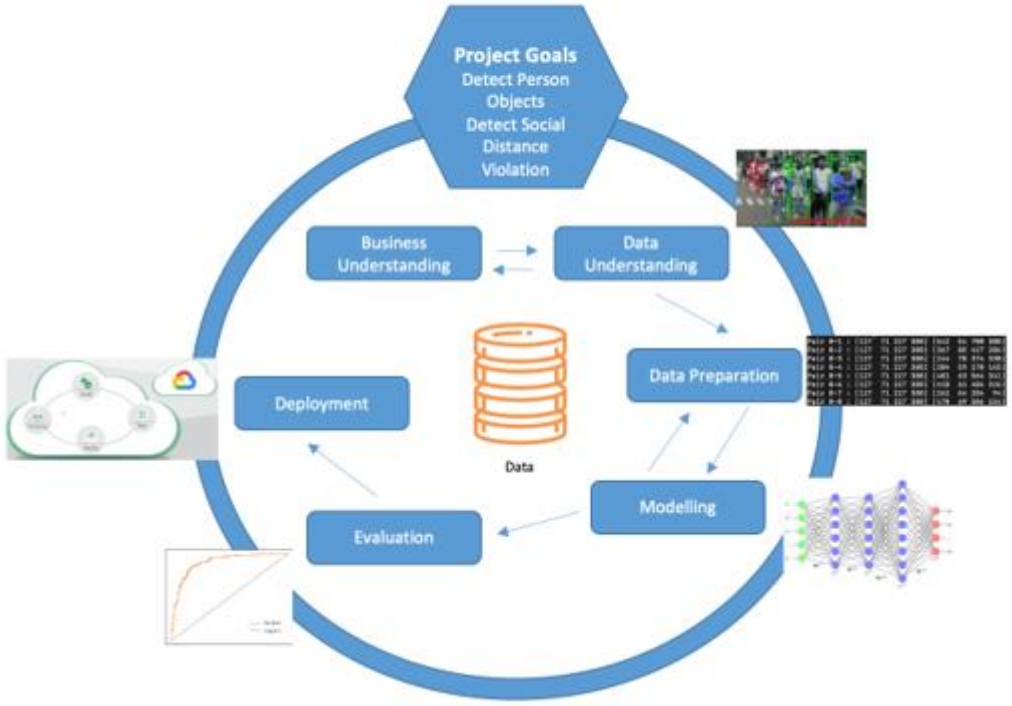
Chapter 4: Objectives of the Study

Previous studies use software that are then manually designed to identify tumour areas. Detection of certain tumours such as gliomas and glioblastomas are very difficult because identifying symptoms are rare. These standardised traits are also less likely to function. Since artificial innovations display immense promise in the field of medicine. Two patchwise architectures of the CNN for patch-by-patch binary tumour and non-tummour areas and full-image architecture of CNN have been proposed.

1. To study and identify the symptoms of brain tummour. Explore new possibilities for brain tummour segmentation algorithm.
2. To analyze the feasibility of brain tummour. Make system enable to differentiate between tummour and non- tummour images.
3. To design and develop the classification model. Comparative study between different algorithms at each stage and finally uses the algorithm which one is best.
4. To evaluate the performance for CNN. Calculate area and volume of tummour.

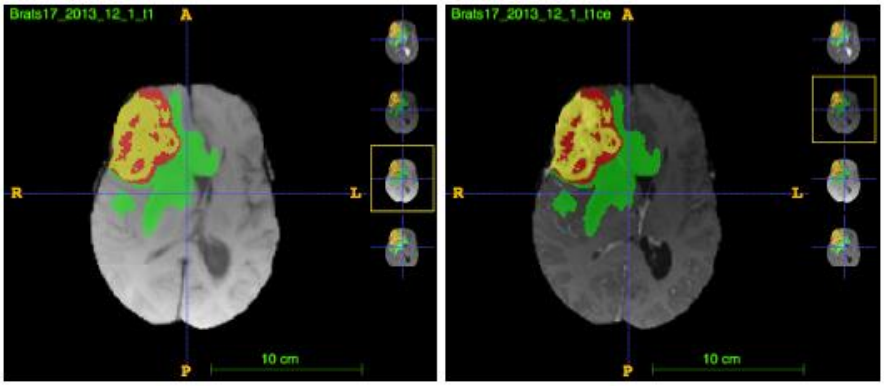
Chapter 5: Project Methodology

This computer vision research project uses CRISP-DM Project execution Methodology.

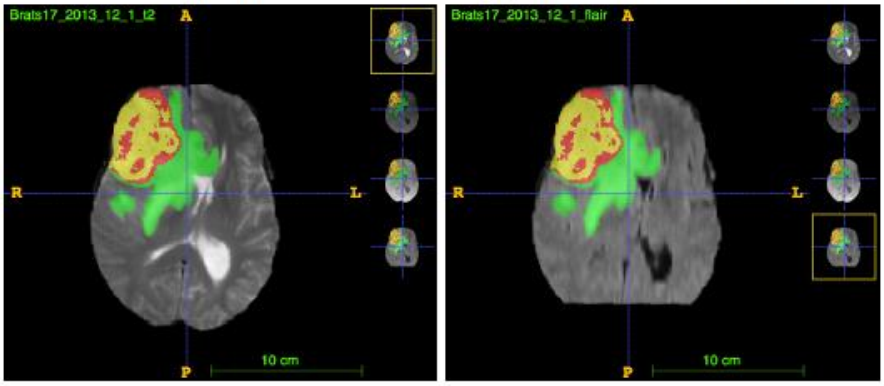


Chapter 6: Data Understanding

A total of 243 individuals (135 glioblastoma and 108 lower-grad gliomas), manually segmented in the tumour, tumour development, and the background regions, in the data collection of the BraTS data system, have photos T1, T1 improved contrast, T2 and FLAIR. Below are outstanding images. The BraTS dataset comprises clusters of necrotic core tumours, core tumour enhancements, core tumour non-enhancements and edoema areas. The entire BraTS data collection is built to achieve the highest possible segmentation score for all four regions, but it focuses here on segmenting tumour regions from the past exclusively. In this article, a total of 178 photographs are used for training models and 44 for assessment. The normalisation of images to zero and unit deflection is needed to pre-process the input images. The images in the BraTS dataset are 240 x 240 x 155 voxels in uniform format, so the images are redundant for zero pads or otherwise processed further.



(a) T1 (b) T1 contrast enhanced



(c) T2 (d) FLAIR

Fig 5: BraTS Dataset Images

MR sequences

In the study or discovery of brain tumours, Magnet Resonance Imaging (MRI) is widely used. There are a number of MR loops, each appropriate for various illustration purposes. It is now standard practice to use a mixture of multiple MR sequences to generate useful and reliable results in automated analysis (Bauer et al., 2011). In this article, three sequences of MR-T1, T2, and FLAIR images-are used and are shortly described. In other words, the following sequences have been used.

T1-WeightedImage

In MRI, T1 is the moment that protons in a tissue have to return to the original magnetization condition caused by the static magnetic field. Easy T1 weighted images (T1 short) provide more anatomical detail than T2 images; usually, they don't provide any useful information for analysing brain tumours.They are also used in conjunction with the liquid balancing agent injection into the vascular system of the recipient (Darmawan, 2019). In T1-weighted photos, the opposing agent displays blood supply. This results in a hyper-intense and readily distinguishable portion of the active tumour and vessels from the surrounding tissues. For the tumour malignancy examination, the active tumour appearance is also used. These images are referred to as "contrast enhanced images of T1," and T1C is used for this purpose.

T2-Weighted Image

T2 means the time to break this coherence as protons interrupted by the pulse of radiofrequencies into coherent oscillation. T2 pics (shortly T2 pictures) are more open to content than T1 pictures Water and thus disease that tends to be hyper-intensive as well as cerebrospinal fluid (CSF) (Cates et al., 2005).

FLAIR Image

FLAIR is a sequence that can eliminate fluids and is used to inhibit CSF in brain imaging. Fluid attenuated inversion recovery This influence makes it possible to distinguish from CSF, which is hypo-intense here, lesions, which remain highly intimate as in T2 images. This is why it is primarily used in measuring brain tumours (Hariharan et al., 2014).

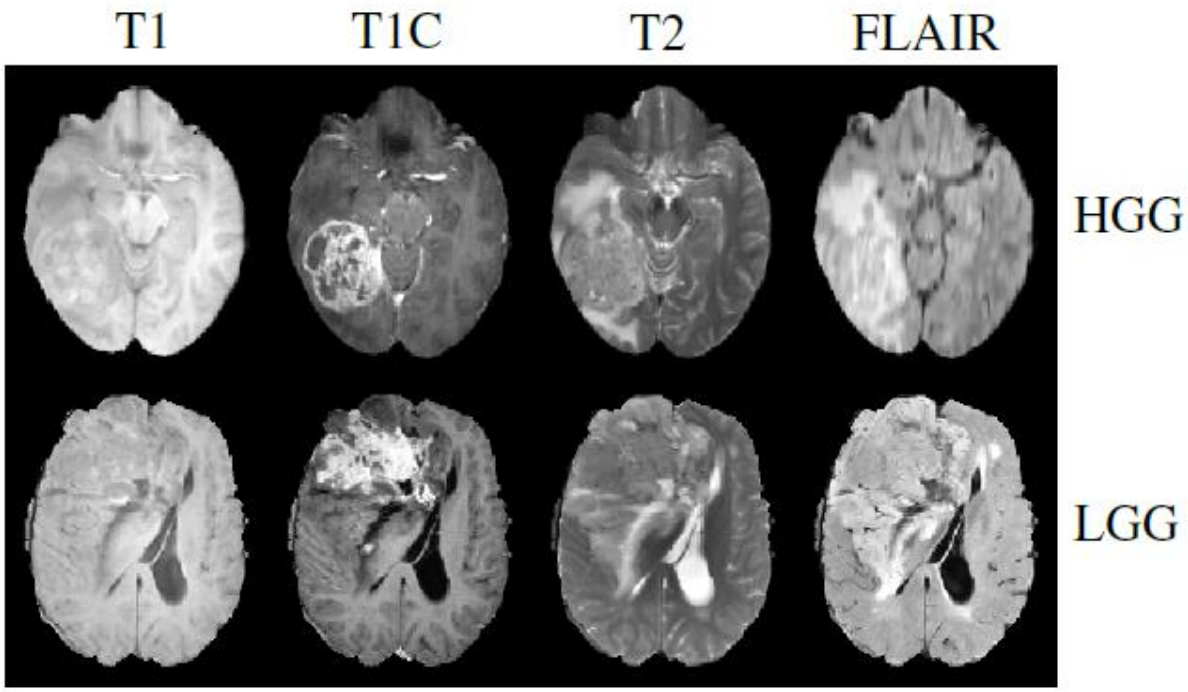


Fig 6: MR Image Types

A comparison of tumour current images of T1, T1C, T2 and FLAIR is shown in Fig. 2. Notice the high-resolution active tumour image T1C, the high intensity tumour and edoema images T2 and FLAIR and hypointensive CSF images FLAIR (Capelle et al., 2000).

Brain Anatomy

The human brain, which serves as a centre for the integration of all parts of the human body, is a highly trained organ that makes a person understand and avoid complex world environments. The brain helps a person to tell things, behave, and exchange thoughts and feelings (Prastawa et al., 2004). In this segment, the objectives of this investigation are identified as the structure of the tissue and its anatomical sections. There are two tissue forms in the brain: grey (GM) and white (WM). Glial and neuronal cell grey matter is also known as glia and neuroglia which controls brain function and the centre of grey matter in white matter. The grey matter is made up. The primary nuclei are caudate nuclei, putamen, pallid and clauster. White matter fibres contain numerous elined axons that bind the cortex of the brain to other brain regions. Callosum corpus is a thick band of fibres of white matter. The hippocampus is connected to the left and right hemisphere.

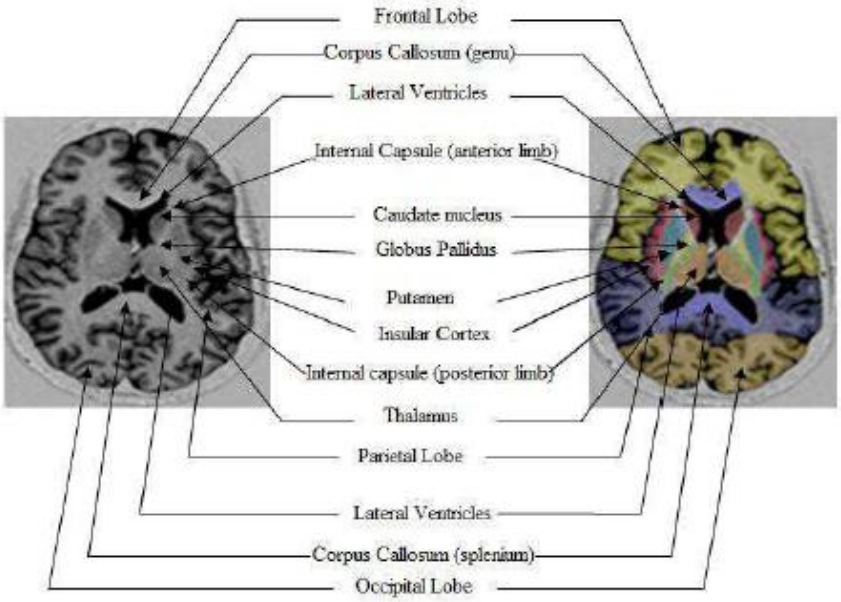


Fig 7: Left side overview: AxialSlice MR Image, Right hand, Image left arm, colourcoded edition

This covers the hippocampus, which contains of cerebrospinal fluid, blood glycosis, calcium, hormones, and white blood cells. The blood circulates into channels (ventricles) in the brain and spinal cord to prevent injuries. The hippocampus and the spinal cord are often referred to as meninges (4). There's a different tissue too. It consists of the cortex and the cortical stem. The hippocampus is primarily in the brain. There is a link between conscious thoughts, behaviour and expectations. (Cates et al., 2005). It has two parts, right and wrong. It is two parts. The other half of the one is handled by others. In addition, the dorsal, temporal, parietal and occipital lobes are grouped into four different lobes in each hemisphere. The cerebellum is the second largest brain tissue. The body's motor functions are related, including walks, balance, balancing and overall muscle power. It lies on the brain behind and is connected to the blocks of stumbling. The liver and hippocampus have a thick grey outer cortex with a minimum of white but large cells. The backbone aligns with the nucleus of the skull. It is behind the head. It's back. Brainstem influences main body functions including brain, sensory receptors, heart, reflex and bibliography. Three elements include the midbrain, pons and medulla oblongata.

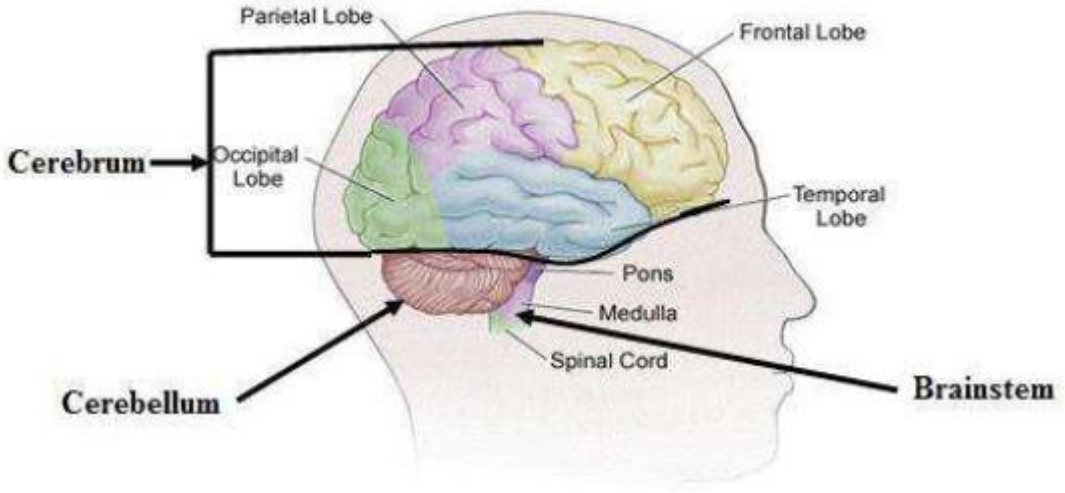


Fig 8: Divisions of Human Brain

Brain Tummour

Although the mechanism that controls normal cells can not regulate the growth of brain cells for some reasons, the brain cells in certain situations enlarge and uncontrollably evolve. Abnormal brain tissue mass is the brain tumour whose place in the skull prevents normal brain-stressing processes (Chen et al., 2016). Increased brain pressure causes certain brain tissues to travel, firmly lock into the skull or inflict some stable nerve tissue damage.

Scientists classify a classified brain tumour based on tumour position, whether it is cancer or tumour. The place of origin (primary secondary) and other factors. The World Health Organisation (WHO) has identified the brain tumour in 120 ways. The cell's source and cell activity is based on this classification, from less violence to more violent (Bauer et al., 2011). In particular, some types of tumours differ between Grade I (not malignant) and Grade IV (more malignant). This means that growth is high due to improvements in ranking schemes that rely on tumours. Primary brain tumours are tumours that arise in the brain that are defined by the types of cells involved. They can be benign (cancerous) and malignant. Benign tumours grow slowly and do not spread or join anywhere (Bauer et al., 2013). The least active tumour in a small room may therefore be very pressurised and inefficient. More active tumours can, by contrast, spread to other tissues more rapidly. Per tumour has its own properties in terms of medicine, radiography and biology.

A separate segment of the body causes secondary brain tumours. Prostate cancer, breast cancer , melanoma, ovarian cancer, leukaemia, sarcoma and some testicular and genital cancers are the most common sources of secondary brain tumours (Bengio, 2012). These tumours are primary cancers that are metastatic in the brain or spread to another part of the body.

MRI Brain Visualisation And Brain Tumour Symptoms:

Different imaging techniques are used to study MRI, CT, CST (Single Photon Emission Computer Tomography) and cerebral angiograpy brain tumours. Different approaches are used in brain tumours (Bengio et al., 2013). In the field of neuroscience there are also several imaging techniques. Duo CT and MR imaging have been the best approaches in recent years due to their widespread availability and versatility for high-resolution photographing of normal and anomalous tissues. A medium to visually image pathogen or other biochemical modifications to living tissue is a magnetic resonance imaging (MRI) system. The resolution of the comparison is higher than the above strategies. The MRI instruments' ability to produce 3D images helps them to retain an outstanding location of the tumour and the ability to collect functional and anatomical information of the tumour (Cireşan et al., 2012).

Until addressing the picture characteristics of MR tumours, it is important to describe the working theory of MR imagery. The patient is held under a strong magnetic field during MR visualisation, creating a parallel (low energy) or parallel (high energy) path for the protons in the water molecule with the magnetic field (Clark et al., 1998). The spinning protons are then separated from the power by using a radiofrequency pulse. The protons rest again as the radio frequency pulse ceases, then at a frequency which is dependent on the magnetic field a sinusoidal sign produces. Finally, the signal is sensed and the image produced by the spindles or resonators of the radio frequency inside the scanner.

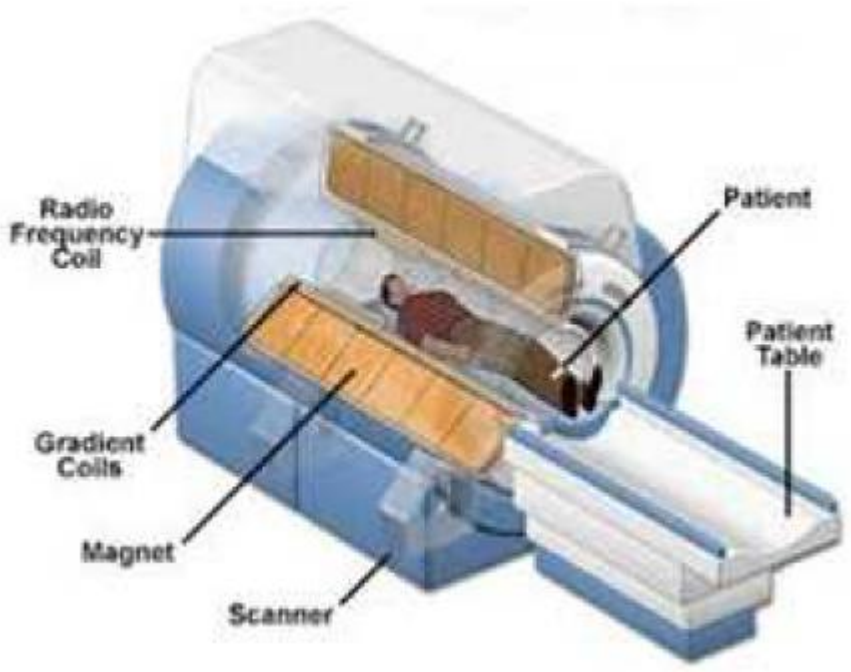


Fig 9: MRI Scanner

Magnatic resonance imaging (MRI) is the imaging technique used primarily in medical facilities to obtain high-grade images inside the human body. Yet X-rays are not associated for a CT-like MRI. Instead, a strong magnetic field defines the orientations of protons that are small magnets that correspond to the outer space (Cobzas et al., 2007).

MR Imaging (MRI)

Raymond V. Damadian used RMI to research the human body in 1969. Finally, since RIM has allowed a person to see the internal structures in some detail, RIM has become the chosen radiological imaging method. A good comparison between different soft body tissues can be seen with MRI. MRI can also have better pictures of brain , muscle, heart or tumours, of comparison to other approaches such as computed tomography (CT) or x-rays (Havaei et al., 2017).

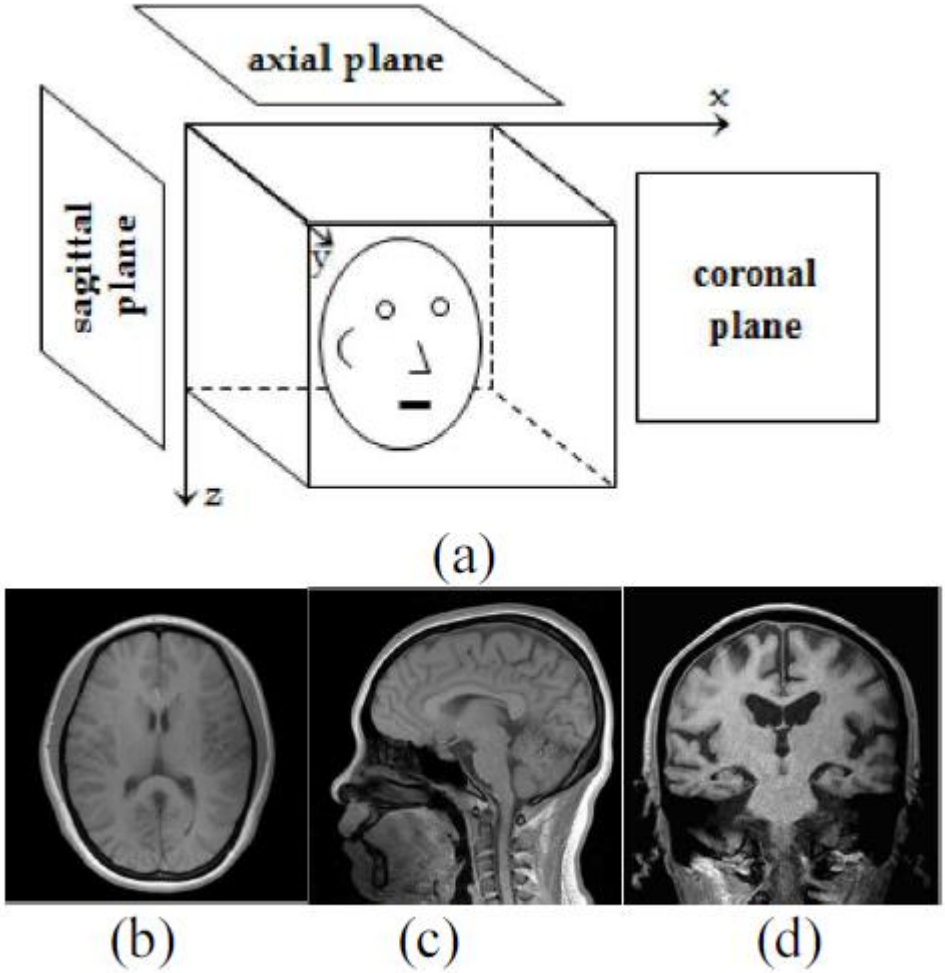


Fig 10: Brain MR Images Form

In MRI signal processing, signal errors are considered. Both are marked by independent magnetic signals, which measure the abticual values of echo length (Tg) and repeat length (Tr). For signal processing from the same body, three different images are needed: T1 , T2, and PD (Proton density) images. T3 images are needed (Bauer et al., 2011). In three medical clinics: flat, sagittal and coronal, the head of the patient is seen above.

Based on the recorded echo structure, there are two main MR imaging families: the spin echo sequence and the gradient echo sequences. The standard MRI pulse sequences were spin echo (SE) with its FSE variant series for anatomical and pathological details. Normal or abnormal brain scans may be done in the MRI scan. The normal brain has components of grey (GM), white (WM) and blood components. Haut. Haut. The pathologic brain normally involves inflammation, necrosis

and edoema in addition to normal brain tissues (Couprie et al., 2013). Necrosis, where an edoema occurs near the outer limits of the tumour, is a dead cell within an outer tumour. Edemas caused by superficial disruption to the brain membrane frequently mix with typical tissues and are difficult to differentiate from other tissues.

The grey strength values are an MRI scan representation in pixel spaces. The grey dependency of the scanned volume on the cell concentration. A blurred region indicates that there is some abnormality (Glorot et al., 2011). Standard brain MR pictures demonstrate the sensitivity of the brain tissu image to improve transparency of CSF, GM to WM, T1-w, and WM, GM to CSF for T2-wighted photographs. for daily brain MR pictures.

MR photos indicate various intensity levels in the brain of the tumour due to the type of tumour on the images T1-w and T2-w. For T1-w, the majority of tumours have low to moderate signal power, but not the amplitude of glioblastomamultiform signal, for example (Goodfellow et al., 2013). Some of the tumours of the T2-w system are great but tumours are of low severity and lymphoma is common.

Chapter 7: Data Preparation

The segmentation of photographs is one of the primary and most involved research fields in the field of medical imaging. The delineation of one or more involved structures in the picture can be described. Automated techniques are tried to prevent manually contouring the structures and taking a long time to stop. In the case of brain tumours, the issue is particularly complicated. In fact, the majority of tumours are heterogeneous, and their severity overlaps with healthy tissues. The appearance of a necrotic centre is always the product of good comparison with the "healthy" tumour (in particular, for glioblastomas, although it appears even with diffuse low-grade gliomas. Previous tumour structure details cannot be used because it has different sizes and textures. In fact, because of their infiltrative nature, DLGGs have highly fluid and erratic borders (Gotz et al., 2014). Since the DLGG's slow-growing characteristics [Sanai 2011], oedema (swelling of tissue around the tumour) and mass effects (drive tissue caused by tumour) are not normal. The simplest segmentation processes like thresholding or area production are not adequate in this sense [Gibbs 1996]. A detailed and accurate segmentation of tumours remains a challenging challenge in lieu of comprehensive and successful progress in the field of tumour segmentation.

Segmentation methods can be categorised in two categories: surface and surface area. The goal of techniques for the surface is the position of a curve / surface with a flow determined by the distribution of the organ or tumour border by the curvature and the restriction of the images (usually gradient) (Hamamci et al., 2012). Regional methods take a broad view of the issue of segmentation. It is here where all voxels must be identified and separated from the rest of the code.

Brain Imaging

There are numerous imaging methods available to allow the brain to be examined. This section gives a concise description of the numerous imaging techniques then focuses on MRI, which is the most popular technique for detecting brain tumours (Hariharan et al., 2014).

Imaging Modalities

The modalities of brain imaging can be divided according to the details collected by them into two categories: structural and functional imaging. Structural or structural approaches are designed to image multiple brain regions and structures. The Computed Tomography (CT) and MRI are among the most common neuroimaging studies (Havaei et al., 2014). MRI is based in the massive quantities in the human body on the magnetic activity of the hydrogen nuclei. Because of the high tissue and specifics and not ionising agents, it is the tool of choice for brain research. New techniques like Diffusion Tensor Imaging (DTI) have been developed in recent years. This modality allows the reconstruction of tracts linked by calculating the anisocoric diffusion of water within tissues to link the various parts of the neural networks in the brain. A tumour may directly affect the structures of the fibres by causing disruption, displacement or infiltration of the fibres [Wei 2007] (Jarrett et al., 2009). Further information on the diagnosis of and analysis of brain tumours can be found in methods like MR spectroscopy (measurement of key tumour tissue metabolites) and Perfusion MRI (measurement by a contrasting agent on relative cerebral blood volume).

The purpose of functional imagery is to analyse the structure of the human brain based on the structural changes induced by the activation of the brain. Electroencephalography (EEG) and magnetizephalography (MEG) are techniques that include direct brain function assessment. EEG feels electrical impulse in the brain by electrodes mounted on the scalp due to neuronal activation. MEG uses sensors located in the scalp to measure the magnetic flux shifts. The methods are common because they are simple , non-invasive and very high time resolution. However, it is difficult to ascertain the exact spatial origin of the signal detected (Khotanlou et al., 2009).

Blood oxygen level based (BOLD) comparison [Ogawa 1990] is used by functionalMRI(fMRI) to diagnose changes in sensorimotor or cognitive functional roles of neuronal activity. Neuronal stimulation induces an increase in blood supply to compensate for oxygen absorption which decreases the number of haemoglobin molecules deoxygenated. The identification of changes in haemoglobin oxygenation is based on the deoxygenated haemoglobin's paramagnetic characteristics that affect the calculated NMR signal (Szegedy et al., 2015). FMRI is also used for the preparation of tumour surgery to define the spatial connexion between the lesion and the functional region and to determine the risks in question. Time limits in signal

processing result in lower-quality fMRI images compared to structural MRIs. For the preparation of brain tumour surgery, functional imaging is especially relevant. It allows the lesion to be connected to the operating region and the procedure to be scheduled accordingly.

The locations of the tumour are:

a) the whole tumour area (including the four components of the tumour).

b) Central area of the tumour (including all of the "edoema" except tumour structure).

c) Improved tumour field (including the structure of the 'enhanced tummour').

For each tumour area, dice, sensitivity and specificity are calculated as follows:

$$\begin{aligned} Dice(P,T) &= \frac{|P_1 \wedge T_1|}{(|P_1| + |T_1|)/2}, \\ Sensitivity(P,T) &= \frac{|P_1 \wedge T_1|}{|T_1|}, \\ Specificity(P,T) &= \frac{|P_0 \wedge T_0|}{|T_0|}, \end{aligned}$$

Chapter 8: Modeling

In this paper, we are proposing three innovative architectures for multimodal RIM brain tummour segmentation: a voxel-wise baseline CNN, a patch-wise fully-convolutionary fully-image CNN and the considerable performance of deep neural networks in a wide range of visual recognition tasks , such as image detection and semanthropical segmentation (Jia Deng et al., 2009).

Baseline Convolutional Network (BCN)

In each picture sub-region, the Base Convolution Network (BCN) model operates by calculating voxel probabilities. We use a 3D coevolutionary kernels CNN architecture illustrated below. A ReLU nonlinearity accompanies any convolutionary layer that is omitted for brevity in the graph (Szegedy et al., 2015). The two last layers are totally connected, 13 kernel convolutions are introduced. A sample randomly from the input picture tensor I is the entry to the baseline model of a 253x4 cube and the result is a 93 volume that gives projections for an inside volume of 9x9. We practise the model by using the L2 control and the Adam optimization of Softmax cross-entropy loss. This figure demonstrates the architecture of the house.

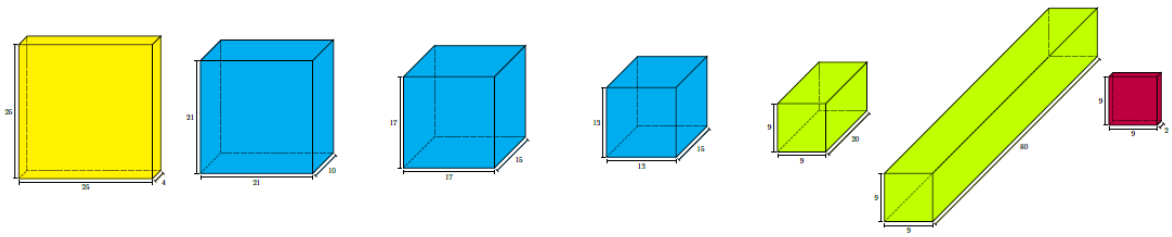


Figure 11: Baseline Architecture Diagram.

The data is 253 x 4 volumes (single slice reflected, slice depth channels). 3D layers of convolution (cyan), accompanied by activation of ReLU. Completely linked layers (green) used as 13-kernel convolutions; ReLU and drop-out obey first line. History and foreground ratings (red) (Urbanska et al., 2014).

Fully Convolutional Network (FCN)

Even a completely reversible patch-wise (FCN) network is introduced. A 243-patch input volume is taken from the FCN and a 243-segmented volume is released. Two 3D convolution layers and 2 3D deconvolution layers are in the network. The chaos is accompanied by an initiation, drop-off, normalisation of batches and max 2x2 pooling. ReLU activation and batch-normalization obey the deconvolution steps. (Bauer et al., 2011). The batch normalisation of transforming models has been demonstrated to minimise internal covariance, while drop-out is infamous for helping to avoid over-setting and boost model generalisation in test dataset evaluation. The model uses L2 regularisation and is conditioned by optimisation with Adam. Softmax cross-entropy loss. In the figure below you can see the model architecture.

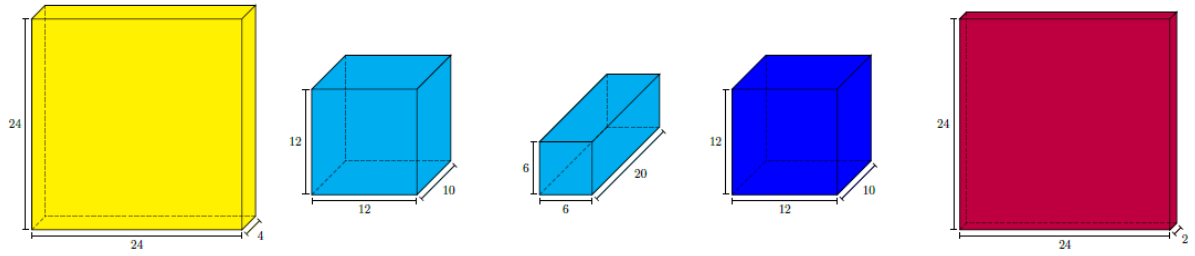


Figure 12: Patchwise FullyConvolutionalNetwork (FCN) Architecture Diagram.

Output similar to the model baseline. 3D convolution layers (cyan) with a ReLU induction, decomposition and max-pooling of two x 2 layers. Three-dimensional (blue) layers followed by reLU and drop-out. History and foreground scores (red) for the whole patch.

FullimageFullyConvolutional Network (FIFCN)

Finally, it is proposed as a design based on the full image of the FIFCN. A duplicate of the forward network of unpools and unwrappable filters, substituted for unpool layers is the deconvolutionary network. The architecture was VGG16 but it has used the network of 11 layers of CNN. To minimise the number of parameters, we have first reduced the number of layers to just 5 coalescing layers, from 8 coalescing and 3 fully connected layers. Second, we use ~4x less philtres for any film. It hold the ReLU activation mechanism and 2x2 max pooling for each convolutionary sheet. In order to mitigate the bias against the background class, the input volume from 240X240x155 until 160x160x144 and mitigate the sampling through 2x2 max pooling has

reduced (Cronin et al., 2018). The image moves through the layers of convolution and disintegration. The output of the network is 160 x 160 x 144, with estimates in tumour and history. This volume is then filled to generate the final display volume with historical pixels. The softmax cross-entropy loss and the loss of the dice score is now evaluated for the loss function. For regulating the L2 and train, Adam optimization is available. FIFCN architecture is seen in the following diagram.

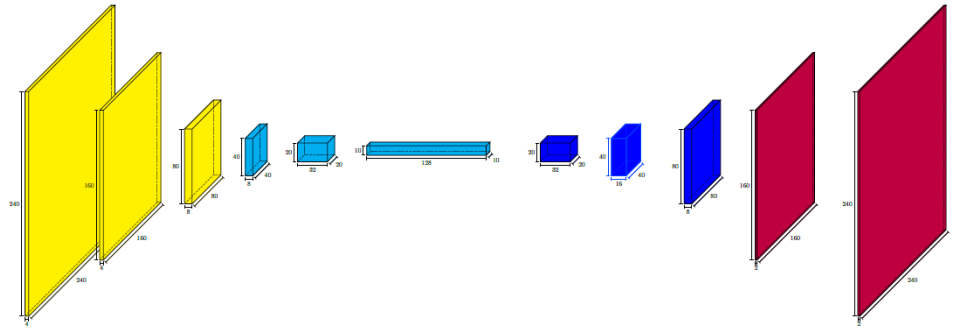


Figure 13: Full-Image FCN Architecture (FIFCN) Diagram.

The entry is a complete picture (one slice, channels displayed as a slice depth). 3D levels of convolution (cyan) and 2 x 2 of the full volume. 3D layers followed by ReLU deconvolution and standardisation of the array. Class ratings for history and four tumour regions are present in Performance (red). Potential speed-up is the biggest benefit of FIFCN.

Chapter 9: Model Evaluation

The key output measurement for all functions on both versions is the mean value, which is a common assessment metric for medical imaging and computer vision. The dice score for output predictions P$\varepsilon\P\prime\Leftrightarrow\infty\diamondsuit$ and the expert's consensus ground truth T$\varepsilon\P\prime\Leftrightarrow\infty\diamondsuit$ is defined as:

$$\text{Dice Score} = \frac{2|P_1 \cap T_1|}{|P_1| + |T_1|}$$

where P1 and T1 are the voxels of which P = 1 and T = 1 are representing (Menze et al., 2015)

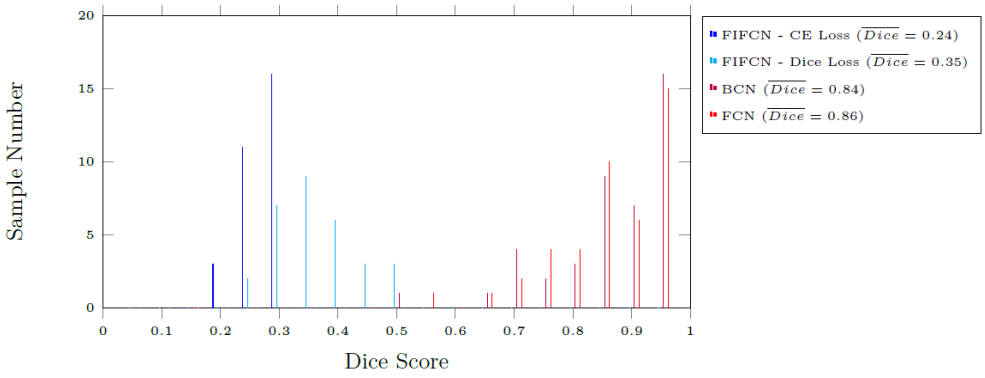
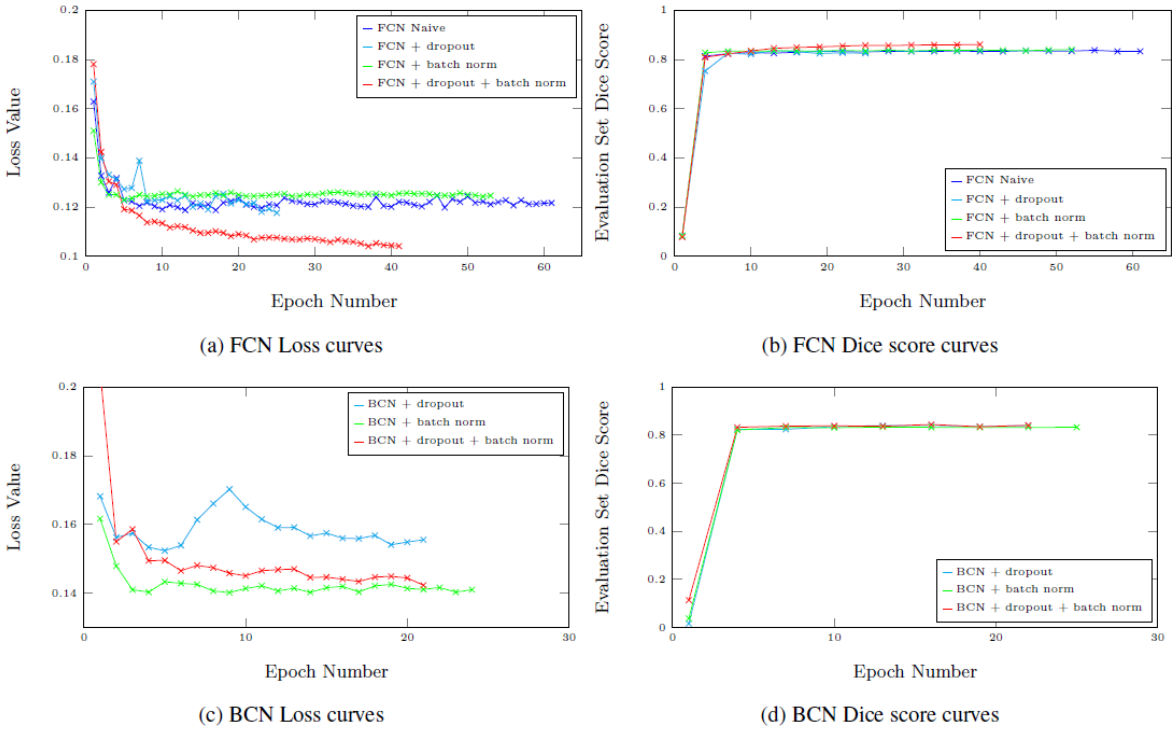


Figure 14: Histogram of dice scores

The inclusion of dropouts and batch normalisation layers typically substantially increases the model dice performance. There are outstanding loses and dice ratings shown in Fig 15. We note that the regular and drop-out models have a comparatively lower final failure rate, but with dropout and batch standardisation the mean dice rate is significantly improved on the other models.



(a) FCN Loss curves

(b) FCN Dice score curves

(c) BCN Loss curves

(d) BCN Dice score curves

Fig 15: Loss and evaluation set dice score curves

Chapter 10: Analysis and Results

The outcomes are taken three primarily. Firstly, the fineness of the model increases the segregation efficiency substantially, both with respect to the histogram (the distribution is even more distorted for the finely tuned models) and the mean dial performance. The initial model is expected to work for the Bra TS dataset, and the BCN has significantly exceeded the FCN. This transition is expected to happen. This is most probably because the BCN is less reliant on those image properties than the FCN. The convolutionary layers in the FCN compact the picture as high-level characteristics and then recreate the segmented image from these attributes. While the design can be quite different, it's implicitly expected the high level functionality would also be similar for similar photos. The consistency and resolution variations of the photographs of BraTS however mean that this statement can not be quite accurate and thus a decrease in the segmentation rate is seen in the BraTS datasets. Similarly, variations between the images may also cause the B CNN model to drop dramatically when shifting from the BraTS dataset because of the high-level characteristics however, some of this drop off is assumed to be offset by using entirely linked layers rather than deconvolutionary layers for prediction purposes.

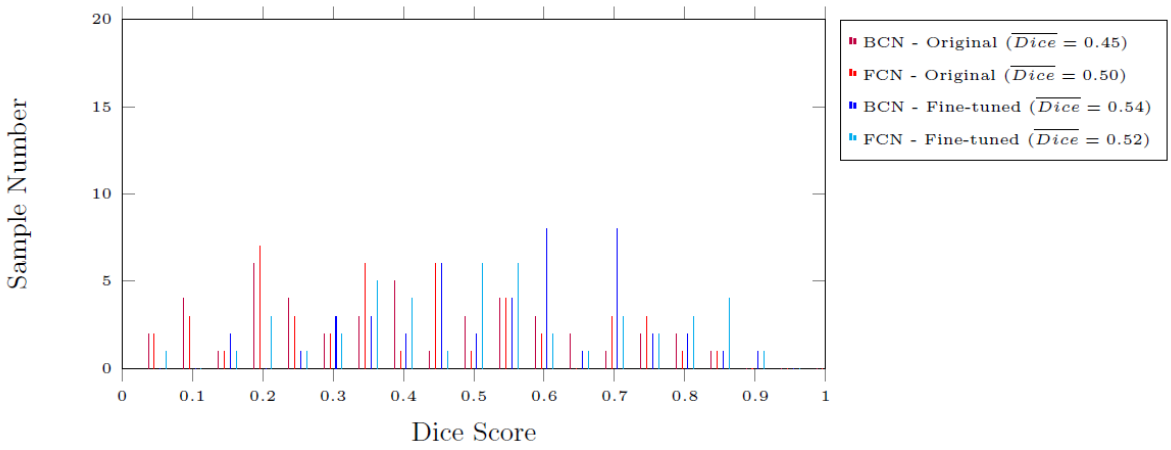


Fig 16: Histogram of Dice Scores

The findings also reveal that the consistency of the segmentation in the validation collection is quite incompatible. For both models, the high and very low segmentation efficiency is observed. In fact, in every bin around the histogram, the fine-tuned FCN model includes samples. This shows that transmitting learning is not always accurate, but highly dependent upon

image quality and resolution, although it can be encouraging for applications into glioma segmentation. Fig-17 below displays an FCN model segmented picture proof. Fine-tuning the model greatly increases the segmentation efficiency, as seen in Table 1. In fact, when using the fine tuning model, the number of unspecified voxels (red) is decreased dramatically. The segmentation efficiency of the BraTS dataset is now higher.

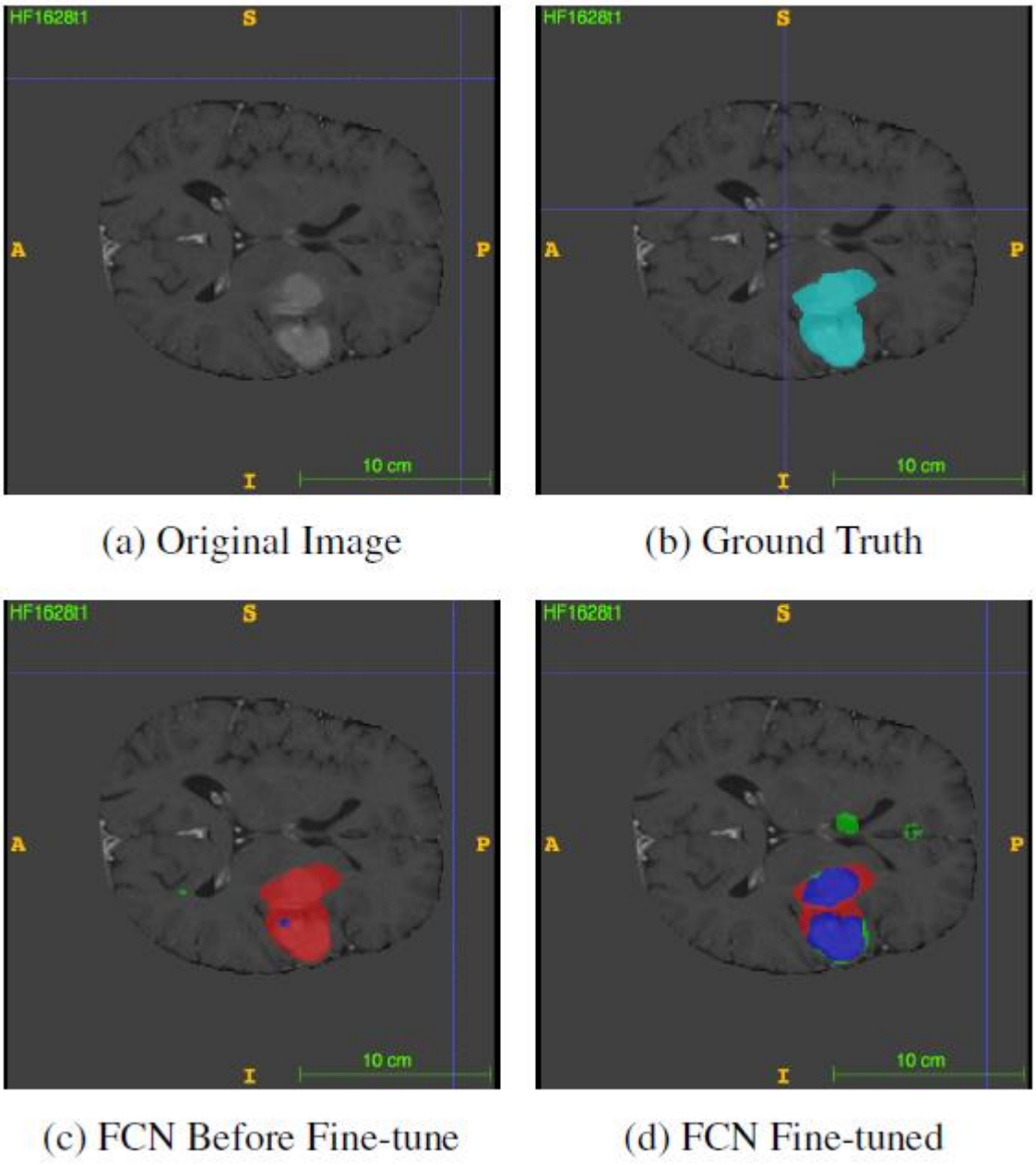


Fig 17: segmentations via transfer learning.

In general, transfer learning is a good method for CNNs for the segmentation of brain MRI by using small datasets but more work needs to be done. In pre-processing and preparing for pictures with varying sizes, in fact, there are several difficulties to address. However, as shown by the fine tuning findings, a model 's accuracy can also be substantially increased by fine tuning, which in turn implies that transmitting modelling can effectively be extended in future work to brain segmentation.

Model	Method	Average Dice Score (%)
BCN	Pre-trained	80.8
BCN	Re-trained	89.2
BCN	Fine-tuned	**97.9**
FCN	Pre-trained	78.2
FCN	Re-trained	90.7
FCN	Fine-tuned	**92.2**

Table 1: Ccomparison for transfer learning experiments of both BCN and FCN.

Chapter 11: Conclusions and Recommendations for future work

A separation and detection mechanism focused on profound convolutionary neural networks has been introduced in this article. Different architectures are considered and their effect on results has been studied (Al et al., 1999). Dice and negative log Hausdorff are available for all three forms of tumours. The strategies are semi-automatically shown in colour. The methods are graded in each subfigure according to their average value. The average is found in green, red and blue, the highest.

A high efficiency is achieved by the modern architecture and modelling local mark dependencies by storing 2 CNNs (which can model both local and world information). A two-stage training process is performed which helps to effectively train CNNs if the distribution of labels is unbalanced (Zeiler & Fergus, 2012).

In order to increase the overall network generalisation potential, other approaches to the segmentation of the datasets will be considered (e.g., the number of subjects). The architecture would be tailored to use during a brain procedure, to identify and to find the tumour correctly, among the other changes. In real-time and in real-world environments, the identification of tumours in storage space should be carried out and therefore the network should also be modified to a 3D device in that case (Zikic et al., 2012). It may be possible to detect in real time by keeping the network architecture simple. In future, our built neural network and enhanced network output in other medical images are to be investigated.

Bibliography

Ain, Q., Sciences, E., Mehmood, I., Naqi, M., & Jaffar, A. (2010). Bayesian Classification Using DCT Features for Brain TummourDetection. *Researchgate.Net*, *September*. https://doi.org/10.1007/978-3-642-15387-7

Akselrod-Ballin, A., Galun, M., Gomori, M. J., Filippi, M., Valsasina, P., Basri, R., & Brandt, A. (2006). An integratedsegmentation and classification approach applied to multiple sclerosisanalysis. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, *1*, 1122–1129. https://doi.org/10.1109/CVPR.2006.55

Al, X. E. T., Xing, E. P., & Jordan, M. I. (1999). A generalizedmean field algorithm for variational inference mexponential families. *XING ET AL.*, 583–591.

Ambrosini, R. D., Wang, P., & O'Dell, W. G. (2010). Computer-aided detection of metastatic brain tummours using automated threedimensional template matching. *Journal of Magnetic Resonance Imaging*, *31*(1), 85–93. https://doi.org/10.1002/jmri.22009

Bauer, S., Nolte, L. P., & Reyes, M. (2011). FullyAutomaticSegmentation of Brain Tummour Images Using Support Vector MachineClassification in Combination withBauer, S., Nolte, L. P., & Reyes, M. (2011). Fully AutomaticSegmentation of Brain Tummour Images Using Support Vector Machine Classification i. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *6893 LNCS*(PART 3), 354–361. https://doi.org/10.1007/978-3-642-23626-6_44

Bauer, S., Wiest, R., Nolte, L. P., & Reyes, M. (2013). A survey of MRI-based medical image analysis for brain tummour studies. *Physics in Medicine and Biology*, *58*(13). https://doi.org/10.1088/0031-9155/58/13/R97

Bengio, Y. (2012). Practical recommendations for gradient-based training of deep architectures. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *7700 LECTU*, 437–478. https://doi.org/10.1007/978-3-642-35289-8-26

Bengio, Y., Courville, A., & Vincent, P. (2013). Representation learning: A review and new perspectives. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *35*(8), 1798–1828. https://doi.org/10.1109/TPAMI.2013.50

Capelle, A. S., Femandez, C., Lefevre, S., Ferrie, J. C., Poitiers, B. P., & Spmi, B. (2000). UNSUPERVISED SEGMENTATION FOR AUTOMATIC DETECTION OF BRAIN

TUMMOURS IN MRI. *IEEE Transactions on Medical Imaging*, 613–616.

Cates, J. E., Whitaker, R. T., & Jones, G. M. (2005). Case study: An evaluation of user-assisted hierarchical watershed segmentation. *Medical Image Analysis*, *9*(6), 566–578. https://doi.org/10.1016/j.media.2005.04.007

Chen, H., Dou, Q., Yu, L., & Heng, P.-A. (2016). VoxResNet: Deep Voxelwise Residual Networks for Volumetric Brain Segmentation. *ArXiv*, 1–9. http://arxiv.org/abs/1608.05895

Cireşan, D. C., Giusti, A., Gambardella, L. M., & Schmidhuber, J. (2012). Deep neural networks segment neuronal membranes in electron microscopy images. *Advances in Neural Information Processing Systems*, *4*(January), 2843–2851.

Clark, M. C., Hall, L. O., Goldgof, D. B., Velthuizen, R., Reed Murtagh, F., & Silbiger, M. S. (1998). Automatic tummour segmentation using knowledge-based techniques. *IEEE Transactions on Medical Imaging*, *17*(2), 187–201. https://doi.org/10.1109/42.700731

Cobzas, D., Birkbeck, N., Schmidt, M., Jagersand, M., & Murtha, A. (2007). 3D variational brain tummour segmentation using a high dimensional feature set. *Proceedings of the IEEE International Conference on Computer Vision*, 0–7. https://doi.org/10.1109/ICCV.2007.4409130

Couprie, C., Najman, L., & Lecun, Y. (2013). Learning Hierarchical Features for Scene Labeling. *Pattern Analysis and Machine Intelligence, IEEE Transactions On*, *35*(8), 1915–1929. https://doi.org/10.1109/TPAMI.2012.231

Cronin, K. A., Lake, A. J., Scott, S., Sherman, R. L., Noone, A. M., Howlader, N., Henley, S. J., Anderson, R. N., Firth, A. U., Ma, J., Kohler, B. A., & Jemal, A. (2018). Annual Report to the Nation on the Status of Cancer, part I: National cancer statistics. *Cancer*. https://doi.org/10.1002/cncr.31551

Darmawan, D. (2019). On a measure of divergence between two statistical populations defined by their probability distribution. *Journal of Chemical Information and Modeling*, *53*(9), 1689–1699. https://doi.org/10.1017/CBO9781107415324.004

Glorot, X., Bordes, A., & Bengio, Y. (2011). Domain adaptation for large-scale sentiment classification: A deep learning approach. *Proceedings of the 28th International Conference on Machine Learning, ICML 2011*, *1*, 513–520.

Goodfellow, I. J., Warde-Farley, D., Lamblin, P., Dumoulin, V., Mirza, M., Pascanu, R., Bergstra, J., Bastien, F., & Bengio, Y. (2013). Pylearn2: a machine learning research

library. *ArXiv*, 1–9. http://arxiv.org/abs/1308.4214

Gotz, M., Weber, C., Blocher, J., Stieltjes, B., Meinzer, H., & Maier-Hein, K. (2014). Extremely randomized trees based brain tummour segmentation. in Proceedings of BRATS Challenge - MICCAI. 2014. *Researchgate.Net*, *March 2015*, 1–6. http://people.csail.mit.edu/menze/papers/proceedings_miccai_brats_2014.pdf

Hamamci, A., Kucuk, N., Karaman, K., Engin, K., & Unal, G. (2012). Tummour-cut: Segmentation of brain tummours on contrast enhanced mr images for radiosurgery applications. *IEEE Transactions on Medical Imaging*, *31*(3), 790–804. https://doi.org/10.1109/TMI.2011.2181857

Hariharan, B., Arbeláez, P., Girshick, R., & Malik, J. (2014). Simultaneous detection and segmentation. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *8695 LNCS*(PART 7), 297–312. https://doi.org/10.1007/978-3-319-10584-0_20

Havaei, M., Davy, A., Warde-Farley, D., Biard, A., Courville, A., Bengio, Y., Pal, C., Jodoin, P. M., & Larochelle, H. (2017). Brain tummour segmentation with Deep Neural Networks. *Medical Image Analysis*, *35*, 18–31. https://doi.org/10.1016/j.media.2016.05.004

Havaei, M., Jodoin, P. M., & Larochelle, H. (2014). Efficient interactive brain tummour segmentation as within-brain kNN classification. *Proceedings - International Conference on Pattern Recognition*, 556–561. https://doi.org/10.1109/ICPR.2014.106

Jarrett, K., Kavukcuoglu, K., Ranzato, M., & LeCun, Y. (2009). What is the best multi-stage architecture for object recognition? *Proceedings of the IEEE International Conference on Computer Vision*, 2146–2153. https://doi.org/10.1109/ICCV.2009.5459469

Jia Deng, Wei Dong, Socher, R., Li-Jia Li, Kai Li, & Li Fei-Fei. (2009). ImageNet: A large-scale hierarchical image database. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 248–255. https://doi.org/10.1109/cvprw.2009.5206848

Jones, T. L., Byrnes, T. J., Yang, G., Howe, F. A., Bell, B. A., & Barrick, T. R. (2015). Brain tummour classification using the diffusion tensor image segmentation (D-SEG) technique. *Neuro-Oncology*. https://doi.org/10.1093/neuonc/nou159

Khotanlou, H., Colliot, O., Atif, J., & Bloch, I. (2009). 3D brain tummour segmentation in MRI using fuzzy classification, symmetry analysis and spatially constrained deformable models. *Fuzzy Sets and Systems*, *160*(10), 1457–1473. https://doi.org/10.1016/j.fss.2008.11.016

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2017). ImageNet classification with deep convolutional neural networks. *Communications of the ACM*. https://doi.org/10.1145/3065386

Menze, B. H., Jakab, A., Bauer, S., Kalpathy-Cramer, J., Farahani, K., Kirby, J., Burren, Y., Porz, N., Slotboom, J., Wiest, R., Lanczi, L., Gerstner, E., Weber, M. A., Arbel, T., Avants, B. B., Ayache, N., Buendia, P., Collins, D. L., Cordier, N., … Van Leemput, K. (2015). The Multimodal Brain Tummour Image Segmentation Benchmark (BRATS). *IEEE Transactions on Medical Imaging*, *34*(10), 1993–2024. https://doi.org/10.1109/TMI.2014.2377694

MuhammadWaqas Nadeem, Mohammed A. Al Ghamdi, Muzammil Hussain, Muhammad Adnan Khan, Khalid Masood Khan, S. H. A. & S. A. B. (2020). Brain Tummour Analysis Empowered with Deep Learning: A Review, Taxonomy, and Future Challenges. *Brain Sciences Www.Mdpi.Net*, 1–33.

Prastawa, M., Bullitt, E., Ho, S., & Gerig, G. (2004). A brain tummour segmentation framework based on outlier detection. *Medical Image Analysis*, *8*(3), 275–283. https://doi.org/10.1016/j.media.2004.06.007

Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., & Rabinovich, A. (2015). Going deeper with convolutions. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*. https://doi.org/10.1109/CVPR.2015.7298594

Urbanska, K., Sokolowska, J., Szmidt, M., & Sysa, P. (2014). Glioblastoma multiforme - An overview. *Wspolczesna Onkologia*, *18*(5), 307–312. https://doi.org/10.5114/wo.2014.40559

Zeiler, M. D., & Fergus, R. (2012). Visualizing and Understanding Convolutional Networks. *ArXiv*.

Zikic, D., Glocker, B., Konukoglu, E., Criminisi, A., Demiralp, C., & Shotton, J. (2012). Decision Forests for Tissue-specific Segmentation of High-grade Gliomas in Multi-channel MR. *Springer-Verlag Berlin Heidelberg*, *m*, 1–8.

Appendix

Plagiarism Report[1]

AUTO-DETECTION OF BRAIN TUMOUR FROM MRI DEEP LEARNING

by Satyajit Pal

Submission date: 17-Sep-2020 10:14AM (UTC+0530)
Submission ID: 1389215138
File name: in_Tumour_from_MR_Images_Using_Deep_Learning_Project_Report.docx (1.32M)
Word count: 8199
Character count: 45526

[1] Turnitn report to be attached from the University.

AUTO-DETECTION OF BRAIN TUMOUR FROM MRI DEEP LEARNING

ORIGINALITY REPORT

6%	%	%	%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	westfieldcomics.com Internet Source	1%
2	Submitted to Galgotias University, Greater Noida Student Paper	1%
3	"Detection of Brain Tumor using Image Processing Techniques", International Journal of Engineering and Advanced Technology, 2019 Publication	1%
4	Submitted to Federal University of Technology Student Paper	1%
5	student-friendly.blogspot.com Internet Source	<1%
6	www.termpaperwarehouse.com Internet Source	<1%
7	digitalcommons.mtu.edu Internet Source	<1%
	Submitted to Sogang University	

8	Student Paper	<1%
9	Submitted to St. John Fisher College Student Paper	<1%
10	www.ijitjournal.org Internet Source	<1%
11	link.springer.com Internet Source	<1%
12	Submitted to Chandigarh University Student Paper	<1%
13	www.inmybangalore.com Internet Source	<1%
14	Lecture Notes in Computer Science, 2015. Publication	<1%
15	Submitted to Institute of Technology Blanchardstown Student Paper	<1%
16	researcharchive.vuw.ac.nz Internet Source	<1%

Exclude quotes	On	Exclude matches	< 10 words
Exclude bibliography	On		

Publications in a Journal/Conference Presented/White Paper[2]

Paper -1: Ransomware Auto-Detection in IOT Devices Using Machine Learning

https://ijesc.org/upload/a20f8938071ad706c04baf48bf5c98b9.Ransomware%20Auto-Detection%20in%20IoT%20Devices%20using%20Machine%20Learning.pdf

[2] URL of the white paper/Paper published in a Journal/Paper presented in a Conference/Certificates to be provided.

ISSN XXXX XXXX © 2018 IJESC

IJESC

Research Article Volume 8 Issue No.12

Ransomware Auto-Detection in IoT Devices using Machine Learning

Anshuman Dash[1], Satyajit Pal[2], Chinmay Hegde[3]
REVA Academy for Corporate Excellence, REVA University, Bengaluru, India

Abstract:
The term Internet of Things (often abbreviated IoT) was coined by industry researchers but has emerged into mainstream public view only more recently. The IoT is a massive group of devices containing sensors or actuators connected over wired or wireless networks. IoT has been rapidly growing over the past decade and, during the growth, security has been identified as one of the weakest areas in IoT. There are over six billion estimated devices currently connected to the Internet and an estimate of over 25 billion connected by 2020. IoT and its applications propagate to majority of life's infrastructure ranging from health and food production to smart cities and urban management. While efficiency and prevalence of IoT are increasing, security issues remain a necessary concern for industries. Internet connected devices, including those deployed in an IoT architecture, are increasingly targeted by cybercriminals due to their pervasiveness and the ability to use the compromised devices to further attack the underlying architecture. In the case of ransomware, for example, devices that can store a reasonably amount of data are likely to be targeted. Thus, ensuring the security of IoT nodes against threats such as malware is a topic of ongoing interest. While malware detection and mitigation research are now new, ransomware detection and mitigation remain challenging. Ransomware is a relatively new malware type that attempts to encrypt a compromised device's data using a strong encryption algorithm. The victim will then have to pay the ransom (usually using bitcoins) to obtain the password or decryption key. Consequences include temporary or permanent loss of sensitive information, disruption of regular operations, direct/indirect financial losses. In this paper, we present a machine learning based approach to detect ransomware of IoT devices. Specifically, our proposed approach outperforms K-Nearest Neighbors, Neural Networks, Support Vector Machine and Random Forest, in terms of accuracy rate, recall rate and precision rate.

Keywords: Ransomware detection, Internet of Things (IoT) security, Machine learning, Malware detection, DDoS

I. INTRODUCTION

The concept of Internet of things(IoT) was introduced by the growth of the widely used global network known as the internet along with the deployment of ubiquitous computing and mobiles in smart objects which brings new opportunities for the creation of innovative solutions to various aspects of life [1].

The concept of Internet of things(IoT)creates a network of objects that can communicate, interact and cooperate to reach a common goal [2].

IoT devices can enhance our daily lives, as each device stops acting as a single device and become part of an entire full connected system. This provides us with the resulting data to be analyzed for better decision making, tracking our businesses and monitoring our properties while we are far away from them [3].

While efficiency and prevalence of IoT are increasing, security issues remain a necessary concern for industries. Internet-connected devices, including those deployed in IoT architecture, are increasingly targeted by cybercriminals due to their pervasiveness and the ability to use the compromised devices to further attack the underlying architecture [4].

In the case of ransomware, for example, devices that can store a reasonably amount of data (e.g., Android and iOS devices) are likely to be targeted. Thus, ensuring the security of IoT nodes against threats such as malware is a topic of ongoing interest.

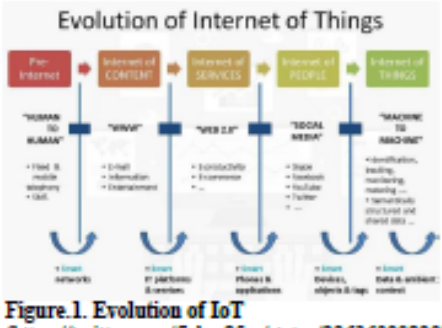


Figure.1. Evolution of IoT
(https://twitter.com/fisher85m/status/926360908900773889)

II. IOT APPLICATIONS

As the paradigm of IoT is growing, it is stepping into every aspect of our lives. This leads to an easier life through wider range of applications such as smart city and its basic utilities.

A. Basic Utility Systems in Smart City

- ***Smart Water Management System:*** In this innovative day and age, the Internet of Things (IoT) is offering new solutions for improving water management to maximize efficient use. Comprehensive strategies utilizing the IoT can reduce water costs up to 20 percent. One of the reasons is how cheap IoT sensors are becoming with new battery-powered networking solutions (like LORA).
- ***Smart Energy Management Systems:*** Energy is a very important aspect for any household, industries, agriculture and

so. Managing the energy efficiently and conserving it intelligently for appliances is very much important. The energy usage is directly affected with Coal, oil and so towards power generation.

- *Smart Gas Management Systems:* Urban natural gas provisioning has made significant progress in recent years. In the smart gas field, IoT enables stable, real-time traffic data collection from gas meters, device status monitoring, command delivery, and additional remote operations.
- *Smart Sewage Management Systems:* IOT in wastewater facility installs smart sensors at various points in their water management system. These sensors collect data on water quality, temperature variations, pressure changes, water, and chemical leaks, and they send those data back to a web application that synthesizes the information into actionable insights.
- *Smart Lighting Management Systems:* Improving safety within the city is one of the major benefits of smart lighting solution. The way smart lights work is that they derive power from the power grid and the street light poles have small cells that are a key enabler for smart solutions.

Figure.2. An illustration of IoT based Smart City with Smart Utility Systems

B. Safety & Security (Surveillance)

A safer, mobilized version of our current cities where everything is automated through technologies such as artificial intelligence (AI), cloud, Big Data and analytics is what we envisage. But one of the key components that is almost forgotten is the importance of surveillance—24/7 round the clock monitoring of citizens, enabled with a robust infrastructure to ensure the safety and security of civilians.

- *Environmental sensors:* Sensors can measure environmental conditions such as air quality, temperature and humidity, water quality and noise levels, which are useful for city authorities to direct resources towards any mitigated imbalances.
- *Support to law enforcement:* Video surveillance feed has often been useful tool for law enforcement agencies to gather primary evidence during investigations. It is therefore critical for civic authorities to deploy high-quality CCTV cameras with required software interface to ensure that the recorded feed is of high-resolution.

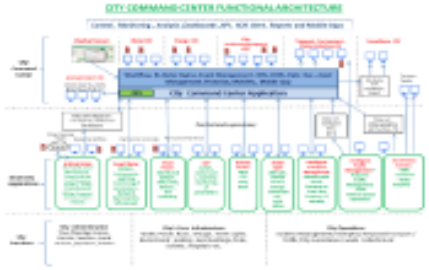


Figure.3.An illustration of IoT based Smart City Functional Architecture
(https://docs.google.com/presentation/d/1ck3o4M21qEvvXTLgXL56mdCwi6pvYlmITXNyv7JV0TU/edit#slide=id.g3e3c360a4a_0_27)

C. Mobility

- *Smart Traffic Management Systems:* Traffic management is one of the biggest infrastructure hurdles faced by developing countries today. Thus, the increased use of vehicles has caused an immense amount of traffic congestion. Several countries are overcoming this traffic bottleneck by fetching information from CCTV feeds and transmitting vehicle-related data to city traffic management centers to help create improvements.

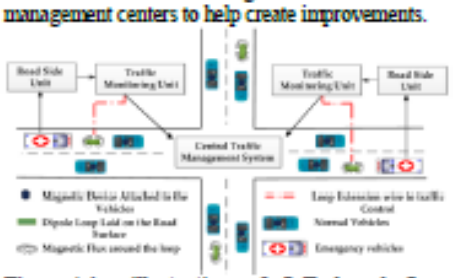


Figure.4.An illustration of IoT based Smart Traffic Managements System
(https://www.mdpi.com/1424-8220/16/2/157)

- *Smart Transport Management Systems:* Smart transportation is developed on the base of smart infrastructure that includes not only multi-modal connected conveyance but also automated traffic signals, tolls and fare collection. Smart services offer different benefits, from smart parking and vehicle locating systems, to route diversion alerts.

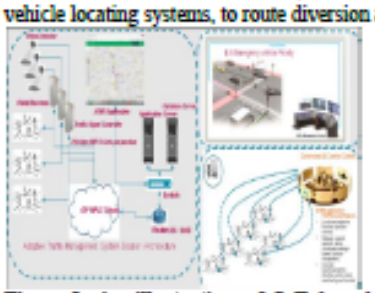


Figure.5. An illustration of IoT based Smart Transport Management System
(https://docs.google.com/presentation/d/1ck3o4M21qEvvXTLgXL56mdCwi6pvYlmITXNyv7JV0TU/edit#slide=id.g3e3c360a4a_0_27)

D. Governance

Smart governance or good governance are two sides of the same coin. The use of the internet and digital technology is creating a progressive government- public partnership, strengthening government institutions and integrating all sections of society. To effectively manage these segments of society, our cities need smart administration and governance. Web portals, online forums, mobile apps and their unified services have helped public to directly share their questions, suggestions and grievances to government authorities. The forum on the website gives fellow applicants a space to share ideas, suggestions and know the status of other applications filed.

III. IOT TECHNOLOGIES AND PROTOCOLS

Several Communication Protocols and Technology used in the internet of Things. Some of the major IoT technology and

protocol (IoT Communication Protocols) are discussed here. These IoT communication protocols cater to and meet the specific functional requirement of an IoT system.

A. 6LOWPAN

6LoWPAN is connecting more things to the cloud. Low-power, IP-driven nodes and large mesh network support make this technology a great option for Internet of Things (IoT) applications. As the full name implies – "IPv6 over Low-Power Wireless Personal Area Networks" – 6LoWPAN is a networking technology or adaptation layer that allows IPv6 packets to be carried efficiently within small link layer frames, such as those defined by IEEE 802.15.4 [7].

- ***Network Architecture:***The uplink to the Internet is handled by the Access Point (AP) acting as an IPv6 router. Several different devices are connected to the AP in a typical setup, such as PCs, servers, etc. The 6LoWPAN network is connected to the IPv6 network using an edge router.
- ***Security:***Security is a must for IoT systems and always presents a challenge. Due to the nature of IoT with many nodes that in many cases have very constrained performance, there are also more entry points for an outside attacker. Another critical aspect is that data flowing in a typical IoT system is not just "data," the damage potential is much higher since the data flowing in the system can be used to open the door to you house or turn on/off alarms remotely.

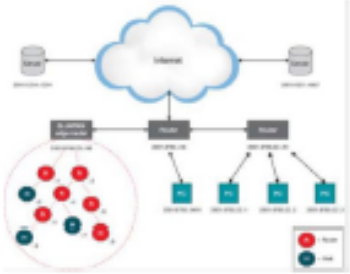


Figure.6. 6LoWPAN Network Architecture
(https://www.researchgate.net/figure/An-example-of-an-IPv6-network-with-a-6LoWPAN-mesh-network-26_fig2_318075856)

B. DASH7

DASH7 Alliance protocol is an Actuator network protocol and it is an open source wireless sensor network. DASH7 operates in 433 MHz, 868 MHz and 915 MHz unlicensed ISM bands. DASH7 can provide a range up to 2km. For security AES 128-bit shared key encryption is used. Data rate offered by DASH7 is 28 kbps and it uses wakeup signal to achieve low power, low latency and elegant architecture. DASH7 uses BLAST networking technology. BLAST means Bursty Light data Asynchronous and Transitive. In bursty, data transfer is abrupt and it does not contain contents like audio and video. In light, multiple consecutive packet transmission is generally avoided and the packet size is limited to 256 bytes. DASH7 communicates through command- response and therefore periodic hand shaking is not required.

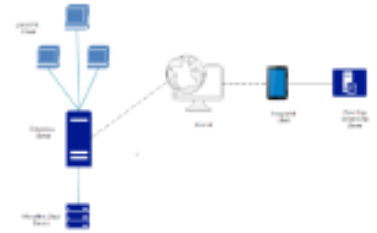


Figure.7. DASH7 Network Architecture
(http://iopscience.iop.org/article/10.1088/1757-899X/396/1/012027/pdf)

C. LoRa &LoRaWAN

- ***LoRa:***It is a radio modulation technology by Semtech Corporation. LoRa [3] provides connectivity about 15 to 20 km using chirp spread spectrum technique in which entire band width is used to transmit single signal. LoRa [8] uses 868 MHz to 900 MHz ISM bands for its operation and data rate is 0.3 kbps. Since LoRa have a long battery life, cost of replacement of devices can be reduced and its deployment is not so complex. LoRa use symmetric key cryptography to ensure security to its devices. Chirp spread spectrum used by LoRa can reduce the signal degradation, noise etc. while transmitting signal.

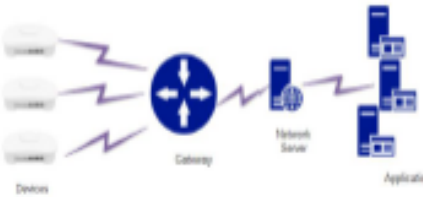


Figure. 8. LoRa Network Architecture
(http://iopscience.iop.org/article/10.1088/1757-899X/396/1/012 027/pdf)

- ***LoRaWAN:*** LoRaWAN defines the communication protocol and system architecture for the network while the LoRa physical layer enables the long-range communication link. The protocol and network architecture have the most influence in determining the battery lifetime of a node, the network capacity, the quality of service, the security, and the variety of applications served by the network.
- ***LoRaWAN Network Architecture:*** Many existing deployed networks utilize a mesh network architecture. In a mesh network, the individual end-nodes forward the information of other nodes to increase the communication range and cell size of the network. While this increases the range, it also adds complexity, reduces network capacity, and reduces battery lifetime as nodes receive and forward information from other nodes that is likely irrelevant for them. Long range star architecture makes the most sense for preserving battery lifetime when long-range connectivity can be achieved.
- ***Security:*** LoRaWAN™ utilizes two layers of security: one for the network and one for the application. The network security ensures authenticity of the node in the network while the application layer of security ensures the network operator does not have access to the end user's application data. AES encryption is used with the key exchange utilizing an IEEE EUI64 identifier.

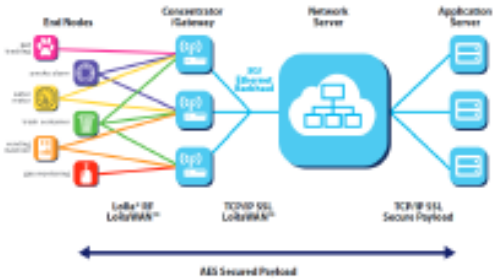


Figure.9. LoRaWAN Network Architecture
(https://www.profissionaisti.com.br/2017/11/iot-protocolo-lorawan-e-principais-placas-de-desenvolvimento-lora/)

D. SigFox

Sigfox is a LPWAN technology which provides low power low data and low-cost communication devices. It operates in global network and is used for IoT communications. Even though there are plenty of wireless technologies Sigfox is widely accepted due its cheaper connection and an extended battery life. Sigfox enables new IoT applications and it can provide backup connectivity for higher bandwidth devices.

- *SigFox Network Architecture:*Communication using Sigfox can be performed when it detects an event or measure something, it will power on the communication module and send the message. The message is then picked up by the network and the data is received on your server. Sigfox is designed to maximize the energy efficiency. Sigfox consumes very low power when it transmits a data and no maintenance is required.

- *Security:*In Sigfox security challenge is addressed through a systematic process. Sigfox is one of the most secure LPWAN technology and it is unique in design. Sigfox devices predominantly operate in offline with a built-in behavior. When Sigfox needs to transmit or receive data from the internet, the Sigfox device will broadcast a radio message. This broadcasted message is received by base stations and the message is then transmitted to Sigfox core network, which is then delivered to corresponding IoT applications. This Sigfox network architecture provides an air gap and it is not possible to access an end through internet maliciously.

Figure. 10. SigFox Network Architecture
(http://www.rfwireless-world.com/Tutorials/Sigfox-network -architecture.html)

E. Zigbee

Zigbee system structure consists of three different types of devices such as Zigbee coordinator, Router and End device. Every Zigbee network must consist of at least one coordinator which acts as a root and bridge of the network. The coordinator is responsible for handling and storing the information while performing receiving and transmitting data operations.

- *Zigbee Network Architecture:*Zigbee routers act as intermediary devices that permit data to pass to and from through them to other devices. End devices have limited functionality to communicate with the parent nodes such that the battery power is saved as shown in the figure. The number of routers, coordinators and end devices depends on the type of network such as star, tree and mesh networks.

- *Security:*The ZigBee standard includes complex security measures to ensure key establishment, secure networks, key transport and frame security. Those services are implemented at the Network and the Application Support Sublayer (APS), a sub layer of the Application Layer. The ZigBee protocol is based on an "open trust" model. This means all protocol stack layers trust each other. Therefore, cryptographic protection only occurs between devices. Every layer is responsible for the security of their respective frames.

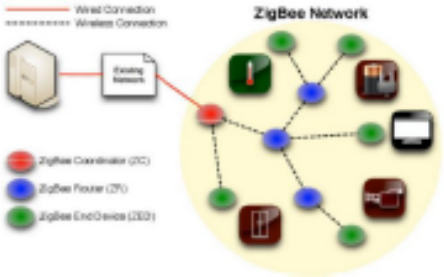


Figure.11. Zigbee Network Architecture
(https://www.elprocus.com/what-is-zigbee-technology-architecture-and-its-applications/)

IV. RANSOMWARE

Unlike traditional malware threats, a ransomware attack in IoT can be more devastating as it may affect an entire landscape of security services i.e., confidentiality, integrity, and availability, which may not only result in financial losses but may also result in an important information breach [5]. A ransomware may take entire control of data or a system and allow limited access for user interaction with the devices, ask for a hefty sum as a ransom, and release data to the user only after successful payment. In case a user does not pay, ransomware either extends the payment periods and ransom amount or deletes the data from the devices [6]. Initially, ransomware was named "AIDS", as reported in 1989, when Joseph Popp distributed 20,000 infected floppy disk drives to the participants of World Health Organizations' AIDS conference [8]. AIDS monitored the systems and counted number of times for which the systems were rebooted. It used to either encrypt data files or hide directory folders in C drive of infected computers. AIDS used to silently stay in the systems and get activated after a system reboots for 90 times. A typical ransomware attack scenario involves infection of victim computer through penetration of an attack vector whereby the malware resulting from the attack contains a payload that, unbeknownst to the victim, engages in rendering important files as unusable, through their encryption with a key that is unknown to the victim [9]. Upon completion of the initial silent encryption phase, the original unencrypted files are deleted, and the victim is alerted that their files are now inaccessible and will remain so until a ransom is paid [10].

A. Crypto Ransomware

- Crypto-ransomware is a type of harmful program that encrypts files stored on a computer or mobile device to extort money. Encryption 'scrambles' the contents of a file, so that it is unreadable. To restore it for normal use, a decryption key is needed to 'unscramble' the file. Crypto-ransomware essentially takes the files hostage, demanding a ransom in exchange for the decryption key needed to restore the files.
- Unlike other threats, crypto-ransomware is neither subtle nor hidden. Instead, it prominently displays lurid messages to call attention to itself, and explicitly uses shock and fear to pressure you into paying the ransom. A few so-called crypto-ransomware do not perform the encryption at all, and just use the threat of doing so to extort money. In most cases however, the threat is carried out.

B. Locker Ransomware

- Locker ransomware is a virus that infects PCs and locks the user's files, preventing access to data and files located on the PC until a ransom or fines are paid. Locker demands a payment of $150 via Perfect Money or is a QIWI Visa Virtual Card number to unlock files. This variant affects Windows including Windows XP, Windows Vista, Windows 7, and Windows 8. Locker ransomware is a copycat of another very nasty ransomware that has infected over 250,000 computer systems named Crypto Locker. Although the Locker ransomware is simple, it can pack a devastating blow to one's computer.

C. Hybrid Ransomware

- Hybrid ransomware attacks that enable encryption and locking mechanisms are more dangerous because the device data and functionality could be compromised. A hybrid ransomware attack could become more vicious because it can target front-end and back-end IoT devices and systems.

V. MACHINE LEARNING ALGORITHMS

First, the overview of the machine learning field is discussed, followed by the description of methods relevant to this study. These methods include outperforms K-Nearest Neighbors, Neural Networks, Support Vector Machine and Random Forest, in terms of accuracy rate, recall rate and precision rate. The rapid development of data mining techniques and methods resulted in Machine Learning forming a separate field of Computer Science. The basic idea of any machine learning task is to train the model, based on some algorithm, to perform a certain task: classification, cauterization, regression, etc. Training is done based on the input dataset, and the model that is built is subsequently used to make predictions. The output of such model depends on the initial task and the implementation.

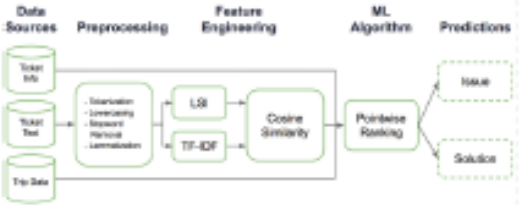


Figure.12. General Workflow Of The Machine Learning Process (https://eng.uber.com/cota/)

A. Supervised and Unsupervised Learning

- There are two machine learning approaches - supervised and unsupervised learning. In Supervised Learning, learning is based on labeled data. In this case, we have an initial dataset, where data samples are mapped to the correct outcome. The model is trained on this dataset, where it" knows" the correct results. In contrast to Supervised Learning, in Unsupervised Learning, there is no initial labeling of data. Here the goal is to find some pattern in the set of unsorted data, instead of predicting some value.

B. Classification Methods

- From machine learning perspective, Ransomware detection can be seen as a problem of classification or cauterization: unknown Ransomware types should be cauterized into several clusters, based on certain properties, identified by the algorithm. On the other hand, having trained a model on the wide dataset of malicious and benign files, we can reduce this problem to classification. For known Ransomware families, this problem can be narrowed down to classification only – having a limited set of classes, to one of which Ransomware sample certainly belongs, it is easier to identify the proper class, and the result would be more accurate than with cauterization algorithms.

- ***K-Nearest Neighbors***: K-Nearest Neighbors (KNN) is one of the simplest, though, accurate machine learningalgorithms. KNN is a non-parametric algorithm, meaning that it does not make any assumptions about the data structure. KNN can be used for both classification and regression problems. In both problems, the prediction is based on the k training instances that are closest to the input instance. In the KNN classification problem, the output would be a class, to which the input instance belongs, predicted by the majority vote of the k closest neighbors.

Figure. 13. KNN Example
(https://medium.com/@adi.bronshtein/a-quick-introduction-to-k-nearest-neighbors-algorithm-62214cea29c7)

$$Hamming\ Distance:\ d_{ij} = \sum_{j=1}^{n} |x_{ik} - x_{jk}|$$

$$Manhattan\ Distance:\ d_1(p,q) = \|p - q\|_1 = \sum_{i=1}^{n} |p_i - q_i|$$

$$Minkowski\ Distance = \left(\sum_{i=1}^{n} |x_i - y_i|^p\right)^{1/p}$$

- The most used method for continuous variables is generally the Euclidean Distance, which is defined by the formulae below:

$$EuclidianDistance = \sqrt{\sum_{i=1}^{n} (q_i - p_i)^2}\ ;\ p\ and\ q\ are\ the\ points\ in\ n-space$$

- Euclidian distance is good for the problems, where the features are of the same type. For the features of different types, it is advised to use. The value of k plays a crucial role in the prediction accuracy of the algorithm. However, selecting the k value is a non-trivial task. Smaller values of k will most likely result in lower accuracy, especially in the datasets with much noise, since every instance of the training set now has a higher weight during the decision process. As a general approach, it is advised to select k using the formula below:

$$k = \sqrt{n}$$

- ***Support Vector Machines:*** Support Vector Machines (SVM) is another machine learning algorithm that is generally used for classification problems. The main idea relies on finding such a hyperplane, that would separate the classes in the best way. The term 'support vectors' refers to the points lying closest to the hyperplane, that would change the hyperplane position if removed. The distance between the support vector and the hyperplane is referred to as margin.

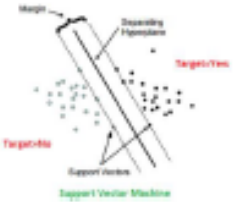


Figure.14. SVM Example
(http://www.1progn.info/support-vector-machine-algorithm-389128593a10a109ec0dd4/)

- In the above figure, there is a dataset of two classes. Therefore, the problem lies in a two-dimensional space, and a hyperplane is represented as a line. In general, hyperplane can take as many dimensions as we want.

- **The algorithm can be described as follows:**
  - We define *X* and *Y* as the input and output sets respectively. $(x1, y1), \ldots, (xm, ym)$ is the training set.
  - Given x, we want to be able to predict y. We can refer to this problem as to learning the classifier y=f(x, a), where a is the parameter of the classification function.
  - F(x, a) can be learned by minimizing the training error of the function that learns on training data. Here, L is the loss function, and *Remp* is referred to as empirical risk.

$$R_{emp}(a) = \frac{1}{m}\sum_{i=1}^{m} l(f(x_i, a), y_i) = Training\ Error$$

- We are aiming at minimizing the overall risk, too. Here, P(x,y) is the joint distribution function of x and y.

$$R(a) = \int l(f(x, a), y)dP(x, y) = Test\ Error$$

- We want to minimize the Training Error + Complexity term. So, we choose the set of hyperplanes, so f(x) = (w□x)+b:

$$\frac{1}{m}\sum_{i=1}^{m} l(w \cdot x_i + b, y_i) + ||w||^2 \text{ subject to } min_i|w \cdot x_i| = 1$$

- SVMs are generally able to result in good accuracy, especially on" clean" datasets. Moreover, it is good with working with the high-dimensional datasets, also when the number of dimensions is higher than the number of the samples. However, for large datasets with a lot of noise or overlapping classes, it can be more effective.
- ***Random Forest:***Random Forest is one of the most popular machine learning algorithms. It requires almost no data preparation and modelling but usually results in accurate results. More specifically, Random Forests are the collections of decision trees, producing a better prediction accuracy.That is why it is called a "forest" – it is basically a set of decision trees. The basic idea is to grow multiple decision trees based on the independent subsets of the dataset. At each node, n variables out of the feature set are selected randomly, and the best split on these variables is found.
  - Multiple trees are built roughly on the two third of the training data (62.3%). Data is chosen randomly.
  - Several predictor variables are randomly selected out of all the predictor variables. Then, the best split on these selected variables is used to split the node. By default, the amount of the selected variables is the square root of the total number of all predictors for classification, and it is constant for all trees.
  - Using the rest of the data, the misclassification rate is calculated. The total error rate is calculated as the overall out-of-bag error rate.
  - Each trained tree gives its own classification result, giving its own" vote". The class that received the most" votes" is chosen as the result.

- ***Neural Network:*** Neural networks are one type of model for machine learning; they have been around for at least 50 years. The fundamental unit of a neural network is a node, which is loosely based on the biological neuron in the mammalian brain. The connections between neurons are also modelled on biological brains, as is the way these connections develop over time (with "training").

VI. USE CASE

A. Discovery & Attack Surface Assesment

- As connected IoT devices communicate, usually via HTTP initially, it's possible to continuously, automatically mine the involved user agent strings within the traffic, even if the devices communicate over nonstandard TCP ports, in order to discover these active devices and understand their device type, make, model, version, etc.

B. Detection

Continuously analysing the network traffic generated by IoT devices can help to identify potential compromise and misuse [11]. Here are some examples of traffic characteristics that can be monitored for:

- Repeated login attempts to IoT devices, which are using factory default credentials (e.g. admin/admin). This may indicate an attacker or malicious process is attempting to gain access via brute force.
- If IoT devices are moved to their own dedicated network segment / VLAN, monitor for unexpected traffic between the IoT devices and your other internal hosts on the main internal network.
- Outbound SOCKS proxy traffic from IoT devices.
- Communication from IoT devices to public IPs / Domains known to be associated with botnets, command and control, proxy usage, crypto mining, etc.
- The SSL/TLS certificates used to establish secure connections to public servers.
- Dynamically generated domains being looked up from IoT devices.
- Changes in traffic rates and patterns from an established baseline.

As noted earlier, IoT devices must communicate across the network to function and attackers must communicate across the network to repurpose them. With network traffic analysis, security analysts have the means to identify devices of interest and examine the full extent of their activity, even amongst the din of an internet of things.

C. Metrics And Cross Validation

We use the following four common performance indicators for malware detection:

- True positive (TP): indicates that a ransomware is correctly predicted as a malicious application.
- True negative (TN): indicates that a goodware is detected as a non-malicious application correctly.
- False positive (FP): indicates that a goodware is mistakenly detected as a malicious application.
- False negative (FN): indicates that a ransomware is not detected and labelled as a non-malicious application.

To evaluate the effectiveness of our proposed method, we used machine learning performance evaluation metrics that are commonly used in the literature, namely: Accuracy, Recall, Precision and AUC. *Accuracy* is the number of samples that a classifier correctly detects, divided by the number of all ransomware and goodware applications:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

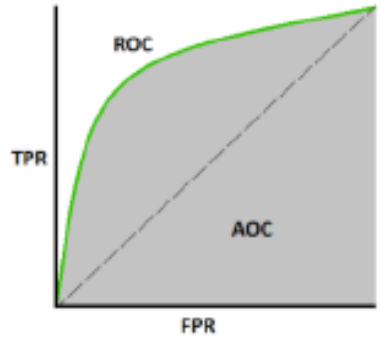
Recall or detection rate is the ratio of ransomware samples that are correctly predicted, and is defined as follows:

$$Recall = \frac{TP}{TP + FN}$$

Precision is the ratio of predicted ransomware that are correctly labelled a malware. Thus, Precision is defined as follows:

$$Precision = \frac{TP}{TP + FP}$$

AUC (Area Under the Curve) represents the probability that a true positive (green) example is positioned to the right of a true negative (red) example. AUC ranges in value from 0 to 1. A model whose predictions are 100% wrong has an AUC of 0.0; one whose predictions are 100% correct has an AUC of 1.0.



- ***Performance of Algorithms***: We will now use the leave-one-out technique for cross validation. Below figure illustrate the network traffic usage graph due to ransomware.

D. Data Flow

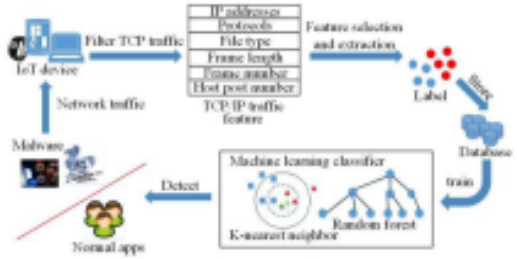
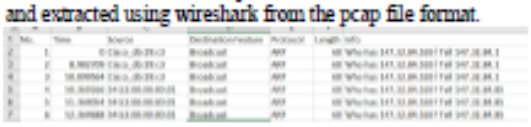


Figure. 14. Data Flow Diagram

(http://www.1progn.info/support-vector-machine-algorithm-389128593a10a109ec0dd4/)

The IoT device filters the TCP packets and selects the features among various network features including the frame number and length, labels them and stores these features in the database. The K-NN based ransomware detection assigns the network traffic to the class with the largest number of objects among its K nearest neighbours. The random forest classifier builds the decision trees with the labelled network traffic to distinguish ransomware. According to the experiments in [14], the true positive rate of the K-NN based ransomware detection and random forest-based scheme with dataset are 93% and 69%, respectively. In a ransomware detection scheme as developed in [15], an IoT device can apply the Q-learning to achieve the optimal offloading rate without knowing the trace generation and the radio bandwidth model of the neighbouring IoT devices.

- Network Data Set collected from the IOT devices in pcap file format. Below is a sample of the data which was converted and extracted using wireshark from the pcap file format.



Threats are clustered according to attack type: Denial of service (DoS), where some resource is swamped, causing DoS to legitimate users. Probes, gathering network information to bypass security. The data set is divided into two sets while executing the model as train and test. The division of the data is chosen in a way most demanding for classifiers. Train and test data do not show the same attack probability distribution; moreover, 16 out of the 38 labelled threats are only present in the test data set.

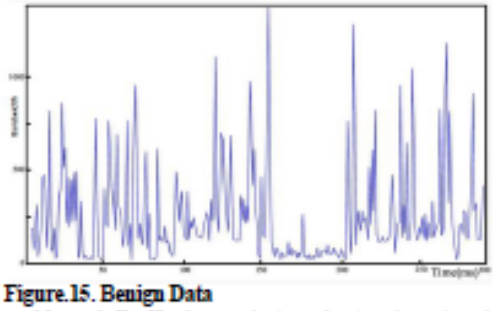


Figure.15. Benign Data

- Network Traffic data analysis on benign data given below. [Fig-15]

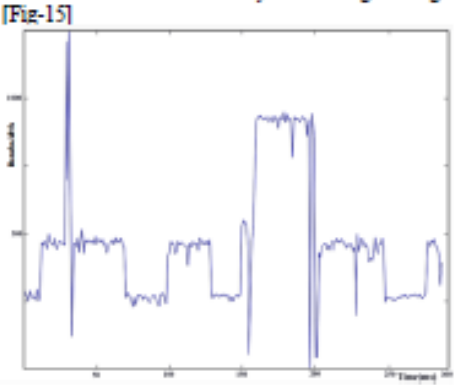


Figure.16. Infected Data with Ransomware

- Network Traffic data analysis on infected data with Ransomware given above. [Fig-16].

Comparison of the above two figures reflects a significant difference between patterns of network usage for ransomware versus benign applications.

- ***Performance of Machine Learning Techniques: A Comparative Summary***

Model	Accuracy %	Recall %	Precision %	AUC %	F-Measure %
KNN (K = 1)	72.08	73.18	57.6	67.53	68.64
KNN (K = 5)	72.82	71.29	60.48	68.51	68.63
KNN (K = 10)	72.65	73.18	58.1	67.83	68.94
KNN (K = n)	88.98	70.86	75.42	81.12	77.28
Neural Network	76.16	74.6	63.14	71.79	67.9
Random Forest	81.97	77.74	70.53	77.41	79.53
SVM	78.75	75.51	67.15	76.57	78.68

- Evaluation Metrics for different Window Sizes, KNN and Euclidean distance: A Comparative Summary given below.

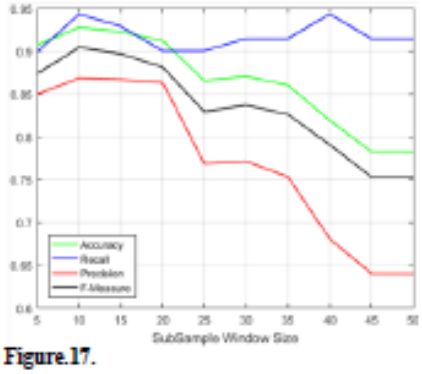


Figure.17.

- Evaluation metrics for different window sizes, KNN and n distance: a comparative summary

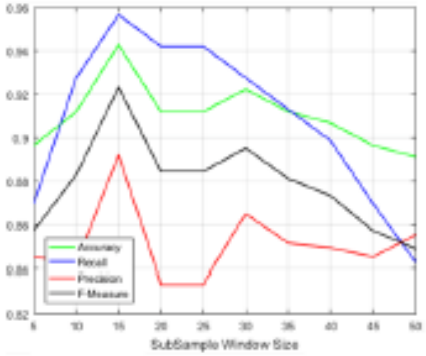


Figure.18.

Furthermore and in practice, KNN's requirement for concurrent distance calculations between training and testing objects can be implemented using parallel processing (so distances can be independently computed). Subsamples dictionary can be partitioned into sperate IoT nodes and each subsample is sent to nodes. They return a label and a similarity value and the label having less similarity value is final subsample's label. This approach reduces the classification time and mitigates the need for storage capacity in every node.

VII. FUTURE DIRECTIONS OF IOT TECHNOLOGIES IN SMART CITY

IoT technologies contribute significantly to the majority of the detailed aspects of smart city technologies and infrastructure. Because the fundamental concepts and ideas of IoT technologies are shared with those of smart city technologies and infrastructure, a large number of business opportunities and extensive growth potential exist. Moreover, for the efficient and successful development of future IoT technologies in smart cities, the following points require focus.

Importance and essentiality: Sensor-oriented technologies for wireless networking are considered the top priority of IoT technologies for a smart city infrastructure.

- **Importance:** In addition to sensor-oriented technologies, technologies for network services are considered the most important IoT technologies for a smart city infrastructure.
- **Essentiality:** Compared to other technologies, energy-related technologies are considered the most essential IoT technologies for the smart city infrastructure. Because the technologies applied in a smart home environment are fundamental aspects of a smart city, the technologies and infrastructure for a smart home network should be swiftly developed and prepared.

In addition to these points, there are some challenging aspects that should also be resolved. First, because IoT technologies should provide various operating systems, which includes low to high-capacity processors, providing appropriately distributed resources is one of the most important tasks required in devices employing IoT technologies. Second, data management solutions are required for the massive amounts of data collected by various IoT technology devices because the majority of such data is unstructured or atypical. This means that technologies for data categorization and intelligent analysis should be developed and introduced. Third, the current IoT technology services are provided through independent specialized solutions, which are oriented and operated within a specific environment. Therefore, compatible integrated Sustainability applications for providing various IoT technology services should be developed and prepared by using appropriate network technologies. Fourth, both appropriate solutions and plans for data security and privacy should be established. When users connect to an IoT technology service, reliable data processing and storage should be applied with confidentiality, integrity, and privacy. This means that reliable and safe communications and connections from each IoT technology device to the smart city infrastructure should be provided.

VIII. CONCLUSION

With increasing prevalence of Internet-connected devices and things in our data-centric society, ensuring the security of IoT networks is vital. Successfully compromised IoT nodes could hold the network to ransom. For example, in the case of ransomware, denying availability to data in an IoT network could adversely affect the operation of an organisation and result in significant financial loss and reputation damage. In this paper, we presented an approach to detect ransomware, using network traffic flow analysis. Specifically, we utilize the unique local fingerprint of ransomware's network usage pattern to distinguish ransomware from non-malicious applications. Our set of experiments demonstrated that our approach achieved a detection rate of 93.76% and a precision rate of 89.85%. We have shown that ML is a viable and effective approach to detect new variants and families of ransomware for subsequent analysis and signature extraction, and as a complement forAV. Mutual Information has shown to be an effective way of automatically selecting the features, while Regularized K Nearest Neighbors has shown to be an accurate algorithm, easy to train and update, and fast.

IX. REFERENCES

[1] [Online] Available: https://iot-analytics.com/internet-of-things-definition/.

[2] [Online] Available: http://www.gartner.com/newsroom/id/3165317.

[3] Tankard, Colin. "The security issues of the Internet of Things." Computer Fraud & Security 2015, no. 9 (2015): 11-14.

[4] Watson S, Dehghantanha A (2016) Digital forensics: the missing piece of the internet of things promise. Comput Fraud Secur 2016(6):5–8

[5] O'Gorman G, McDonald G (2012) Ransomware: a growing menace. Tech. rep., Symantec Corporation. http://www.symantec.com/content/en/us/enterprise/media/security_response/whitepapers/ransomware-a-growing-menace.pdf. Accessed 12 Feb 2017

[6] Caviglione L, Gaggero M, Lalande JF, Mazurczyk W, Urbański M (2016) Seeing the unseen: revealing mobile malware hidden communications via energy consumption and artificial intelligence. IEEE Trans Inf Forensics Secur 11(4):799–810

[7] Jonas Olsson, Texas Instrument, 6LoWPAN demystified, http://www.ti.com/lit/wp/swry013/swry013.pdf

[8] E. Bertino and N. Islam, "Botnets and internet of things security," Computer, vol. 50, pp. 76–79, 2017.[24] B. Nassi,

A. Shamir, and Y. Elovici, "Oops!... i think I scanned a malware," arXiv preprint arXiv:1703.07751, 2017.

[9] R. Richardson and M. North, "Ransomware: Evolution, mitigation and prevention," International Management Review, vol. 13, no. 1, p. 10, 2017.

[10] The rise of ransomware and emerging security challenges in the Internet of Things. Available from: https://www.researchgate.net/publication/319527564_The_rise_of_ransomware_and_emerging_security_challenges_in_the_Internet_of_Things [accessed Oct 28, 2018].

[11] [Online] https://www.corvil.com/blog/2018/discovering-iot-devices-and-monitoring-their-network-traffic

[12] Haykin S (1998) Neural networks: a comprehensive foundation, 2nd edn. Prentice Hall, 20(8):2481–2501 Kim H, Smith J, Shin KG (2008) Detecting energy-greedy anomalies and mobile malware variants. In: Proceedings of the 6th international conference on mobile systems, applications, and services. ACM, pp 239–252

[13] Kohavi R et al (1995) A study of cross-validation and bootstrap for accuracy estimation and model selection. Ijcai 14:1137–1145 and mobile malware variants. In: Proceedings of the 6th international conference on mobile systems, applications, and services.

[14] F. A. Narudin, A. Feizollah, N. B. Anuar, and A. Gani, "Evaluation of machine learning classifiers for mobile malware detection," Soft Computing, vol. 20, no. 1, pp. 343–357, Jan. 2016.

[15] L. Xiao, Y. Li, X. Huang, and X. J. Du, "Cloud-based malware detection game for mobile devices with offloading," IEEE Trans. Mobile Computing, vol. 16, no. 10, pp. 2742–2750, Oct. 2017.

Paper -2: Auto-Detection of Click-Frauds using Machine Learning algorithms

https://ijesc.org/upload/c85b77baa8fb8e66d83d5a80fd11a744.Auto-Detection%20of%20Click-Frauds%20using%20Machine%20Learning.pdf

IJESC INTERNATIONAL JOURNAL OF ENGINEERING SCIENCE AND COMPUTING

ISSN 2250-1371
www.ijesc.org

Certificate of Publication

Awarded to

Satyajit Pal

(MBA in Business Analytics REVA Academy for Corporate Excellence, REVA University, Bengaluru, India)

For publishing Research Article in International Journal of Engineering Science and Computing (IJESC)

Volume 10 Issue No.9, September 2020

Manuscript titled

"Auto-Detection of Click-Frauds Using Machine Learning"

Dr. Aneesh Thomas
Editorial Head, IJESC

ISSN 2321 3361 © 2020 IJESC

IJESC

Research Article Volume 10 Issue No.9

Auto-Detection of Click-Frauds using Machine Learning

Anshuman Dash[1], Satyajit Pal[2]
MBA in Business Analytics
REVA Academy for Corporate Excellence, REVA University, Bengaluru, India

Abstract:
In the current web advertising activities, the fraud increases the number of risks for online marketing, advertising industry and e-business. The click-fraud is considered one of the most critical issues in online advertising. On-line advertisement has become one of the most important funding models to support Internet sites. Given that large sums of money are involved in on-line advertisement; malicious parties are unfortunately attempting to gain an unfair advantage. Even if the online advertisers make permanent efforts to improve the traffic filtering techniques, they are still looking for the best protection methods to detect click-frauds. Click-Fraud occurs by intentional clicking of online advertisements with no actual interest in the advertised product or service. Click Fraud is an important threat to advertisement world that affects the revenue and trust of the advertisers also. Click-fraud attacks are one instance of such malicious behavior, where software imitates a human clicking on an advertisement link. Hence, an effective fraud detection algorithm is essential for online advertising businesses. The purpose of our paper is to identify the precision of one of the modern machine learning algorithms in order to detect the click fraud in online environment. In this paper, we have studied click patterns over a dataset that handles millions of clicks over few days. The main goal was to assess the journey of a user's click across their portfolio and flag IP addresses who produce lots of clicks, but never end up in installing apps. We have focused on the issue while using various single and ensemble-typed classification algorithms for the fraud detection task. As our single classifiers, we employed the Support Vector Machine, kNN algorithms. We have also employed decision tree-based ensemble classifiers, which have been used in data mining. These algorithms are Random Forest and Gradient Tree Boosting.

Keywords: Click-Fraud Detection, Advertisements, Internet Spammers, Machine learning, Ensemble Models.

I. INTRODUCTION

Online marketing has exposed the world to everyone. Where small companies were struggling to impact in the local areas once, now-a-days the world has become very small while using the concepts of pay per click and digital marketing tools [1]. More than "4 billion people use internet on daily basis and more than 2 billion people" use internet for shopping online. A targeted pay per click campaign is the difference between sinking and swimming as more than 5 billion clicks happen in Google every day. But there are always more than a few rats in any busy marketplace. Click fraud is one of the most harmful and successful practices in the online marketplace[2]. This technique works by manipulating your PPC campaigns, causing you to lose money, miss valuable sales opportunities, and possibly even destroy your business [3]. There is an entire industry that has been set up to defraud web marketers and consumers. Some mischievous ones, such as hackers; some created for the profit of another group fraudulently, some deliberately vindictive and with the intention of stealing ads from certain networks. By default, click fraud does not produce an advertiser's profits, but losses "hundreds of millions of dollars" a year to "tens of thousands" of online advertisers [4]. Normally, malicious applications (apps) and malware produce click fraud and account for about "30% of click traffic in ad networks". The number of click frauds has increased significantly with mobile malware. Fraudsters obviously create legitimate apps or buy respectable men [5]. Such applications perform a legitimate operation, like torch control, but also function as a tool to undermine the clicking behaviour of the user of the computer. In addition, attackers laundered clicks again via their installed user base [6]. As click fraud is based on valid traces, ad-network filters may pass through the clicks. Exclude the use of a small pool of IP addresses to execute the attack. The attack violates a threshold, for example. This ultimately leads to the need for automated techniques for detecting click scams, thus guaranteeing the credibility of the digital advertising ecosystem [7].

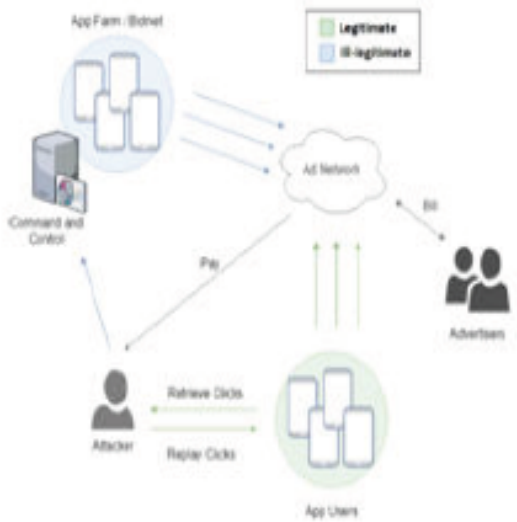


Figure.1.Legitimate & Ill-legitimate Click Fraud

A. History of Advertisements Click Business

In an online advertising market, advertisers pay ad networks for each click on their advertisements, and ad networks pay publishers a share of the revenue [8]. When online advertising has grown into a multi-billion dollar business, click fraud has become a serious and widespread problem. For example, the "Chameleon" botnet infected more than 120,000 host machines in the U.S. and siphoned $6 million a month [9].

Click fraud occurs when miscreants make HTTP requests for destination URLs found in the ads being deployed. Such HTTP requests with malicious intent are called fraudulent clicks. The motive for fraudsters is to increase their own income to the detriment of other parties [10]. A fraudster is typically a publisher or an advertiser. Publishers may place excessive advertising banners on their sites and then fake clicks on the ads to get more money. Unscrupulous advertisers are clicking heavily on a competitor's advertisements in order to deplete the victim's advertising budget. Click fraud is mainly done by using click bots, recruiting human clickers, or tricking users into clicking ads [11].

Figure.2. Advertisements Click Business (https://www.digitalvidya.com/blog/what-is-ppc/)

[Click fraud is not trivial. Click fraud systems have been growing continuously in recent years[12–15]. Existing detection approaches aim to classify click fraud behaviours from different perspectives, but each has its own limitations. The solutions suggested in [16–19] conduct a traffic analysis on ad network traffic logs to detect publisher inflation fraud. Nonetheless, an advanced clickbot can perform a low-noise attack, which makes these unusual behavioural detection mechanisms less successful.]

B. Examples Of Click-Fraud Attacks

Major search engines such as Google and Bing are aware of how serious click fraud detection is. Back in 2005, Lane's Gifts & Collectibles sued Google along with Yahoo! and Time Warner in a collective action case resulting in $90 million settlement with an agreement to improve their tracking and identification of fraudulent clicks [20]. While things have definitely improved in the past 10 years, every PPC advertiser–or ad network–probably thinks that the problem is gone. Detecting search engine click fraud such as Google and Bing means that the big moneymakers in the industry secure their advertisers and the entire network.

C. Purpose

The goal of this project is to build an adaptive and scalable feature for Rich in fraud detection. This component is able to deal with the large quantity of data that is downgraded via the system and to provide output to improve the accuracy of the reports produced.

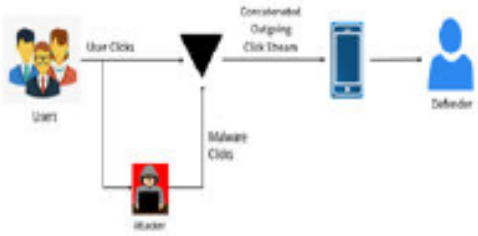


Figure.3. Click-Fraud Detection Problem

D. Overview

The paper provides major application of "machine learning" and "data mining" to solve real-world issues of fraud detection as valuable resources for industry and researchers. So far, the "data mining / machine learning" approach to fraud detection in ads has not been thoroughly studied. This research includes university-based data, which are collected over 1 month and present many data mining and machine learning algorithms with a difficult problem [21]. The solutions presented in this report answer some important questions in data mining and machine-learning science, including a highly imbalanced output variable distribution, heterogeneous data (mixing number and class variables) and noisy patterns of missing / unknown values. The analysis and feature engineering of exploratory data were shown to be crucial milestones for the detection of fraud. In general, there has been a systematic study of spatial and temporal factors at various granularity rates leading to the creation of nice, predictive characteristics to detect specific fraud [22]. A wide range of algorithms for single and ensemble learning have been tested in the detection of fraud, with a significant improvement over the single algorithms [23]. Coupling ensemble learning with evaluation of the feature rating often shows the key features to differentiate fraudulent from ordinary. In this paper, the overview of the captured dataset, challenges, and evaluation procedures have been presented.

II. THEORY

A. Terms & Concepts

•**Click-through And Click-Through Rate:** CTR stands for the click-through level of Internet marketing: a measure calculating the number of click-throughs that advertisers earn per experience. Achieving a high click rate is crucial for the success of Pay-Per-Click, as it affects both value and compensation at any time anyone clicks on an ad request [24].The rate at which your PayPal-Click advertisements are clicked is the click-through rate. This number represents the proportion of people who watch announcements (impressions) and then click on the ad. The click rate can usually be viewed on the PPC account dashboard. This is the formula for CTR:

(Total Clicks on Ad) / (Total Impressions) = Click Through Rate

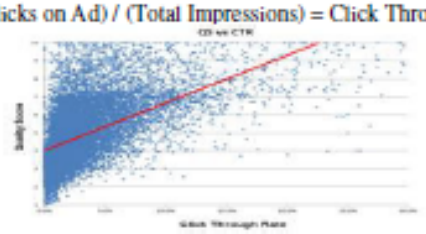


Figure.4. Click Through Rate (https://www.wordstream.com/click-through-rate)

•**Pay-Per-Click:** This marketing is an advertisement network in which advertisers do not pay for printing or ad positioning alone [25]. The bid can impact positioning, but only when an advertiser clicks on an online client. The advertiser charges. On search results pages of search engines such as Google and Bing the most popular PPC ad format appeared. Advertisers may position their brand, product or service in the form of an ad to a specific keyword or behavior [26].

•**Google AdWords:** It is Google's advertising service for companies wishing to show ads on the Google network. The AdWords program allows businesses to set an advertising budget and charge only by clicking on the ads [27]. The ad network concentrates mainly on keywords. Corporate users of AdWords may build advertisements with keywords that will be used by people searching the Internet through the Google search engine. The keyword will show your ad when it is checked. AdWords in the top marketing headings that appear on the right or above Google search results under the heading "Sponsored Links." Google search users are then forwarded to your website if your AdWords ad is clicked upon.

•**Click Fraud:** It is an unethical practice when individuals click an ad from a page (banner advertising or paid text links) to increase the number of clicks payable to the advertiser. Click fraud is an illegal practice. Illegal clicks can either be achieved by clicking on advertisement hyperlinks by someone manually or by using automated software or programmed on-line bots to click on those banner ads to pay for text ad links per click. Research has shown that clicking fraud is committed by persons using click fraud to maximize personal banner ad profits, and businesses using click fraud to deplete the budget of a competitor's publicity. Pay-per-click ads (PPC) is commonly associated with click fraud [28].

•**Impression Fraud:** It is when an ad cannot be seen in the eye, but it still takes account of experiences. Pixel filling, ad stacking and fraudulent traffic are the most common fraudulent methods. Nevertheless, malware may also happen in mobile and fraud-creating websites.

Figure.5. Impression Fraud (https://blog.anura.io/blog/what-is-impression-fraud-and-how-does-it-work)

•**Click Bot:** It is a traffic bot breed that aims at spiking the ad count. Whenever a fraud ad is in operation, a click bot normally belongs to the crime scene. This attracts a real user who visits the site and clicks on an ad. From here, it is easy to see how the PPC project could make money bleed from this tiny piece of software [29].

III. MACHINE LEARNING ALGORITHMS

A. Classification Approach

First, the overview of the machine learning field is discussed, followed by the description of difference between unsupervised and supervised classification and relevant methods. These methods include outperforms K-Nearest Neighbors classification, Classification Trees, Support Vector Machine, Random Forest and Gradient Tree Boosting, in terms of accuracy rate, recall rate and precision rate. The rapid development of data mining techniques and methods resulted in Machine Learning forming a separate field of Computer Science. The basic idea of any machine learning task is to train the model, based on some algorithm, to perform a certain task: classification, cauterization, regression, etc. Training is done based on the input dataset, and the model that is built is subsequently used to make predictions. The output of such model depends on the initial task and the implementation.

Figure.6. General Workflow of The Machine Learning Process (https://tibacademy.in/machine-learning-training-in-marathahalli/)

B. Supervised and Unsupervised Learning

There are two approaches to machine learning-supervised and unsupervised learning. Learning is based on labelled data in supervised training. There is an initial dataset in this case, in which data samples are mapped to the correct result.On this dataset the model is trained, where "the correct results are known." Unlike Supervised Learning, there is no initial data labelling in Unsupervised Learning. Instead of predicting a certain value, the aim is to find some pattern in a set of unsorted data.

C.Classification Methods

The question of classification or cauterization can be seen from a machine learning point of view: unidentified click-fraud forms are cautioned into several clusters based on specific algorithmic characteristics. On the other hand, we can reduce this problem to classification after training a model with the large dataset of malicious and benign files. This issue can be reduced for known click-fraud types to classifications with only a limited group of classes, which certainly include the click-fraud model, which can be used more easily to identify the correct class, and result is more accurate than with cauterization algorithms.

D. K-Nearest Neighbors

[K-Nearest Neighbours (KNN) is one of the simplest, though, accurate machine learning algorithms. KNN is a non-parametric algorithm which means that the data structure is not assumed. For classification problems as well as regression problems, KNN can be used. The prediction in both cases is based on the instances of k training which are closest to the input example. The result would be a group, to which the input instance belongs, foreseen by the majority of the votes of k closest neight.]

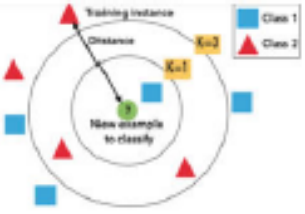


Figure.7. KNN Example (https://medium.com/@adi.bronshtein/a-quick-introduction-to-k-nearest-neighbors-algorithm-62214cea29c7)

$$Hamming\ Distance:\ d_H = \sum_{i=1}^{k} |x_{1i} - x_{2i}|$$

$$Manhattan\ Distance:\ d_1(p,q) = \|p-q\|_1 = \sum_{i=1}^{n} |p_i - q_i|$$

$$Minkowski\ Distance = \left(\sum_{i=1}^{n} |x_i - y_i|^p\right)^{1/p}$$

The most used method for continuous variables is generally the Euclidean Distance, which is defined by the formulae below:

$$EuclidianDistance = \sqrt{\sum_{i=1}^{n} (q_i - p_i)^2}\ ;\ p\ and\ q\ are\ the\ points\ in\ n-space$$

[Euclidian distance is good for the problems, where the features are of the same type. For the features of different types, it is advised to use. The value of k plays a crucial role in the prediction accuracy of the algorithm. However, selecting the k value is a non-trivial task. Smaller values of k will most likely result in lower accuracy, especially inthe datasets with much noise, since every instance of the training set now has a higher weight during the decision process. As a general approach, it is advised to select k using the formula below]:

$$k = \sqrt{n}$$

E. Support Vector Machines:

Support Vector Machines (SVM) is another machine learning algorithm that is generally used for classification problems. The main idea relies on finding such a hyperplane, that would separate the classes in the best way. The term 'support vectors' refers to the points lying closest to the hyperplane, that would change the hyperplane position if removed. The distance between the support vector and the hyperplane is referred to as margin.

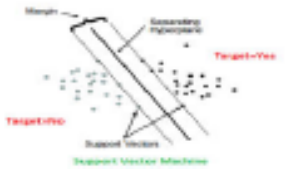


Figure.8. SVM Example (http://www.1proga.info/support-vector-machine-algorithm-389128593a10a109ec0dd4/)

In the above figure, there is a dataset of two classes. Therefore, the problem lies in a two-dimensional space, and a hyperplane is represented as a line. In general, hyperplane can take as many dimensions as we want.

The algorithm can be described as follows:

- We define X and Y as the input and output sets respectively. $(x1,y1),\ldots,(xm,ym)$ is the training set.
- Given x, we want to be able to predict y. We can refer to this problem as to learning the classifier y=f(x, a), where a is the parameter of the classification function.
- F(x, a) can be learned by minimizing the training error of the function that learns on training data. Here, L is the loss function, and *Remp* is referred to as empirical risk.

$$R_{emp}(a) = \frac{1}{m}\sum_{i=1}^{m} l(f(x_i,a),y_i) = Training\ Error$$

We are aiming at minimizing the overall risk, too. Here, P(x,y) is the joint distribution function of x and y.

$$R(a) = \int l(f(x,a),y)dP(x,y) = Test\ Error$$

We want to minimize the Training Error + Complexity term. So, we choose the set of hyper planes, so f(x) = (w·x)+b:

$$\frac{1}{m}\sum_{i=1}^{m} l(w\cdot x_i + b, y_i) + \|w\|^2 \text{ subject to } \min_i |w\cdot x_i| = 1$$

SVMs are generally able to result in good accuracy, especially on" clean" datasets. Moreover, it is good with working with the high-dimensional datasets, also when the number of dimensions is higher than the number of the samples. However, for large datasets with a lot of noise or overlapping classes, it can be more effective.

F. Random Forest

Random Forest is one of the most popular machine learning algorithms. It requires almost no data preparation and modelling but usually results in accurate results. More specifically, Random Forests are the collections of decision trees, producing a better prediction accuracy. That is why it is called a "forest" – it is basically a set of decision trees. The basic idea is to grow multiple decision trees based on the independent subsets of the dataset. At each node, n variables out of the feature set are selected randomly, and the best split on these variables is found.

- Multiple trees are built roughly on the two third of the training data (62.3%). Data is chosen randomly.
- Several predictor variables are randomly selected out of all the predictor variables. Then, the best split on these selected variables is used to split the node. By default, the amount of the selected variables is the square root of the total number of all predictors for classification, and it is constant for all trees.
- Using the rest of the data, the misclassification rate is calculated. The total error rate is calculated as the overall out-of-bag error rate.
- Each trained tree gives its own classification result, giving its own" vote". The class that received the most" votes" is chosen as the result.

F. Classification Tree

For classification of instances, a decision tree is a simple representation. It is a supervised learning machine in which the information are separated continuously by a certain parameter.

A decision tree is a method used to support decision-making that uses a tree-like graph or template of decisions, including chance outcomes, the value of resources and utility. The decision tree is a flowchart structure in which each inner node is a "test" in a particular attribute (e.g. if the coin pad appears on the heads or tails). Every branch is the outcome of the test and every leaf node is a class tag (decision made after all attributes have been computerized). The root-to-leaf paths are category rules. One of the popular and mostly used supervised learning methods is trees-based learning algorithms. Tree-based methods allow high accuracy, stability and easy analysis of predictive models. They map non-linear relations rather well, unlike linear modelling. These are ideal for the resolution of any question (classification or regression). CART (Classification and Regression Trees) algorithms for Decision Tree. Classification tree is a predictive model that maps an item's observations to its final value conclusions. The leaves are classifications of the tree structures (also called label), features are non-leaf nodes, and branches represent conjunctions of features leading to classifications [30]. It is simple to build a decision tree that fits a particular data set. The goal is to build good decision-making bodies, usually the smallest decision-making bodies. Overfitting can be used to avoid overfitting the tree for the training set only. This technique produces the tree for unmarked data and can accommodate some erroneously labelled training data.

Figure.9. Classification Tree (https://www.digitalvidya.com/blog/classification-and-regression-trees/)

G. Gradient Tree Boosting

From the application of boosting methods to regressions trees the algorithm for boosting trees was created. The general idea is to measure a series of simple trees, in which each subsequent tree is built to estimate the remains of the previous tree. This approach constructs binary trees, i.e., divides data into two samples at every divided node. At every step in the boost (algorithm for boosting trees), the data are simply (best) partitioned and variations in the observed values (residuals for each partition) measured from the respective means are calculated. In order to find another partition that reduces the rest (error) variance for the results, given the previous trees sequence, the next three node tree will then be fitted to these rest products [31].

It is shown that such "additive weighted expansions," even though the specific nature of the relationships between the predictor variables and the dependent interest variable is very complicated (not linear in nature), can eventually lead to an outstanding match of the expected values to the observed values. Therefore, a very common and efficient learning process is the gradient boosting approach – the adaptation of a weighted, additive distribution of simple trees.

IV. USE CASE

A. Dataset

The click information containing both valid and fraudulentclick spam has been identified. First, it acquired a pre-label data set, consisting in controlled proportions both of legitimate clicks and of fraudulentclick spam. In order to achieve this, traffic click spam has been processed within the university network; it has been filtered and distributed to test beds. As a consequence, clicks from both true and false clicks comprise the traffic leaving the Testbed.

B. Dataset Collection

The traffic monitors on backbone routers of the campus university network were set up to collect legal ad-click files. The following information was recorded in the application for each click: the URL, the IP address of the ad server, the publishing page (referrer URL), the IP address of source, the User agent string, and the time stamp. In addition, between August-2019 and October-2019, a total of 32,119 unique clicks were registered. Data was collected and all stored data were encrypted following the due process of receiving ethical approval.

C. Data Preparation

The data is prepared for effective analyses after data collection. The data set obtained consists of several attributes which are not required for study, so the data should be prepared according to the requirements so that the algorithm produces accurate results. Data is prepared by the data.table kit and fusion method in this research work. The data.table kit is supported with a data.frame upgrade version. This allows the user to manipulate data extremely quickly, and is commonly used for large data sets[32]. Merge function allows two databases to be merged by calling the data.frame method based on common columns or row names. When columns have been defined, names of columns are given by.x (first file column names) and by.y (second file column names)[33]. Next, the original data set is charged and, by setting the Order Date and Product ID, the number of occurrences per velocity parameter is determined. Then the output of all events of every velocity variable is combined by the merge function with the original data collection.

D. Metrics & Cross Validation

The following four common performance indicators for click traffics detection are used in this research paper:

- [True positive (TP): indicates that a click traffic is correctly predicted as a fraudulent ad.]
- [True negative (TN): indicates that a click traffic is detected as a legitimate ad correctly.]
- [False positive (FP): indicates that a click traffic is mistakenly detected as a fraudulent ad.]
- [False negative (FN): indicates that a click traffic is not detected and labelled as a legitimate ad.]

The effectiveness of our proposed methods are evaluated by using machine learning performance evaluation metrics which are "Accuracy, Recall, Precision and AUC". Accuracy is defined as the number of samples that a classifier can correctly detect, divided by the addition of number of all ransomware and good ware applications.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

Recall value or the detection rate is the ratio of ransomware samples that are correctly predicted

$$Recall = \frac{TP}{TP + FN}$$

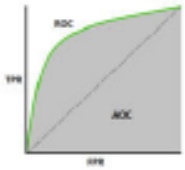
Precision is the calculated ratio of predicted ransomware that are correctly identified as a malware. Precision is defined below

$$Precision = \frac{TP}{TP + FP}$$

["AUC (Area Under the Curve) represents the probability that a true positive is positioned to the right of a true negative."] AUC ranges in values from 0 to 1. A model which predicts 100% wrong values has an AUC of 0.0 and one which predicts 100% correct values has an AUC of 1.0.

E.Performance of Algorithms

The leave-one-out technique for cross validation is used in this research paper. Below figure illustrate the network traffic usage graph due to fraudulent ads.



F. Exploratory Data Analysis

The following are the specifics of the button history and fraud tap. In this case, however, the data collection time is too short to display trends. So the attribute hour or minute is not here extracted from the time function of the click. The dataset is therefore distributed without regard to bias.

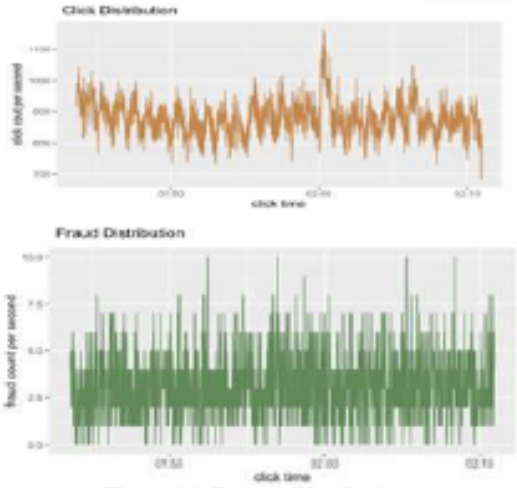


Figure.10. Dataset Distribution

The fraudulent versus non-fraudulent rate of traffic is measured as a fraudulent versus non-fraudulent proportion. Filtration speed x-axis is time, y-axis is ratio and in the time series described above indexed to ratio on the first date. The numbers show major releases of the material.

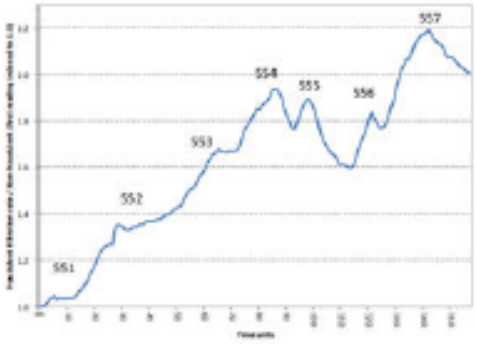


Figure.11. Fraudulent versus Non-Fraudulent Rate

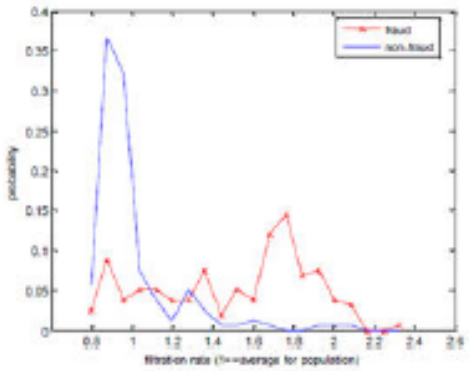


Figure.12. Filtration Rate Ratio for fraudulent versus non-fraudulent click spams

[After a rule update their filtration rates went to 100%. The time-axis shows days leading up to a model update and following the model update.]

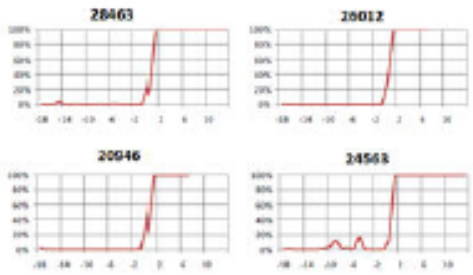


Figure.13. Filtration Rates for Four Fraudulent Click Spam

G.Evaluation

The three-phase analysis of classifier implementations has been done in order to determine the suitability of the various classification approaches for this application scenario. This was also done to address the absence of open classification studies in the area of the issue. In this section the results of the assessment are presented. During the First Evaluation Phase, one candidate

algorithm was evaluated from each approach to determine its exactness (i.e., percentage of properly classified cases) on a small number of prelabelled data. A brief choice was made of candidates who ran the majority of the available classificatory and weakened those who produced too poor results (note that not all classification systems are relevant to the type of data with which we operate, e.g. those which require strict nominal input). The analysis was performed by partitioning a collection of pre-labelling data into two separate sets used for classification learning, whether false or valid, and by assessing the effects of classifies on the pre-labelling data respectively. Rather than doing a simple percentage split, the test results were improved with a so-called n-fold cross-validation technique. A model is built with the same size n-1 partitions in the data set in n-fold cross-validation. On the remaining partition, the template is then evaluated. It is repeated n times, until each partition is used exactly once for evaluation. Listing 3 explains the cross validation algorithm. n=10 has been used for the experiments shown below.

Require: A set D of data points prelabelled with a class
$P = \{p_1, p_2, \ldots, p_n\}$, a set of equally sized partitions of D
for $i = 1$ to n **do**
  $S = \{p_i\}$
  $T = D \setminus S$
  Build a classifier c using T as the training set
  Let r_i be the result of evaluating c on test data S
end for
return The average of all results $\{r_1, r_2, \ldots, r_n\}$

There are some definitions used in the evaluation of this study before the results are reported. In the following text, a positive is the equivalent of a fraudulent instance, while a negative refers to a non-fraudulent example. A true positive is a positive statement, while a false positive is a negative that the classification evidence has been made positive. Likewise, a true negative is an advertised negative, but a false negative is a positive, which is marketed as a negative. The words used are as follows:

$$TPR\ (\text{True Positive Rate}) = \frac{\text{true positives}}{\text{true positives} + \text{false negatives}}$$

$$FPR\ (\text{False Positive Rate}) = \frac{\text{false positives}}{\text{true negatives} + \text{false positives}}$$

$$TNR\ (\text{True Negative Rate}) = \frac{\text{true negatives}}{\text{true negatives} + \text{false positives}}$$

$$FNR\ (\text{False Negative Rate}) = \frac{\text{false negatives}}{\text{true positives} + \text{false negatives}}$$

$$ACC\ (\text{Accuracy}) = \frac{\text{true positives} + \text{true negatives}}{\text{all instances}}$$

	TPR	FPR	TNR	FNR	ACCURACY
Random Forest	95.40%	14.30%	85.70%	4.60%	89.48%
Classification Trees	94.60%	7.60%	92.40%	5.68%	93.28%
Support Vector Machine	95.40%	9.10%	90.90%	4.65%	92.78%
knn Classification	69.00%	69.00%	31.00%	31.00%	45.78%
Gradient Tree Boosting	99.80%	76.90%	98.20%	23.20%	97.28%

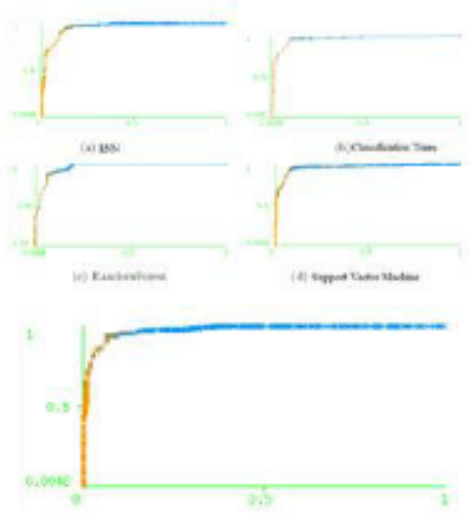
Data set size: 32119 instances (8713 fraudulent, 23406 legitimate)

Figure.14. Classifier Accuracy

A low FPR does not necessarily imply a satisfactory outcome, as the FPR must be taken into account in relation to the actual data positive part. If the actual positive percentage in comparison with the FPR is relatively low, many of the recorded positive warnings can be assumed to be incorrect. For example, it should be presumed that there is a data set of 10,000 users, 100 of whom actually showed an act of fraud (PAP= 1%) and the other 9.900 users show no fraud. The process reports correctly on average0:944 by means of the RandomForest steps. 100= users 94:4, and 0:076 wrongly. As fraudulent, 9900= 752:4 clients. As can be seen, because of the low portion of actual positive data, the number of false-classified positive. Positive elements are significantly higher than that of the correctly classified positives. Thereby, 88:9 percent of all positive reports are false alarm! It can be inferred! More specifically, the following formula can describe the portion of all positive reports that are false alarms:

$$PFA = \frac{\text{false alarms}}{\text{reported positives}} = \frac{FPR \cdot (1 - PAP)}{PAP \cdot TPR + FPR \cdot (1 - PAP)}$$

This is an overall difficulty in detecting fraud. But it's not entirely lost. It should be possible to approach more satisfactory results by rigorously tweaking the parameters of each algorithm. The accuracy of the training data should also be improved when optimized. Because there is a larger number of points in the "black" region between the two categories for the studies, a lower FPR can be predicted when the real data is graded. The PFA is still a concern, however, provided that real data includes signed instances that are considerably less fake than the data set used in those studies.



(e) Gradient Tree Boosting

Figure.15. ROC Curves for Classification Algorithms

V. FUTURE WORK

About future improvements to the process that can be made. The adaptive character of the system means that the learning data are continually improved. Nevertheless, there are additional ways of improving identification system accuracy. In this study, we covered a wide range of classification algorithms to classify who you are [34]. Performance improvements are also available. The current system bottleneck is the move to aggregate user data as shown in the above tests. In addition, this part of the system

should therefore concentrate on efforts to improve overall system performance. We have described some ideas which have been investigated but left out because the necessary data cannot be obtained (such as the analysis of premium clicks or mouse patterns). Those characteristics, such as consumer geographical location, were not included in the existing classification process [35]. To this end, training data would need to be developed for every campaign, so that a warning flag is lifted if most viewers for an ad suddenly comes from a new location. We think these ideas should be discussed further as they may be helpful input attributes to the classification system (when information can be obtained).

VI. CONCLUSION

The financing of millions of websites and mobile apps on-line ads is a template. Digital advertising with special purpose attack methods, called click malware, is constantly targeted by criminals. An important security challenge is click fraud created via malware. The state-of - the-art techniques can easily detect static attacks involving large attack volumes. Nonetheless, current methods fail to detect complex attacks involving steady click-spam that match the app user's actions. Timing analysis has been found to have a crucial role to play in isolating click scams, both static and dynamic. This research paper applies a technique that detects click-spam using relative uncertainty between click-spam and valid clicks-streams. It does this by identifying repeated patterns from valid click-spam in the ad network. A malware corpus is also analysed in an instrumented environment which can handle click-spam generation by exposing malware to legitimate click-spams. We have tested a passive technique that is promising. An effective protection has also been tested, wherein the analytical system is better functioning when injecting watermarked click traffic. Although timing analysis has been well studied for its ability to discover supernatural interaction in the field of data hiding, its potential still has to be fully explored when understanding fraud attacks through stealthy clicks. Our work shows that time analysis may be important in order to improve the detection of fraud by clicking.

VII. REFERENCES

[1]. M. R. Berthold, N. Cebron, F. Dill, T. R. Gabriel, T. K□otter, T. Meinl, P. Ohl, K. Thiel, and B. Wiswedel. KNIME {The Konstanz information miner: Version 2.0 and beyond. SIGKDD Explorations Newsletter, 11(1):26{31, 2009.

[2]. G. E. P. Box. Non-normality and tests on variances. Bio metrika, 30(3/4):318{335, 1953.

[3]. L. Breiman. Bagging predictors. Machine Learning, 24: 123 {140, 1996.

[4].L. Breiman. Random forests. Machine Learning, 45(1):5{32, 2001.

[5]. C. Chambers. Is click fraud a ticking time bomb under Google? Forbes Magazine, 2012. URL http:// www. forbes. com/sites/investor/2012/06/18/is-click-fraud-a-ticking-time-bom b-under-google/.

[6]. C. C. Chang and C. J. Lin. LIBSVM: A library for support vector machines. ACM Transactions on Intelligent Systems Technology, 2(3):27:1{27:27, 2011.

[7].A. Chao and T. Shen. Nonparametric estimation of shannon's index of diversity when there are unseen species in sample. Environmental and Ecological Statistics, 10:429{443, 2003.

[8]. N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegel meyer. SMOTE: Synthetic minority over-sampling technique. Journal of Articial Intelligence Research, 16:321{357, 2002.

[9]. C. Chen, A. Liaw, and L. Breiman. Using random forests to learn imbalanced data. Technical report, Technical Report No. 666, Department of Statistics, University of California, Berk eley, 2004.

[10]. W. Cohen. Fast effective rule induction. In Proceedings of the International Conference on Machine Learning, pages 115{123, Tahoe City, California, 1995.

[11].T. Cover. Nearest neighbor pattern classiffication. IEEE Transactions on Information Theory, 13(1):21{27, 1967.

[12]. V. Dave, S. Guha, and Y. Zhang. Measuring and fingerprinting click-spam in ad networks. In ACM SIGCOMM Computer Communication Review, volume 42, pages 175{186, Helsinki, Finland, 2012.

[13].P. Domingos. MetaCost: A general method for making classiffiers cost-sensitive. In Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pages 155{164, 1999.

[14]. R. E. Fan, K. W. Chang, C. J. Hsieh, X. R. Wang, and C. J. Lin. LIBLINEAR: A library for large linear classiffication. Journal of Machine Learning Research, 9:1871{1874, 2008.

[15]. Amazon EC2 Instance Types. Retrieved March 18, 2010, from http://aws.amazon.com/ec2/instance-types/.

[16].Google AdWords Tra c Estimator. Retrieved February 1, 2010, from https:// adwords. google.com/select/ Traffic Estimator Sandbox.

[17].Invalid Clicks - Google's Overall Numbers. Retrieved May 10, 2010, from http://adwords.blogspot.com/2007/02/invalid-clicks-googles-overall-numbers.html, February 2007.

[18]. Apache Lucene Mahout: k-Means. Retrieved April 6, 2010,fromhttp://cwiki.apache.org/MAHOUT/k-means.html, No v ember 2009.

[19].Dhruba Borthakur. HDFS architecture. Retrieved April 29, 2010, from http://hadoop.apache.org/common/docs/ current/ hdfs_ design.html, February 2010.

[20]. Varun Chandola, Arindam Banerjee, and Vipin Kumar. Anomaly Detection: A Survey. ACM Computing Surveys, 41, 2009.

[21].C.C. Chang and C.J. Lin. LIBSVM: a library for support vector machines, 2001.

[22].D. Chang, M. Chin, and C. Njo. Click Fraud Prevention and Detection. Erasmus School of Economics e Erasmus University Rotterdam, 2008.

[23].N. Daswani and M. Stoppelman. The anatomy of Clickbot. A. In Proceedings of the rst conference on First Workshop on Hot Topics in Understanding Botnets, page 11. USENIX Association, 2007.

[24]. J. Dean and S. Ghemawat. Map Reduce: Simplied data processing on large clusters. Communications of the ACM-Association for Computing Machinery-CACM, 51(1):107114, 2008.

[25]. Peter Eckersley. A primer on information theory and privacy. Retrieved April 28, 2010, from https:// www. eff.org/deeplinks/2010/01/primer-information-theory-and-privacy, January 2010.

[26].Tristan Fletcher. Support vector machines explained. 2009. D. Foregger, J. Manuel, R. Ramirez-Padron, and M. Georg iopoulos. Kernel similarity scores for outlier detection in mixed-attribute data sets. 2009.

[27].M. Gandhi, M. Jakobsson, and J. Ratkiewicz. Badver tisements: Stealthy click-fraud with unwitting accessories. Journal of Digital Forensic Practice, 1(2):131-142, 2006.

[28].Z. He, S. Deng, X. Xu, and J. Huang. A fast greedy algorithm for outlier mining. Advances in Knowledge Discovery and Data Mining, pages 567-576, 2005.

[29].Jackson, C., Barth, A., Bortz, A., Shao, W. and Boneh, D.: Protecting Browsers from DNS Rebinding Attacks, Proceedings of the 14th ACM conference on Computer and communi-cations security, October 26, 2007, pp. 421 – 431 (2007)

[30].Jansen, B. J.: The Comparative Effectiveness of Sponsored and Non-sponsored Results for Web Ecommerce Queries. ACM Transactions on the Web. 1(1), Article 3, http:// ist.psu. edu/ faculty_pages/jjansen/academic/pubs/jansen_tweb_sponsored_links.pdf (2007)

[31]. Jansen, B., Flaherty, T., Baeza-Yates, R., Hunter, L., Kitts, B., Murphy, J.: The Compo-nents and Impact of Sponsored Search, Computer, Vol. 42, No. 5, pp. 98-101. May 2009 http://ist.psu.edu/faculty_pages/jjansen/academic/pubs/jansen_sponsored_search_ieee.pdf (2009)

[32].Kantarcioglu, M., Xi, B., Clifton, C.: A Game Theoretic Approach to Adversarial Learn-ing, National Science Found ation Symposium on Next Generation of Data Mining and Cyber-Enabled Discovery for Innovation, Baltimore, MD, http://www.cs.umbc.edu/~hillol/NGDM07/abstracts/poster/MKantarcioglu.pdf (2007)

[33].Kitts, B.: Regression Trees, Technical Report, http:// www.appliedaisystems.com/papers/RegressionTrees.doc (2000)

[34].Kitts, B. Laxminarayan, P. and LeBlanc, B.: Cooperative Strategies for Keyword Auc-tions, First International Conference on Internet Technologies and Applications, Wales. September 2005. (2005)

[35]. Wellman, M., Greenwald, A., Stone, P. and Wurman, P. (2003a) 'The 2001 Trading Agent Competition', Electronic Markets 13(1): 4–12.