A Project Report on

COVID 19 -

Detection of Social Distancing in Industrial Environment

Submitted in partial fulfilment for award of degree of

MBA

In **Business Analytics**

Submitted by

Anshuman Dash

R18DM006

Under the Guidance of

Sandeep Giri

Founder, Cloudx Lab

REVA Academy for Corporate Excellence

REVA University

Rukmini Knowledge Park, Kattigenahalli,

Yelahanka, Bangalore – 560064

September, 2020

REVA UNIVERSITY
Bengaluru, India

Candidate's Declaration

I, Anshuman Dash hereby declare that I have completed the project work towards the MBA, Business Analytics at, REVA University on the topic entitled COVID 19 - Detection of Social Distancing In Industrial Environment under the supervision of Sandeep Giri, Founder Cloudx Lab. This report embodies the original work done by me in partial fulfilment of the requirements for the award of degree for the academic year 2020.

Anshuman Dash

Place: Bengaluru — Name of the Student:

Date: 11-Sep-20 — Signature of Student

REVA UNIVERSITY
Bengaluru, India

Certificate

This is to Certify that the PROJECT work entitled COVID 19 - Detection of Social Distancing In Industrial Environment carried out by Anshuman Dash with SRN R18DM006, 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 2020. The Project report has been tested for plagiarism and has passed the plagiarism test with the similarity score less than 15%. The project report has been approved as it satisfies the academic requirements in respect of PROJECT work prescribed for the said Degree.

Sandeep Giri

Guide

Dr. Shinu Abhi

Director

External Viva (Virtually Done)

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: 07/10/2020

Acknowledgement

I would like to thank the following people, without whom I would neither have been able to complete this research, nor would I have made it through my MBA degree:

Dr. P Shayma Raju, Chancellor REVA University
Dr. S.Y Kulkarni, Ex- Vice Chancellor REVA University
Dr. Dr. K. Mallikharjuna Babu, Vice Chancellor REVA University
Dr. M. Dhanamjaya, Registrar REVA University

I would also like to extend my special thanks to Dr. Shinu Abhi, Director REVA Academy of Corporate Excellence for believing in me and for providing all the guidance throughout my MBA degree and for guiding me through all the publications.

I want to thank my guide Sandeep Giri, Founder, Cloudx Lab for all the help in guiding me through the project and for showing me the path I could walk for this research project.

I would also thank all the mentors at REVA Academy of Corporate Excellence for teaching us awesome methodologies throughout the academic year.

Finally, I want to thank my family for patiently supporting me while I was busy working on this research paper, and for believing in me.

Place: Bengaluru
Date: 07/10/2020

REVA UNIVERSITY
Bengaluru, India

Similarity Index Report

Title of the Thesis: COVID 19 - Detection of Social Distancing in Industrial Environment

Total No. of Pages: 39

Name of the Student: Anshuman Dash

Name of the Guide(s): Sandeep Giri

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

Software Used: Turnitin
Date of Report Generation: 16-Sep-2020
Similarity Index in %: 13%
Total word count: 3997

Place: Bengaluru

Name of the Student: Anshuman Dash

Date: 17/09/2020

Signature of Student

Verified By:

Signature

Dr. Shinu Abhi, Director, Corporate Training

List of Abbreviations

Sl. No	Abbreviation	Long Form
1	CNN	Convolutional Neural Network
2	Yolo	You Only Look Once
3	OpenCV	Open Source Computer Vision
4	SVC	Support Vector Classifier
5	ML	Machine Learning

List of Figures

List of Tables

Abstract

Coronavirus (COVID-19) disease is an infectious disease caused by the coronavirus that recently was discovered. The COVID-19 virus is primarily spread by droplets when an affected person is active, sneezed or exhaled. (Yawney & Gadsden, 2020). WHO suggests that we keep at least 1 meter (3 feet) apart to avoid the transmission of the disease between us and others.(Morawska & Cao, 2020) Now as communities and work places are reopening, it has become the responsibility of employers to ensure worker health and they need to implement safety and social measures to prevent the spread of COVID-19 at work place.

With COVID 19 requirements eased in numerous countries and factories preparing to reopen, ensuring employee safety in workplaces or factory shops is of priority for the company. The social isolation at work is incredibly necessary to ensure that employees and workers have a healthy work environment. The problem we are trying to address as part of this project is "How do we monitor and measure social distancing on the factory floor?"

The objective of this project is to build a social distancing detection tool that can detect if people are keeping a safe distance from each other by analyzing real time video streams from the camera. This will help employers to measure and ensure social distancing protocols are being followed on the factory shop floor and employee's safety is being maintained at the workplace by preventing the spread of Corona virus or COVID-19.

We are building a computer vision model that would be able to monitor the factory shop floor or the employee workplace. The model would be able to monitor employees on the shop floor and identify the distance between each individual. If the distance between two individuals are less than 2 feet, the monitoring system would identify and display the individuals in Red bounding boxes. This model can also be used to alert the supervisors about the social distancing violation on the factory shop floors.

The project uses Python packages like SciPy, OpenCV Python, NumPy, Imutils. Yolo CNN is used to identify "Person" objects in images and video files. We are identifying the coordinates of each paired object's bounding boxes and we are checking if there is social distancing violation between objects. We will then extract the paired coordinates into data sheet and label the target column which defines if there is social distance violation or not. This data is then passed through 10 classification machine learning models to validate accuracy and F1 score from each model.

XGBoost appeared as the winning model by giving us Accuracy and F1 Score of 90.5%.

Keywords: COVID-19, Social Distancing, Safe Workplace, Computer Vision, CNN, OpenCV, Yolo

Contents

Chapter 1: Introduction

Coronavirus (COVID-19) was a recently discovered coronaviral infectious condition. Many patients with COVID-19 suffer from mild to severe symptoms.

Current research indicates that COVID-19 is passed through individuals through the means of direct or indirect interaction with infected persons (through infected items or surfaces). The distance from each other should be 1 meter at least (3 feet). It is recommended. When someone is coughing, sneezing or talking, the nose and mouth will release tiny liquid droplets that contain the virus. When we are too close, if the individual has the disease, we will breathe in the droplets, including COVID-19. In other words, as a way to avoid the spread of COVID-19, the WHO recommends social isolation (Lisa Lockerd Maragakis, M.D., 2020).

Now that neighborhoods are being revived and people are more frequently in public, the importance of preserving physical position while in public places is emphasized by social distancing. Companies have begun opening offices and warehouses.

As an employer, it has become their responsibility to ensure worker health and they need to implement safety and social measures to prevent the spread of COVID-19 at workplace.

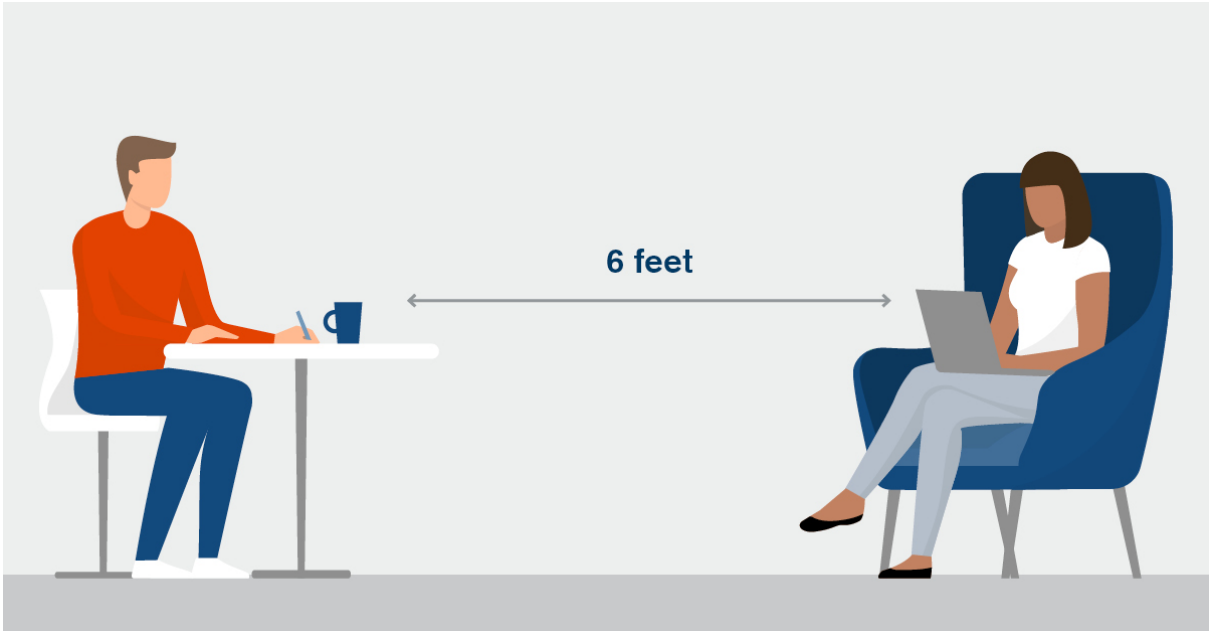


Figure 1

Chapter 2: Literature Review

The life of all humans is now dominated by Coronavirus disease (COVID-19), and its history has been constantly revised. SARS-CoV was the first recognized in China, which first came to the attention WHO in Vietnam in November 2002 (though not identified at the time). It's a flu-like illness, and commonly having diarrhea, caused by Extreme Acute Respiratory Syndrome (SARS). Pneumonia and breathing failure could progress in two weeks. SARS-CoV–2 appears similar to the wild bat virus than SARS-CoV or MERS-CoV indicates it is an introduction of human coronavirus.(Chaplin, 2020)

For all countries, prevention and management of the COVID-19 outbreak is a significant problem. A significant problem for the prevention and control department of all countries is how to estimate the risk grade of the epidemic. The degree of infection risk is taken as the degree of prevention and control when taking decisions on disease management and prevention. There are two research features and patterns according to the present COVID circumstances. One is to pay attention to research into the application and management of epidemics. Secondly, it focuses on the evaluation of disease risk.(He, 2020)

Over 28 million people (Worldometer, 2020) were infected and the number has continued to increase since the beginning of the new COVID-19 pandemic. Various preventive steps to slow the spread of the disease were taken. Therefore, early detection of symptoms and proper sanitary practices are extremely important, especially in sites in which individuals have random or opportunistic contact. Automated systems for body temperature measurement, hygienic adherence assessment and individualized person-to - person monitoring are necessary to ensure economic stability, not just for the spread of diseases intervention and prevention. (Barabas et al., 2020)

Since the disease has not been cured yet, prevention is our best option. Health precautions are the only way to avoid and protect against the infection of Coronavirus. Who is infected and who is cannot easily be detected so the safest choice is to avoid physical contact with the people around (Seniority, 2020). Without adequate treatment, managing infection sources is the only way to cope with the SARS-CoV-2 epidemic. Strategies include precocious diagnosis, monitoring, isolation and therapies of support; prompt dissemination of disease data; and preservation of social orders. Individuals can effectively prevent SARS-CoV-2 infection by protection measures such as improvements in personal hygiene, use of medical masks, proper rest and well ventilated rooms.(Sun et al., 2020)(Технічні et al., 2016)

In that respect, equipment for either treatment (e.g. fans) or automated, efficient identification and tracking of potential carriers of viruses has become important tools for pandemic combat and further prevention of dissemination. A simple measurement of the body temperature at various check spots with the aid of infra - red sensor and the review of proper hygienic standards, coupled with indoor and personalized location data results, may be a fast, effective

way of recognizing and regulating the disease. Polymerase chain reverse transcription or antibody test, is a time consuming and costly activity. (Barabas et al., 2020)

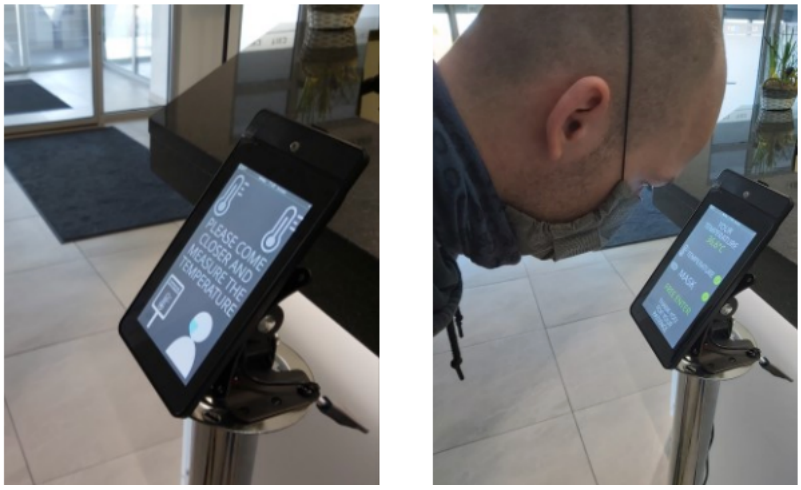


Figure 2

Figure 2 displays the proposed digital temperature and hygienic conformity checks. The unit is placed on a permanent stand at various examination points, e.g. entrance halls, before the main entry to the workplace. The prototype system is linked via LAN or WiFi to the local area network to insert the information needed to SQL database server, which saves the acquired measurement data to the location of the stand-alone RTLS database. (SEWIO, 2019)

Another effective way of managing Corona virus transmission is mask identification using facial recognition technologies. The mask detection algorithm is a hybrid method, based on a Histogram Oriental Gradients (HOG) approach, using both neural networks (NN) and vector definitions (Sagonas et al., 2016). Although images are obtained with spatial resolution, they are decreased for further processing to enhance efficiency. As a framework for face-identification, certified models with a total of about 300 000 faces that are publicly available can be used. This network model will be processed in OpenCV Deep Neural Network modules and stored in Caffe System format (Open CV Dev. Team, 2013).

COVID-19 simulations of social network data in the UK and the United States (Vokó & Pitter, 2020) recently published, noted that aimed to minimize epidemics would entail a complex set of measures involving a social distance between the whole population. (Flaxman et al., 2020). The most powerful approach to limit COVID-19 transmission is through social distancing (Chen et al., 2020). It's very difficult, though, for people to work closely together in a closed environment like a factory, to maintain social distance on a shop floor. (Henry, 2020).

Research gaps & conclusion

Every part of our lives is affected by Coronavirus-19 (COVID-19). COVID-19, which was reported at the end of 2019, was identified by March 2020 as a global pandemic. We propose that COVID-19 need us as a scientific and clinical group to prioritize and mobilize in several key fields: (a) diagnosis, (b) prevention, (c) interpersonal contact, (d) collaboration and participation of non-mental health programs for medical personnel, and (e) COVID-19 basic inquiry into trauma. We expect that leaders in traumatic stress fields will take into account the

limitations of our existing approaches and will devote the strategic and financial capital required desperately to establish collaborations to help people in greater need and to build innovations. (Horesh & Brown, 2020).

Chapter 3: Problem Statement

A pandemic has had a devastating influence on industry at any stage of history. The world's health pandemic and economic pandemic, which have almost crossed billions in sales, are being combated with the 2019 Corona Virus Disease (COVID-19).

World leaders are building war rooms in order to preserve the light to counteract the effects of the COVID-19. Market research analysts actively advise clients on COVID-19 's sales effect on their companies and their subsequent shift to short-term goals and priorities.

Enterprises must tackle coronavirus 'financial and operational problems and meet their staff, consumers and suppliers' demands quickly. With COVID 19 requirements eased in numerous countries and factories preparing to reopen, ensuring employee safety in workplaces or factory shops is of priority for the organization. The aspect of social distancing at work place is utmost importance for ensuring safe work environment for the employees and workers. **The problem we are trying to address as part of this project is "How do we monitor and measure social distancing on the factory floor?"**

Chapter 4: Objectives of the Study

The objective of this study is to build a social distancing detection tool which can

- Monitor the movement and placement of workers on the factory shop floor using video streams from cameras
- Measure if the employees are keeping safe distance between each other while being on the floor
- Identify if any of the employees are violating the defined guidelines for social distancing
- Display the violations on a common screen on the floor
- Provide an alert indicating the violation
- Ensure employee health and safety measures are being maintained on the factory shop floor

Chapter 5: Project Methodology

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

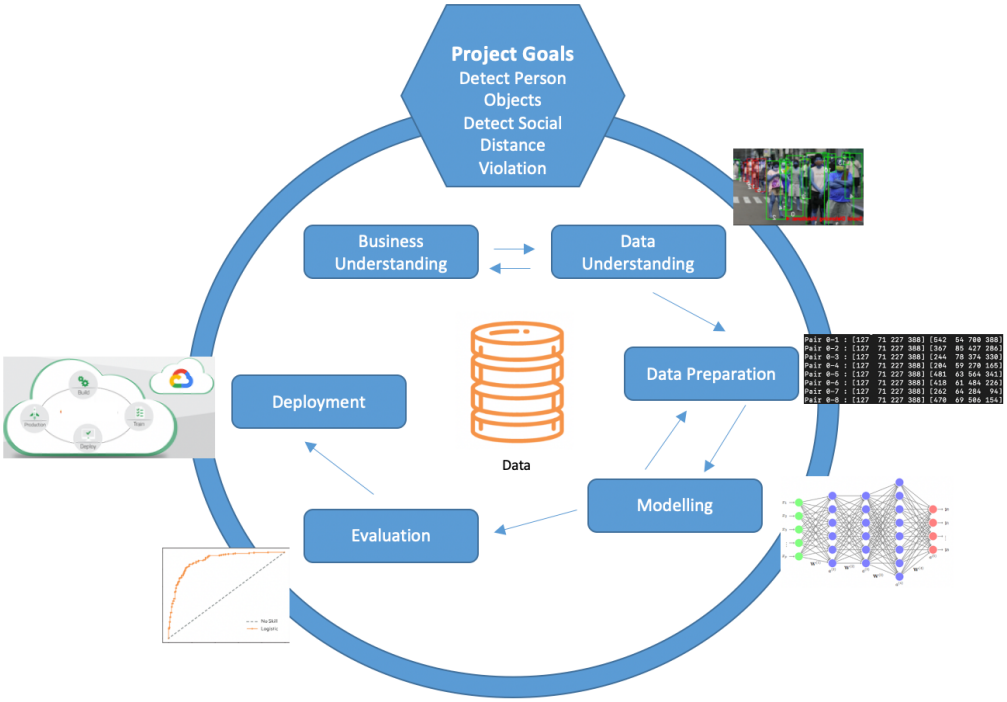


Figure 3

Chapter 6: Business Understanding

The pandemic of COVID-19 already raises a host of problems for all industrial producers. Many that depend on workers whose jobs cannot be remotely conducted. Roughly 80 % of companies believe that the pandemic would negatively harm their business. Most (53 percent) manufacturers expect COVID-19 to have its effect on their market. Some large industries have shut down facilities and have increased layoffs so that disease dissemination and economical profitability can be curbed. Clearly, the market, including some 13 million American workers, will be hit hard by this outbreak (PWC, 2020). According to US Bureau of Labor Statistics, the manufacturing industry saw a decline of 720000 jobs since February 2020. (U. S. Bureau of Labor Statistic, 2020)

In order to reduce the financial impact of COVID, it's important to reopen businesses and restart production. This will need to factories to reopen and start the shop floor. At the same time, it becomes necessary to ensure the virus spread is restricted. This can be possible if the companies adopt to the WHO guidelines and create a safe environment for workers. One way to create a safe working environment is by making sure employees are maintaining social distancing guidelines while they are at the work place.

Chapter 7: Data Understanding

In a typical factory floor, there would be so many employees working together. Many factories are yet to have CCTV cameras installed on their shop floors. It becomes very difficult at this moment to get data from factory shop floors. Hence we have used publicly available data to accomplish the goals of our projects.

- Publicly available images were gathered which comprises of group of people within one image
- Publicly available video files of "Oxford town center".

Sample downloaded image

Figure 4

Chapter 8: Data Preparation

Images were passed through the model and values of 4 coordinates were captured for each object. The coordinates are x, y, Width and Height of the bounding boxes around the "Person" objects. X and Y values were arrived using the below formula:

x = int(centerX - (width / 2))
y = int(centerY - (height / 2))

A function was written to provide paired coordinates for objects in each image. These paired coordinates were manually captured in an Excel sheet and they were labelled to identify if the pair violated social distancing or not. The violation column represents '0' for no social distance violation and '1' for social distance violation.

Below is the snapshot of data output from one such image with 3 human objects.

Image	Pair	Per 1 X	Per 1 Y	Per 1 Width	Per 1 Height	Per 2 X	Per 2 Y	Per 2 Width	Per 2 Height	Violation
5	0-1 :	13	60	144	421	555	82	686	409	0
5	0-2 :	13	60	144	421	285	80	416	402	0
5	1-2 :	555	82	686	409	285	80	416	402	0

Table 1

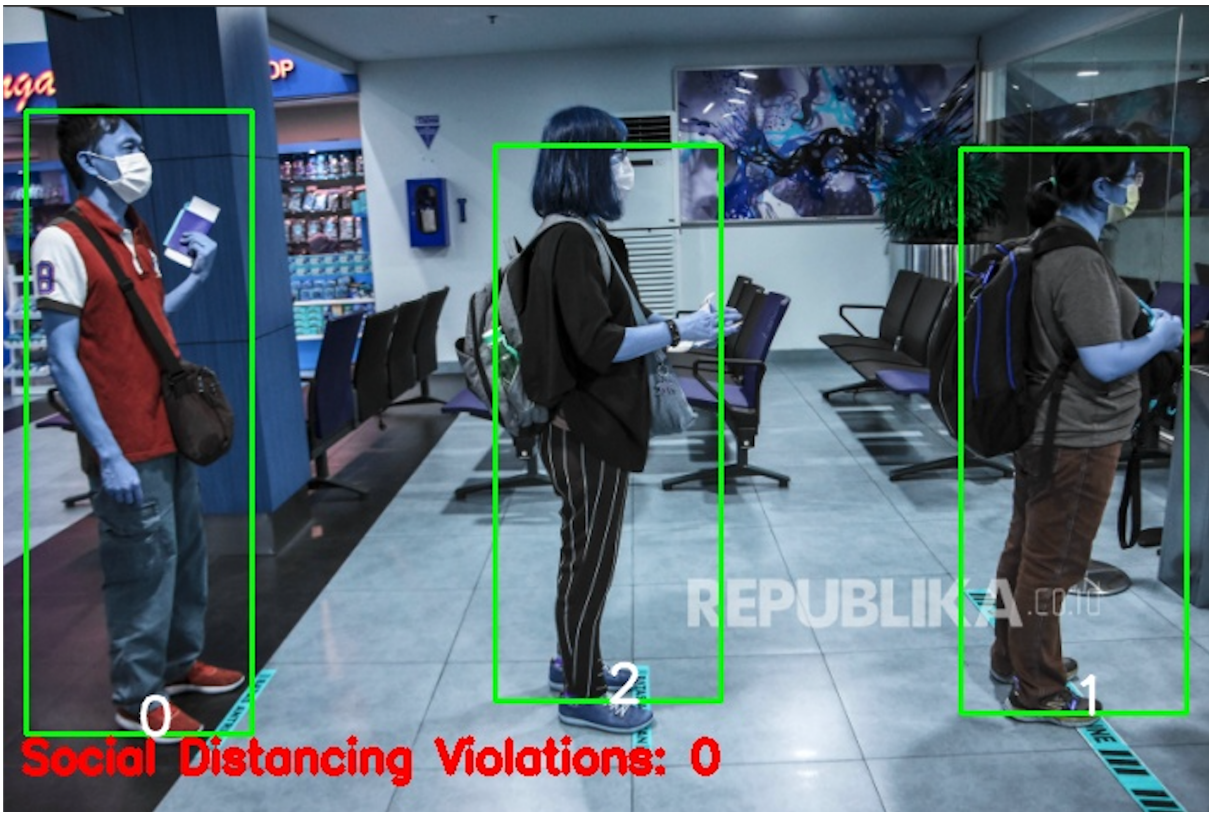


Figure 5

Similarly Oxford Town Centre video file was passed through the model to identify social distance violations between persons in the video

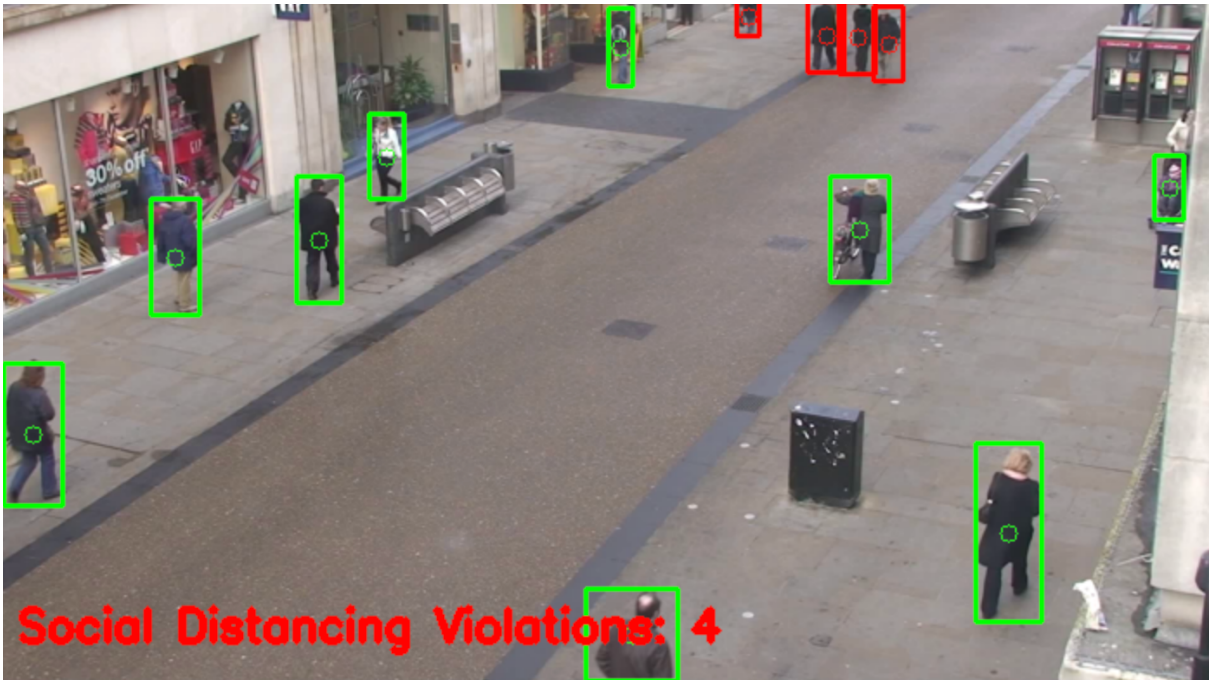


Figure 6

Chapter 9: Data Modeling

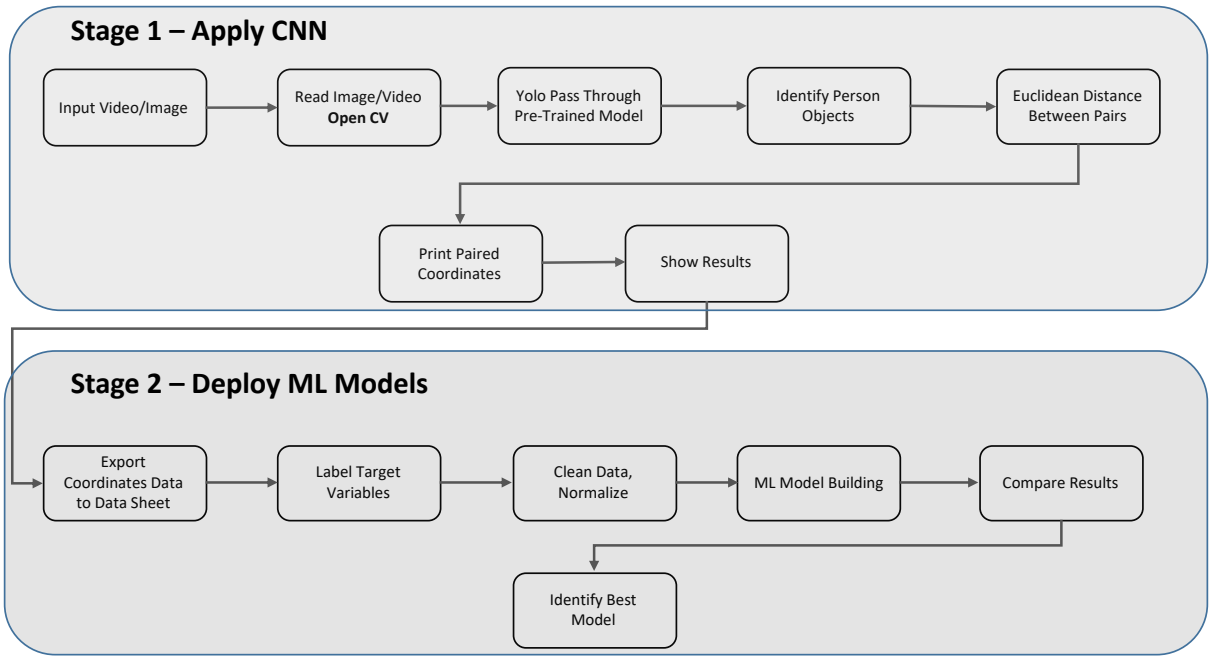


Figure 7

Data model deployment steps:

- Install dependencies
  - SciPy
  - OpenCV Python
  - NumPy
  - Imutils
- Import all the required libraries
- Download the pre-trained Yolo models – Yolov3 weights and config file
- Read an image and pass it on the model for prediction
- Identify the "Person" objects in the images and the bounding boxes around each object
- Print 4 coordinates for each bounding box – X, Y, Height and Width
- Compute the distance between two people in an image using bounding boxes
- Define a function to compute the Euclidean distance between every two "Person" objects in an image
- Define a function that returns the closest people
- Define a function to change the color of the closest people to red
- Print how many social distance violations per image
- Extract the coordinates of each paired objects in the image and label for social distance violation
- Run at least 10 machines learning models on the extracted data

We have defined bounding boxes for each individual, which are valued in the bird's eye view as their (x , y) position. Because the calibration stage transforms the ground pane, we use it to shift into the middle center of the boundary of each individual, resulting in its place in the view of the bird's eye. The final step is to determine the size of the eye of the bird from each pair of

persons and to measure their distances by the measured calibration factor. We highlight people whose distance is less than 50 pixels in red bounding boxes.

Chapter 9: Model Evaluation

The tri-phase review of the implementation of classifiers was carried out to assess the appropriateness of the different classification methods for the scenario. Below metrices were used as KPIs to measure and evaluate the performance of our built models.

- $Accuracy = \frac{True\ Positives + True\ Negatives}{True\ Positives + True\ Negatives + False\ Positives + False\ Negatives}$
- $Precision = \frac{True\ Positives}{True\ Positives + False\ Positives}$
- $Recall = \frac{True\ Positives}{True\ Positives + False\ Negatives}$
- $F1\ Score = 2 * \frac{Precision * Recall}{Precision + Recall}$

Data was passed through 10 classification Machine Learning Models and KPI results were compared to identify the best model.

Algorithm	Accuracy	Precision	Recall	F1 Score
XGBoost	0.905109	0.686041	0.845920	0.905109
RandomForestClassifier	0.897810	0.752568	0.779442	0.897810
GradientBoostingClassifier	0.890511	0.701214	0.767370	0.890511
LinearDiscriminantAnalysis	0.875912	0.527778	0.937500	0.875912
LogisticRegression	0.868613	0.500000	0.434307	0.868613
LinearSVM	0.868613	0.500000	0.434307	0.868613
rbfSVM	0.868613	0.500000	0.434307	0.868613
QuadraticDiscriminantAnalysis	0.861314	0.542951	0.639394	0.861314
KNearestNeighbors	0.839416	0.601074	0.625726	0.839416
DecisionTree	0.795620	0.670168	0.619430	0.795620
GaussianNB	0.766423	0.464753	0.461002	0.766423

Table 2

The final model that would be used in production environment would be XGBoost, due to its highest Accuracy and F1 Score of 90.5%

Chapter 10: Deployment

We tested our solution to be working perfectly with images and offline video files. This solution can now to be deployed to the systems connecting to CCTV cameras on the factory shop floors and ensure worker safety in a much convenient way. It can be deployed either using an on-prem solution, running on a Deep Leaning machine or a cloud-based solution like Azure ML workspace. The overall model would include connecting the CCTV cameras on the Shop Floor to the solutions server and displaying the results on the available screens on the floor. The visual display can identify social distancing violation on the floor and can also alert the supervisors.

For example, at a factory that produces protective equipment, technicians could integrate this software into their security camera systems to monitor the working environment with easy calibration steps. The detector could highlight people using red bounding boxes, if the distance between two or more people is below the minimum acceptable value. The system can also be configured to issue an alert to remind people to keep a safe distance if the protocol is violated.

Chapter 11: Conclusions and Recommendations for future work

As medical experts say, social distancing is our best strategy for minimizing the coronavirus pandemic and opening up the environment before a vaccine is available. Our goal is to support our customers and to inspire other people to try new ideas to defend us at such an early stage that can keep us safe.

The method can be further strengthened. The system's adaptive character allows for the constantly improving learning results. However, there are other approaches to boost the accuracy of the recognition method. In this research the models can be further improved, and "human" object detection capability can be enhanced using a wide variety of classification algorithms.

Bibliography

Barabas, J., Zalman, R., & Kochlan, M. (2020). *Automated evaluation of COVID-19 risk factors coupled with real-time, indoor, personal localization data for potential disease identification, prevention and smart quarantining*. 645–648. https://doi.org/10.1109/tsp49548.2020.9163461

Chaplin, S. (2020). *COVID-19: a brief history and treatments in development*. https://onlinelibrary.wiley.com/doi/10.1002/psb.1843#:~:text=SARS-CoV was the first,in China in April 2004.

Chen, S., Yang, J., Yang, W., Wang, C., & Bärnighausen, T. (2020). COVID-19 control in China during mass population movements at New Year. In *The Lancet*. https://doi.org/10.1016/S0140-6736(20)30421-9

Flaxman, S., Mishra, S., Gandy, A., Unwin, H. J. T., Coupland, H., Mellan, T. A., Zhu, H., Berah, T., Eaton, J. W., Guzman, P. N. P., Schmit, N., Callizo, L., Team, I. C. C.-19 R., Whittaker, C., Winskill, P., Xi, X., Ghani, A., Donnelly, C. A., Riley, S., … Bhatt, S. (2020). *Estimating the number of infections and the impact of non-pharmaceutical interventions on COVID-19 in European countries: technical description update*. *March*, 1–35. http://arxiv.org/abs/2004.11342

He, P. (2020). *Study on Epidemic Prevention and Control Strategy of COVID -19 Based on Personnel Flow Prediction*. 688–691. https://doi.org/10.1109/icuems50872.2020.00150

Henry, B. F. (2020). Social Distancing and Incarceration: Policy and Management Strategies to Reduce COVID-19 Transmission and Promote Health Equity Through Decarceration. *Health Education and Behavior*. https://doi.org/10.1177/1090198120927318

Horesh, D., & Brown, A. D. (2020). Covid-19 response: Traumatic stress in the age of Covid-19: A call to close critical gaps and adapt to new realities. *Psychological Trauma: Theory, Research, Practice, and Policy*, *12*(4), 331–335. https://doi.org/10.1037/TRA0000592

Lisa Lockerd Maragakis, M.D., M. P. H. (2020). *Coronavirus, Social and Physical Distancing and Self-Quarantine*.

Morawska, L., & Cao, J. (2020). Airborne transmission of SARS-CoV-2: The world should face the reality. In *Environment International*. https://doi.org/10.1016/j.envint.2020.105730

Open CV Dev. Team. (2013). OpenCV Introduction. In *OpenCV Online Documentation*.

PWC. (2020). *COVID-19: What it means for industrial manufacturing*. https://www.pwc.com/us/en/library/covid-19/coronavirus-impacts-industrial-manufacturing.html

Sagonas, C., Antonakos, E., Tzimiropoulos, G., Zafeiriou, S., & Pantic, M. (2016). 300 Faces In-The-Wild Challenge: database and results. *Image and Vision Computing*. https://doi.org/10.1016/j.imavis.2016.01.002

Seniority. (2020). *12 Precautions To Protect Yourself From Coronavirus (COVID-19) #7th Is Must Do | How To Protect Yourself From Coronavirus*. https://www.seniority.in/blog/12-precautions-against-coronavirus-covid-19-7-is-a-must-do/

SEWIO. (2019). *Real-Time Location System (RTLS) on Ultra-wideband | Sewio RTLS*. https://www.sewio.net/real-time-location-system-rtls-on-uwb/

Sun, P., Lu, X., Xu, C., Sun, W., & Pan, B. (2020). Understanding of COVID-19 based on current evidence. In *Journal of Medical Virology*. https://doi.org/10.1002/jmv.25722

U. S. Bureau of Labor Statistic. (2020). Current employment statistics highlights. *U.S. Bureau of Labor Statistics*, *August 2020*, 18. http://www.bls.gov/ces/cesprog.htm

Vokó, Z., & Pitter, J. G. (2020). The effect of social distance measures on COVID-19 epidemics in Europe: an interrupted time series analysis. *GeroScience*. https://doi.org/10.1007/s11357-020-00205-0

Worldometer. (2020). *Coronavirus Update (Live): 28,944,152 Cases and 924,580 Deaths from COVID-19 Virus Pandemic - Worldometer*. https://www.worldometers.info/coronavirus/

Yawney, J., & Gadsden, S. A. (2020). A Study of the COVID-19 Impacts on the Canadian Population. *IEEE Access*, *8*, 128240–128249. https://doi.org/10.1109/ACCESS.2020.3008608

Технічні, Ч. І., Дичко, С. М., & Васлович, В. В. (2016). *Оригінальна стаття = Original article = Оригинальная статья*. *3154*, 18–27.

Appendix

Plagiarism Report

Detection of Social Distancing In Industrial Environment

ORIGINALITY REPORT

13%	%	%	%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	Submitted to Sogang University Student Paper	3%
2	www.netnest.com.au Internet Source	1%
3	landing.ai Internet Source	1%
4	J. Barabas, R. Zalman, M. Kochlan. "Automated evaluation of COVID-19 risk factors coupled with real-time, indoor, personal localization data for potential disease identification, prevention and smart quarantining", 2020 43rd International Conference on Telecommunications and Signal Processing (TSP), 2020 Publication	1%
5	www.termpaperwarehouse.com Internet Source	1%
6	worldtechvalley.com Internet Source	1%
7	Submitted to University of Southampton	

Publications in a Journal/Conference Presented/White Paper

Auto-Detection of Click-Frauds using Machine Learning - Paper presented in Seventh International Conference on Business Analytics & Intelligence (BAICONF2019) – Conducted by IIM, Bangalore

भारतीय प्रबंध संस्थान बेंगलूर
INDIAN INSTITUTE OF MANAGEMENT
BANGALORE

IIMB
तेजस्वि नावधीतमस्तु

DCAL
DATA TO DECISIONS

Paper Presentation

This is to certify that the paper titled

AUTO-DETECTION OF CLICK-FRAUDS USING MACHINE LEARNING

authored by

ANSHUMAN DASH & SATYAJIT PAL

was presented at the

"Seventh International Conference on Business Analytics and Intelligence"

5 - 7 December, 2019

U Dinesh Kumar
Conference Chair

INDIAN INSTITUTE OF MANAGEMENT BANGALORE, BANNERGHATTA ROAD, BANGALORE 560 076, INDIA

Paper published in IJESC Journal

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
Anshuman Dash
(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].

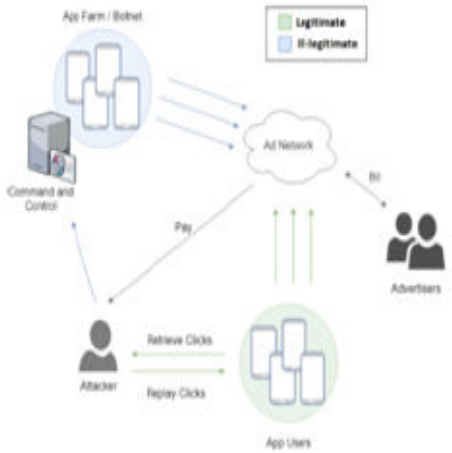


Figure.1.Legitimate & Ill-legitimate Click Fraud

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.

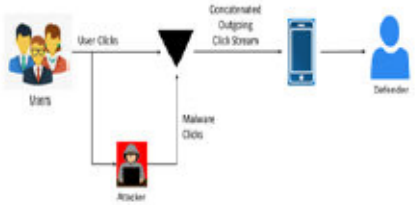


Figure.3. Click-Fraud Detection Problem

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

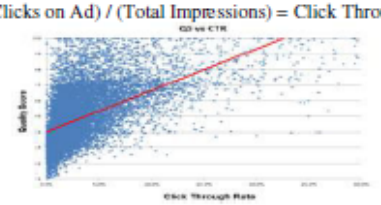


Figure.4. Click Through Rate (https://www.wordstream.com/click-through-rate)

•**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.

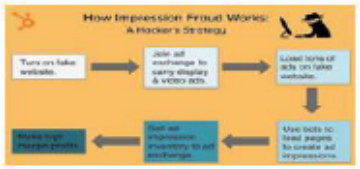


Figure.5. Impression Fraud (https://blog.anura.io/blog/what-is-impression-fraud-and-how-does-it-work)

•**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:

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

[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.

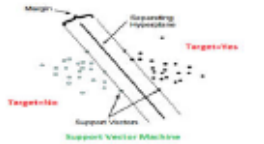


Figure.8. SVM Example (http://www.1proga.info/support-vector-machine-algorithm-38912859 3a10a109ec0dd4/)

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:

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

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.

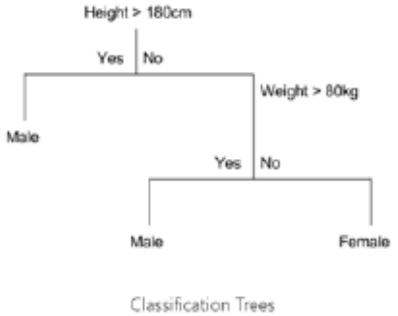


Figure.9. Classification Tree (https://www.digitalvidya.com/blog/classification-and-regression-trees/)

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

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

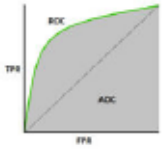
Precision is the calculated ratio of predicted ransomware that are correctly identified as a malware. Precision is defined below

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

["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.

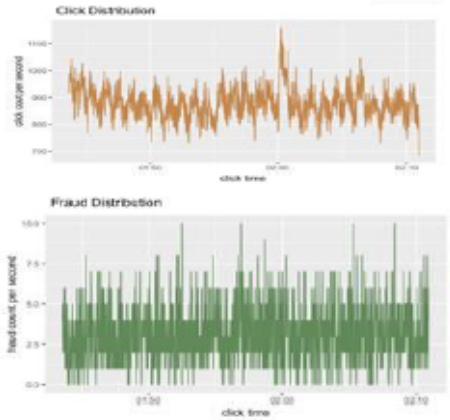


Figure.10. Dataset Distribution

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.

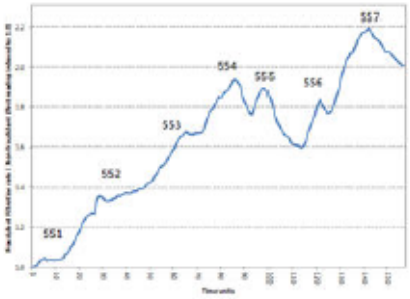


Figure.11. Fraudulent versus Non-Fraudulent Rate

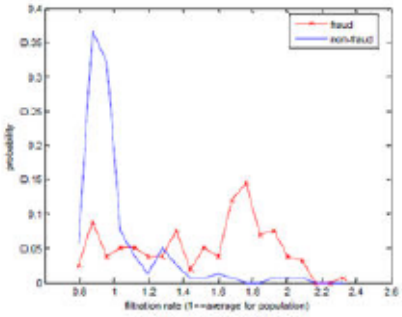


Figure.12. Filtration Rate Ratio for fraudulent versus non-fraudulent click spams

[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.]

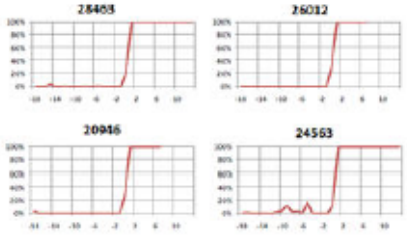


Figure.13. Filtration Rates for Four Fraudulent Click Spam

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%	89.40%
Classification Trees	94.40%	7.60%	92.40%	5.68%	93.20%
Support Vector Machine	95.40%	9.10%	90.90%	4.69%	92.76%
knn Classification	69.80%	69.00%	31.00%	31.00%	43.70%
Gradient Tree Boosting	96.80%	76.90%	93.10%	15.20%	97.20%

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

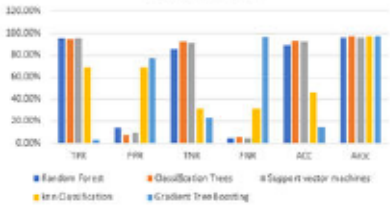
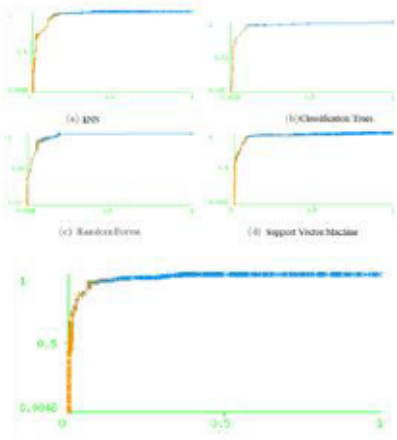


Figure.14. Classifier Accuracy

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.

Any Additional Details

Ransomware Auto-Detection in IoT Devices using Machine Learning - Paper presented in 6th International Conference on Business Analytics and Intelligence 2018 [ICBAI - 2018] -- Conducted by IIM, Bangalore and IISc, Bangalore

Paper published in IJESC Journal -

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

6th International Conference on
Business Analytics and Intelligence 2018
[ICBAI-2018]

Analytics Society of India

PAPER PRESENTATION CERTIFICATE

This is to certify that the paper entitled Ransomware Auto-Detection in IoT Devices using Machine Learning, authored by Anshuman Dash, Satyajit Pal, Girish Yellapa and Chinmay Hegde was presented at **ICBAI-2018** held at J.N. Tata Auditorium, **Indian Institute of Science, Bangalore** and scheduled during 20-22 December 2018.

Dr. M. Mathirajan
ORSI-BC, Secretary

Prof. U. Dinesh Kumar
ASI, President

IJESC INTERNATIONAL JOURNAL OF ENGINEERING SCIENCE AND COMPUTING

ISSN 2250-1371
www.ijesc.org

Certificate of Publication

Awarded to

Anshuman Dash

(REVA Academy for Corporate Excellence, REVA University, Bengaluru, India)

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

Volume 8 Issue No.12, Dec 2018

Manuscript titled

"Ransomware Auto-Detection In IoT Devices Using Machine Learning"

Dr. Aneesh Thomas
Editorial Head, IJESC