Computer Science & Information Technology 149

David C. Wyld,
Dhinaharan Nagamalai (Eds)

Computer Science & Information Technology

International Conference on AI, Machine Learning and Applications (AIMLA 2021),
August 28 ~ 29, 2021, Dubai, UAE.

AIRCC Publishing Corporation

Volume Editors

David C. Wyld,
Southeastern Louisiana University, USA
E-mail: David.Wyld@selu.edu

Dhinaharan Nagamalai (Eds),
Wireilla Net Solutions, Australia
E-mail: dhinthia@yahoo.com

ISSN: 2231 - 5403
ISBN: 978-1-925953-47-3
DOI: 10.5121/csit.2021.111301 - 10.5121/csit.2021.111306

Typesetting: Camera-ready by author, data conversion by NnN Net Solutions Private Ltd., Chennai, India

Preface

The International Conference on AI, Machine Learning and Applications (AIMLA 2021), August 28 ~ 29, 2021, Dubai, UAE, 9th International Conference on Database and Data Mining (DBDM 2021), 8th International Conference on Computer Networks & Communications (CCNET 2021), International Conference on NLP & Text Mining (NLTM 2021) and International Conference on IOT, Big Data and Security (IOTBS 2021) was collocated with International Conference on AI, Machine Learning and Applications (AIMLA 2021). The conferences attracted many local and international delegates, presenting a balanced mixture of intellect from the East and from the West.

The goal of this conference series is to bring together researchers and practitioners from academia and industry to focus on understanding computer science and information technology and to establish new collaborations in these areas. Authors are invited to contribute to the conference by submitting articles that illustrate research results, projects, survey work and industrial experiences describing significant advances in all areas of computer science and information technology.

The AIMLA 2021, DBDM 2021, CCNET 2021, NLTM 2021 and IOTBS 2021 Committees rigorously invited submissions for many months from researchers, scientists, engineers, students and practitioners related to the relevant themes and tracks of the workshop. This effort guaranteed submissions from an unparalleled number of internationally recognized top-level researchers. All the submissions underwent a strenuous peer review process which comprised expert reviewers. These reviewers were selected from a talented pool of Technical Committee members and external reviewers on the basis of their expertise. The papers were then reviewed based on their contributions, technical content, originality and clarity. The entire process, which includes the submission, review and acceptance processes, was done electronically.

In closing, AIMLA 2021, DBDM 2021, CCNET 2021, NLTM 2021 and IOTBS 2021 brought together researchers, scientists, engineers, students and practitioners to exchange and share their experiences, new ideas and research results in all aspects of the main workshop themes and tracks, and to discuss the practical challenges encountered and the solutions adopted. The book is organized as a collection of papers from the AIMLA 2021, DBDM 2021, CCNET 2021, NLTM 2021 and IOTBS 2021.

We would like to thank the General and Program Chairs, organization staff, the members of the Technical Program Committees and external reviewers for their excellent and tireless work. We sincerely wish that all attendees benefited scientifically from the conference and wish them every success in their research. It is the humble wish of the conference organizers that the professional dialogue among the researchers, scientists, engineers, students and educators continues beyond the event and that the friendships and collaborations forged will linger and prosper for many years to come.

David C. Wyld,
Dhinaharan Nagamalai (Eds)

General Chair	Organization
David C. Wyld,	Southeastern Louisiana University, USA
Dhinaharan Nagamalai (Eds)	Wireilla Net Solutions, Australia

Program Committee Members

Abdelhadi Assir,	Hassan 1st University, Morocco
Abdelhak Merizig,	Mohamed Khider University, Algeria
Abdullah,	Adigrat University, Ethiopia
Abir Messaoudi,	lecturer university of Quebec, canada
Addisson Salazar,	Universitat Politècnica de València, Spain
Adrian Olaru,	University Politehnica of Bucharest, Romania
Ágnes Vathy-Fogarassy,	University of Pannonia, Hungary
Ahmed A. Elngar,	Beni-Suef University, Egypt
Ahmed Kadhim Hussein,	Babylon university babylon city, Iraq
Akhil Gupta,	Lovely Professional University, India
Akhlaq Ahmad,	Umm Al Qura University, Saudi Arabia
Alexander Gelbukh,	Instituto Politécnico Nacional, Mexico
Amit Mishra,	Baze university, Nigeria
Amit Prakash Singh,	GGSIPU, Delhi, India
Anand Nayyar,	Duy Tan University, Viet Nam
Anouar Abtoy,	Abdelmalek Essaadi University, Morocco
Aridj Mohamed,	Hassiba Benbouali University, Algeria
Arjav A. Bavarva,	RK University, India
Arjav Bavarva,	RK University, India
Arnaud Soulet,	University of Tours, France
Attila Kertesz,	University of Szeged, Hungary
Bernard Cousin,	University of Rennes, France
Bharat Bhushan Agarwal,	I.F.T.M University, India
Bo Li,	Harbin Institute of Technology, Weihai, China
Bogdan CAUTIS,	University Paris-Saclay, France
Brahim Lejdel,	University of El-Oued, Algeria
Brigitte Jaumard,	Concordia University, Canada
Cagdas Hakan Aladag,	Hacettepe University, Turkey
Carlos Guardado da Silva,	University of Lisbon, Portugal
Cheng Siong Chin,	Newcastle University, Singapore
Chih-Hung Wang,	National Chiayi University, Taiwan
Christina Politi,	University of Peloponnese, Greece
Cristina Rottondi,	Politecnico di Torino, Italy
Daniel Hunyadi,	Lucian Blaga University of Sibiu, Romania
Dário Ferreira,	University of Beira Interior, Portugal
Dhamyaa Saad Khudhur,	Mustansiriyah University, Iraq
Dinesh Reddy,	SRM university, India
Domenico Rotondi,	FINCONS SpA, Italy
El murabet Amina,	Abdelmalek Essaadi University, Morocco
El-Sayed M. El-Horbaty,	Ain Shams University, Cairo, Egypt
Elżbieta Macioszek,	Silesian University of Technology, Poland
Eng Islam Atef,	Alexandria University, Egypt
Ez-zahout Abderrahmane,	Mohamed V University, Morocco

Michele Albano,	Aalborg University, Denmark
Mi-Song Chen,	Da-Yeh University, Taiwan
Mohammad Ashraf Ottom,	Yarmouk University, Jordon
Morris Riedel,	University of Iceland, Iceland
Morteza Alinia Ahandani,	University of Tabriz, Iran
Mu-Song Chen,	Da-Yeh University, Taiwan
Nadia Abd-Alsabour,	Cairo University, Egypt
Naren J,	Professional Researcher, India
Narinder Singh Goria,	Punjabi University, India
Narinder Singh,	Punjabi University, India
Nihar Athreyas,	Spero Devcies Inc, USA
Nikola Ivković,	University of Zagreb, Croatia
Ohyun Jo,	Chungbuk National University, Korea
Okba Kazar,	University of Biskra, Algeria
Osamah Ibrahim Khalaf,	Al-Nahrain University, Iraq
P.V. Siva Kumar,	VNR VJIET, India
Pavel Loskot,	ZJU-UIUC Institute, China
Peiman Mohammadi,	Islamic Azad University, Iran
Pratiyush Guleria,	Nielit shimla himachal pradesh, india
Qi Zhang,	Shandong University, China
Rafik Hamza,	NICT, Japan
Ragupathy,	Annamalai University, India
Rajeev Kaula,	Missouri State University, USA
Rajkumar,	N.M.S.S.Vellaichamy Nadar College, India
Ramadan Elaiess,	University of Benghazi, Libya
Ramayah,	Universiti Sains Malaysia, Malaysia
Ramgopal Kashyap,	Amity University Chhattisgarh, India
Ravikumar CV,	VIT University, India
Renjith V Ravi,	M.E.A Engineering College, India
Ricardo Branco,	University of Coimbra, Portugal
Rodrigo Pérez Fernández,	Universidad Politécnica de Madrid, Spain
S.P.Vimal,	Sri Ramakrishna Engineering College, India
S.Taruna,	JK Lakshmipat University, India
Saad Aljanabi,	Al- Hikma College University, Iraq
Sabina Rossi,	Universita Ca' Foscari Venezia, Italy
Sahar Saoud,	Ibn Zohr University, Morocco
Said Nouh,	Hassan II university of Casablanca, Morocco
Saida Bouakaz,	Claude Bernard University Lyon, France
Samir Kumar Bandyopadhyay,	University of Calcutta, India
Sara M. Mosaad,	Helwan University, Egypt
Sathyendra Bhat J,	St Joseph Engineering College, India
Satish Gajawada,	IIT Roorkee Alumnus, India
Sebastian Floerecke,	University of Passau, Germany
Seppo Sirkemaa,	University of Turku, Finland
Shahid Ali,	AGI Education Ltd, Auckland
Shahram Babaie,	Islamic Azad University, Iran
Shashikant Patil,	SVKMs NMIMS, India
Shervan Fekri-Ershad,	Islamic azad university, Iran
Shing-Tai Pan,	National University of Kaohsiung, Taiwan
Siarry Patrick,	Universite Paris-Est Creteil, France
Smain Femmam,	UHA University, France
Souraya Hamida,	University of Batna, Algeria

Technically Sponsored by

Computer Science & Information Technology Community (CSITC)

Artificial Intelligence Community (AIC)

Soft Computing Community (SCC)

Digital Signal & Image Processing Community (DSIPC)

Organized By

Academy & Industry Research Collaboration Center (AIRCC)

International Conference on AI, Machine Learning and Applications (AIMLA 2021)

9th International Conference on Database and Data Mining (DBDM 2021)

8th International Conference on Computer Networks & Communications (CCNET 2021)

International Conference on NLP & Text Mining (NLTM 2021)

International Conference on IOT, Big Data and Security (IOTBS 2021)

HOW MANY FEATURES IS AN IMAGE WORTH? MULTI-CHANNEL CNN FOR STEERING ANGLE PREDICTION IN AUTONOMOUS VEHICLES

Jason Munger and Carlos W. Morato

Department of Robotics Engineering, Worcester Polytechnic Institute,
Worcester, MA, USA

ABSTRACT

This project explores how raw image data obtained from AV cameras can provide a model with more spatial information than can be learned from simple RGB images alone. This paper leverages the advances of deep neural networks to demonstrate steering angle predictions of autonomous vehicles through an end-to-end multi-channel CNN model using only the image data provided from an onboard camera. Image data is processed through existing neural networks to provide pixel segmentation and depth estimates and input to a new neural network along with the raw input image to provide enhanced feature signals from the environment. Various input combinations of Multi-Channel CNNs are evaluated, and their effectiveness is compared to single CNN networks using the individual data inputs. The model with the most accurate steering predictions is identified and performance compared to previous neural networks.

KEYWORDS

Autonomous Vehicles, Convolutional Neural Network, Deep Learning, Perception, Self-Driving Cars.

1. INTRODUCTION

Though the general problem of autonomous steering is understood, more specific issues arise that prevent AI models from being deployed to a 4 or 5 level of autonomy. Steering angle predictions are directly related to external factors in the AV's surrounding environment: The path of the road (or lack thereof), surrounding vehicles, pedestrians, or objects in the immediate vicinity, etc. Even if an AV has self-steering capability, it requires stability and accuracy to drive in a wide variety of environments and react quickly to changes. While CNN's have allowed for advances in steering angle predictions by automatically learning features from RGB images, it is not enough to address the issues above. Humans steer cars using our eyes: We identify important features ahead, determine the location of the road, and discern relative distances to navigate traffic or in complex urban environments. Driving a car is not an innate skill that humans are born with, but rather a learned skill obtained through a multi-faceted "sensor suite" of our bodies and the experience of training in a variety of driving scenarios. Eventually, driving becomes habitual, almost instinctual, and isn't affected by never-before-seen environments. For vehicles to have the level of autonomy necessary to drive without human involvement, they will require a human level of situational awareness and driving skill. They must interact with the surrounding

David C. Wyld et al. (Eds): AIMLA, DBDM, CCNET - 2021
pp. 01-19, 2021. CS & IT - CSCP 2021 DOI: 10.5121/csit.2021.111301

environment delineating between various scene objects in 3D space and deciphering which are important for steering and navigation decisions. There currently exists a debate between major automotive companies regarding the best method of sensor data AVs should rely on to achieve autonomy. On one side, companies like Tesla believe cameras should be the primary method an AV should sense its surroundings given the advances in camera technology and AI, particularly with image recognition [1]. Most other companies working on autonomous vehicles believe LIDAR is necessary to incorporate necessary depth features. However, there are significant downsides that include stability, cost, volume, and resources for visual recognition [2].

In parallel with the advances of steering angle prediction, deep learning has also made vast strides in other areas of computer vision including image segmentation and depth estimation. Image segmentation allows for the discrete detection and/or classification of objects within an image down to the level of individual pixels. This not only allows the model to detect the presence of a particular object, but also provides an accurate mapping of the location and boundaries of the object in an image. Depth estimations of the environment can be achieved through the use of stereo vision cameras or through LIDAR, however, deep learning has been able to achieve similar depth predictions using only 2D RGB images as demonstrated in [20] and [21].

With these considerations and advances in deep learning in mind, questions arose regarding the amount of useful information 2D mononuclear RBG images can provide for steering angle predictions:

1. Can image segmentation and depth estimates provide enhanced signalling features to a model to improve the accuracy and robustness of steering angle predictions?
2. Can the segmentation and depth estimates generated from RGB images *alone* provide sufficiently significant signalling power to an end-to-end steering angle prediction network?
3. Is it possible to extract and synthesize the outputs of independently developed pre-trained (off-the-shelf) models to use as inputs to another network?
4. What architecture provides the best performance with this extended dataset?
5. What impact do each of the additional features have on the overall steering prediction performance?

These questions are explored in this paper using a proof-of-concept neural network and evidence is provided demonstrating 2D RGB monocular camera images alone can provide sufficient signalling power to perceive the driving environment and provide accurate end-to-end steering angle predictions.

This paper is organized in the following manner: Section 2 provides a literature review of steering angle prediction methods, Section 3 provides a concise description of the steering angle problem this paper addresses; Section 4 details the proposed solution of using Multi-Channel CNNs to provide additional signals from RGB images for the prediction of steering angles; Section 5 presents and discusses the results of the proposed solution; and Section 6 draws conclusions of the work and provides potential future research areas that could expand and improve on this current work.

2. LITERATURE REVIEW

2.1. Computer Vision

Various methods of steering angle predictions of autonomous vehicles have been researched in recent years that include the use of computer vision and deep neural networks [23]. Though the end goal is the same, the approaches differ dramatically. Computer vision techniques have been applied to raw image data to manually extract relevant features from the frame, for example, road boundaries, and fitting curves or points that estimate the deviation of the vehicle orientation concerning the road as described in [3] or in [4]. While computationally light compared to deep learning techniques, these methods do not provide the robustness or accuracy necessary to provide steering commands to an autonomous vehicle in a multitude of environments and driving conditions.

2.2. Convolutional Neural Networks

Deep neural networks, particularly Convolutional Neural Networks, have provided breakthroughs in this area to automatically extract the features required from input images and map them to the steering angle. As early as 1989, researchers at CMU demonstrated the ability of their vehicle, ALVINN, to determine directions of travel using only a 3-layer neural network with artificially simulated road images [5]. Recent history has further demonstrated the power of CNNs as steering angle predictions have vastly improved such as in NVIDIA Corporation's creation of PilotNet. Here, researchers not only showed CNNs can learn pertinent road features from training data automatically, but they also demonstrated the use of images in an end-to-end system for AV steering [6]. This model has been used as the basis for other researchers looking to replicate or enhance the model such as in [7], who recreated the NVIDIA model architecture and trained it on augmented image data for use on a virtual driving simulator, and [8] who trained the same model himself using image data collected from a webcam taped to his car and reading CAN-BUS data into an Arduino microcontroller. Others, as in [9], utilize different architectures such as a 3D CNN with LSTM layers to include temporal data and use the concept of transfer learning to leverage high performing pre-trained models (i.e. ResNet50) for use in the new application of prediction steering angles with great success. Other variations of spatial and temporal type models have produced many of the state-of-the-art (SOTA) steering angle predictions. [27] Implements a combined CNN/LSTM/FC network with two sets of input images. One image sequence is provided by the Ego vehicle and another sequence is shared from a second vehicle ahead of it over a vehicle-to-vehicle (V2V) communication system. Both image sequences are used as inputs to the network to predict the steering angle of the Ego vehicle. [25] Uses spatio-temporal convolutions (ST-Conv) with ConvLSTMs to extract features at multiple levels of video sequences and inputs the information into an LSTM to predict steering angle, torque, and speed of the vehicle. [26] Implements Event Cameras to obtains asynchronous frames of pixels depicting changes in motion. Sequences of frames over a specified time interval are collected and used as inputs to a CNN for feature extraction and Fully Connected Network for steering angle predictions. [28] combines CNN and Conv-LSTMs to extract spatiotemporal features at varying levels and combines them with future steering angle information during training to predict the steering angle of the current time. [24] Utilizes Hierarchical Reinforcement Learning (HRL) through the use of a manager and worker network known as the Feudal Steering Network. This network uses a CNN+LSTM+FC architecture to obtain steering angle predictions from frames of a video sequence.

2.3. Multi-Modal Networks

While steering angle prediction neural networks produced unprecedented results, they still suffered from robustness and optimal accuracy issues due to varying driving conditions, illuminations, shadows, and road geometries that inhibit the ability of the neural network to extract necessary features. Many researchers have begun to utilize the concept of multi-modal end-to-end networks to leverage more information from the surrounding environment from cameras and onboard sensors to bring the AV closer to the context in which it is driving. [10] and [11] utilize auxiliary tasks or networks to include additional side models that perform tasks such as image segmentation and optical flow to be used to input features into a network similar to [9] with a 3D CNN + LSTM. [11] transfers low, middle, and high-level features of each auxiliary branch at the same level as the primary CNN model while [10] combines the optical flow and segmentation information as additional inputs with the original image. Both methods yielded better results than the baseline CNN architectures with raw images as the only inputs.

Alternative methods have been proposed, such as [12], where image data is fused with depth information from LIDAR sensors. [13] uses two separate CNN streams to extract spatial information from a processed image and temporal information from pre-calculated optical flow features and merges them before passing through an MLP regression network. [14] utilizes a multitask network using image inputs as well as speed sequences to predict the steering angle and speed of the AV accurately. Researchers have utilized multi-modal networks for other autonomous vehicle applications such as in [15] where LIDAR front view, bird view, and raw front camera image data are processed and fused for 3D object detection. [29] Uses a combination of ConvLSTMs to generate future frame predictions, thereby obtaining future steering angle estimates, and combines auxiliary data in the form of image segmentation to predict steering angles.

All of the techniques presented have shown that while CNNs have been a powerful tool in end-to-end steering angle predictions, including more contextual information of the AV's environment is crucial to improving the robustness and accuracy of these predictions.

3. PROBLEM DESCRIPTION

A vital problem for autonomous vehicles is the ability to accurately and reliably predict steering angles in any number of different driving situations. A complex task in and of itself, it must also accomplish this with affordable technology for fully autonomous cars to become mainstream. Current technology has enabled the areas of image feature extraction, object detection and recognition, and sensor fusion that can combine to create autonomous systems. Given the expensive nature of this technology in terms of computational resources and price, it is difficult to assemble into a single package ready for level 4 or 5 deployments. While steering angle predictions have become accurate over time, current models still suffer failures due to high vehicle speeds, sub-optimal obstacle avoidance, and the inability to use RGB input as the sole signalling feature as described in [16]. A solution is presented to the steering angle prediction problem to improve accuracy and robustness and enable a reduction in the amount of sensor hardware, data, and software computation necessary to incorporate into an autonomous vehicle. The vehicle will have more capability for fewer resources and cost.

4. Solution

4.1. Strategy

To accurately and robustly predict steering angles, several deep learning models are utilized. By incorporating additional information with the raw RGB image, a better model can be trained by increasing the relevant features of the surrounding environment for the model to use. Humans can identify areas of the environment that constitute a driving surface and determine the distance between ourselves and other objects due to our stereo vision system. To create an artificial model with similar capabilities, several neural networks are stacked together, each performing the function it was trained for, and pass the processed information into a combined "super" network that learns features from the outputs of each individual network. In this case, we want to simulate the ability to identify the driving path and perceive the depth of objects using only the data from an onboard camera system. We obtain this supplemental data using semantic segmentation and depth maps. Neural networks have been developed for each task and leveraged in this model. Figure 1 shows the topology of the proposed full model.

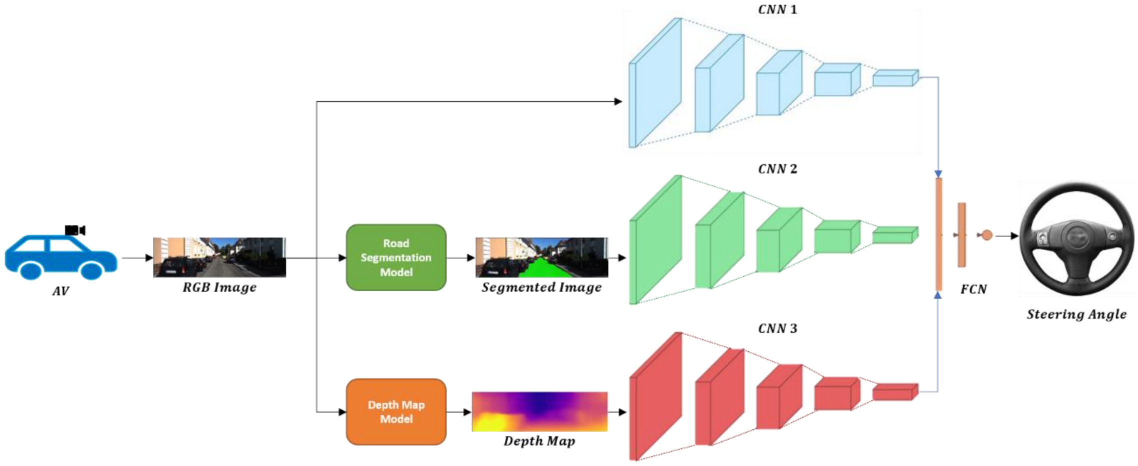


Figure 1. Topology of End-to-End Steering Angle Prediction Neural Network

4.2. Leveraged Pre-Trained Models

Input RGB images taken from a front-view camera are pre-processed through a segmentation model and depth model before passing into individual CNN towers of the primary network. Semantic segmentation allows for pixel-wise object detection and recognition of an image, enabling an autonomous system to gain more information about its environment and allows the system to find only objects for the specific context where the system is used. In the self-driving steering angle case, the AV needs to identify the road and how the driving path is changing. A pre-trained network called RoadSeg parsed through the entire RGB dataset producing another set that includes segmentation arrays in a probabilistic form. Each pixel is assigned a probability as a label of either "road" or "not road." The ability to delineate the road boundaries from the rest of the scene will provide additional signals to the model to learn how the road is changing at different steering angles. A pre-trained depth model trained from monocular camera video data described in [17] allowed for single image frames as input and resulted in an output dataset of depth predictions of objects in the image in the field of view of the camera. The output of this model is an array of values corresponding to the depth predicted for each pixel in the image. This data will require additional processing to scale the data for use as a depth map in the main networks as discussed in 4.4.3. For the multi-channel CNN model demonstration, these models are assumed deployable and capable of generating data at a high enough speed to be utilized in real time and considered "black boxes" that perform a specific function whose details are not the subject of this project. Details of how they work can be reviewed in [18] and [17]. The outputs of the pre-trained models are shown in Figure 2. The top image shows the original input, the middle shows the road segmentation output, and bottom image shows the depth output. Note the image representations are not what the model actually "sees." The outputs of each have been processed to allow the human eye to understand them as an image. The output of the segmentation model consists of an array of probabilities each pixel belongs to the road class or not. The depth model outputs an array of values representing a distance map of the objects in the image. The values are scaled to a reference distance the model was trained with.

4.3. Full Network Architecture

4.3.1. RGB Image CNN

With the RGB images, semantic segmentation images, and depth map datasets created, a CNN model for each type of data is created and optimized to predict steering angles as accurately as possible. The initial inspiration for the models is the PilotNet architecture created by NVIDIA. For the RGB image CNN, a modified version of PilotNet was implemented. The architecture was largely similar, however, given the slight image size difference, a model with less parameters was constructed. The model consists of a 3 channel RGB input, a normalization layer (not part of the main architecture), 5 convolutional layers with decreasing kernel size (5x5 to 3x3) and stride (2x2 to 1x1) and increasing filters (24, 36, 48 and 64), a flattening layer, followed by a fully connected network with 80, 40, 10 and 1 neurons in each layer, respectively. Each layer used the ReLu activation function. The model can be seen in Figure 3 on the left and compared to the PilotNet model on the right.

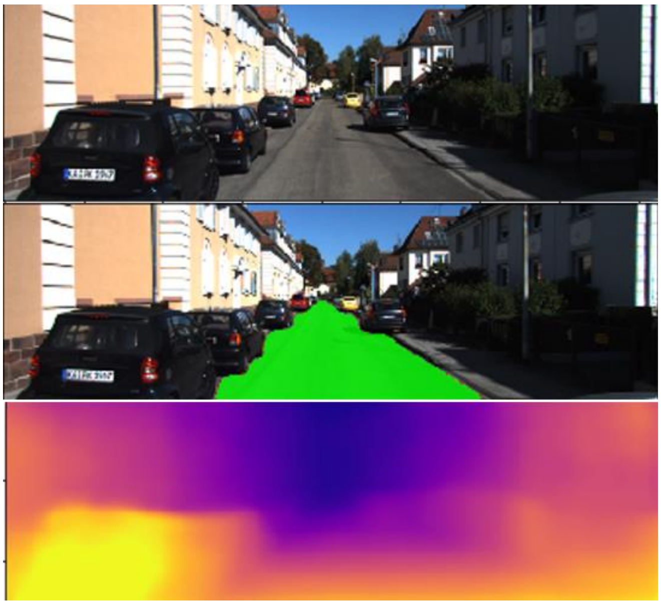


Figure 2. **Top:** Raw RGB Image, **Middle:** Road Segmentation Image, **Bottom:** Depth Map (image representation)

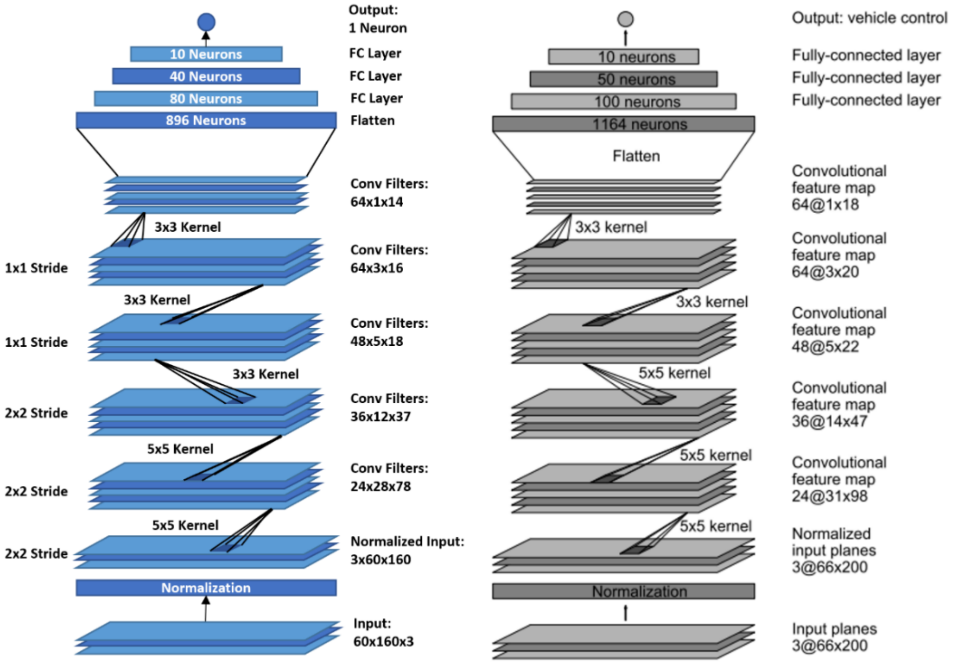


Figure 3. **Left:** Model Architecture for RGB input Multi-Channel CNN, **Right:** NVIDIA PilotNet Model Used as Inspiration and Comparison

4.3.2. Road Segmentation CNN

A CNN for the segmentation input was created following a similar layout as the RGB/PilotNet models though this time a VGG style layer block was followed. This model only required a single normalized channel input, followed by 2-layer blocks consisting of 2 Convolution layers with a kernel size of 3x3 and stride of 1x1 and a Max Pooling layer of size 2x2 and stride of 2x2. The number of filters increased in each subsequent layer from 24, 36, 48, and 64. Again the feature extractor outputs were flattened and input into a small 2 layer fully-connected network of 10 neurons and 1 output neuron. A small dropout of 0.1 was applied to the first FC layer and ReLu activation function was used for each layer. The segmentation CNN model can be seen on the Left of Figure 4.

4.3.3. Depth Map CNN

Finally, a CNN for the depth map input was created in the style of the previous models. Again, this model used a single normalized channel input into 2 VGG style blocks, flattened, and passed into a similar fully-connected network as the RGB model. The primary difference in this model is the number of filters, which range from 32 to 48, and the Dropout rates of 0.5, 0.4 and 0.4 applied to the first 3 fully-connected layers, respectively. ReLu was implemented as the activation function. The depth map CNN model can be seen on the Right of Figure 4.

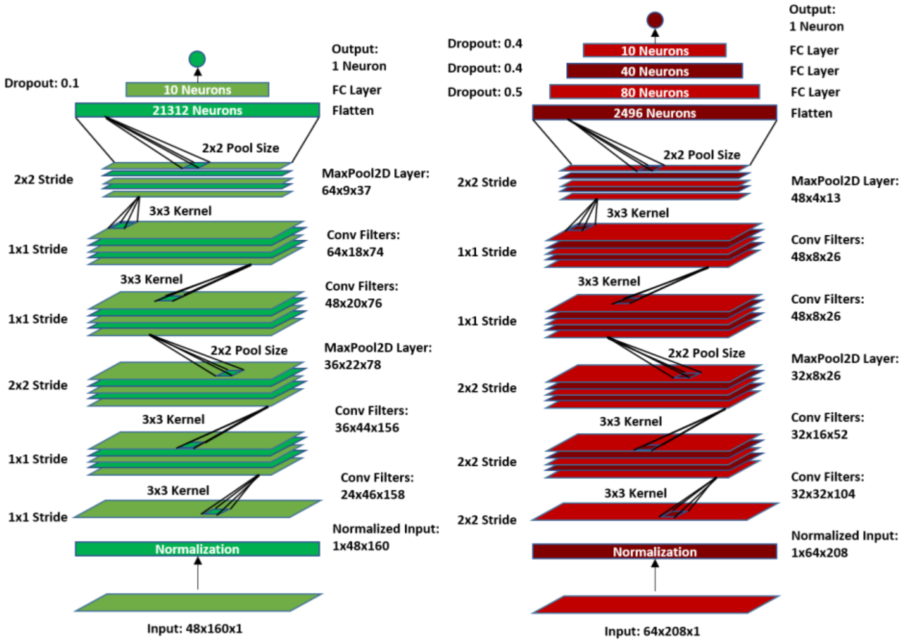


Figure 4. **Left:** Model Architecture for Seginput Multi-Channel CNN, **Right:** Model Architecture for Depth Input of Multi-Channel CNN

4.3.4. Full RGB + Segmentation + Depth Multi-Channel CNN

With the individual CNNs created and trained, a full Multi-Channel CNN can be built. To accomplish this, the fully-connected layers of each CNN was removed leaving only the flattened layers as the outputs. These outputs were merged into a single layer through concatenation. A new fully-connected network was installed consisting of 3 dense layers of 600, 300, and 60 neurons and a single neuron for the steering angle output. A dropout rate of 0.3 was added to each dense layer prior to the output. ReLu activation function was implemented. The weights of the individual networks were left unfrozen to allow for new weights to be computed. The model architecture for the full RGB + Segmentation + Depth CNN network can be seen in Figure 5.

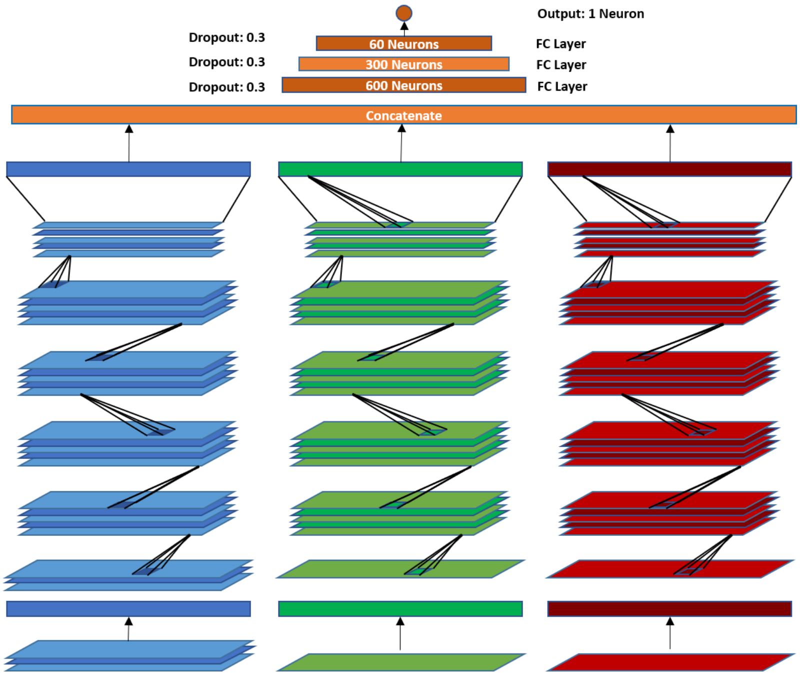


Figure 5. Model Architecture for Full RGB + Seg + Depth Multi-Channel CNN

4.3.5. RGB + Segmentation Multi-Channel CNN

After the full Multi-Channel CNN is built, it is simple to create different combinations of the model to explore how different image representations pair with another. An RGB Image and Segmentation Image Multi-Channel CNN was built nearly identical to the full network only with the depth channel removed. The dropout rate was reduced to 0.2 for this implementation. The model can be seen in Figure 6.

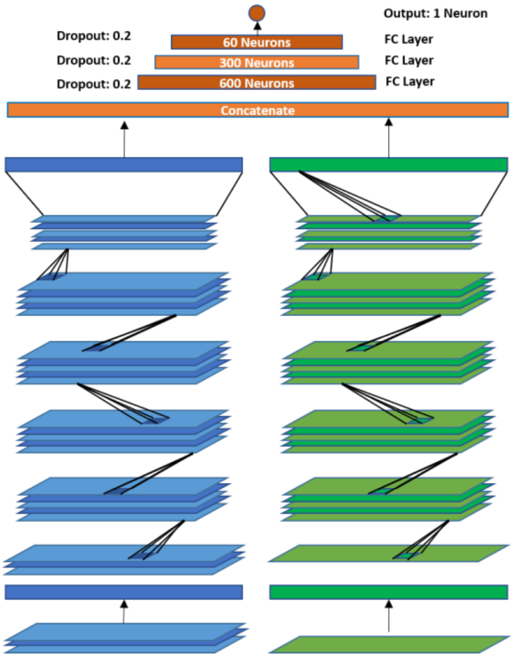


Figure 6. Model Architecture for RGB + Seg Multi-Channel CNN

4.3.6. RGB + Depth Multi-Channel CNN

An RGB + Depth Multi-Channel CNN was built next by removed the Segmentation channel from the full network. All else is identical to the RGB + Segmentation model. The architecture can be seen in Figure 7.

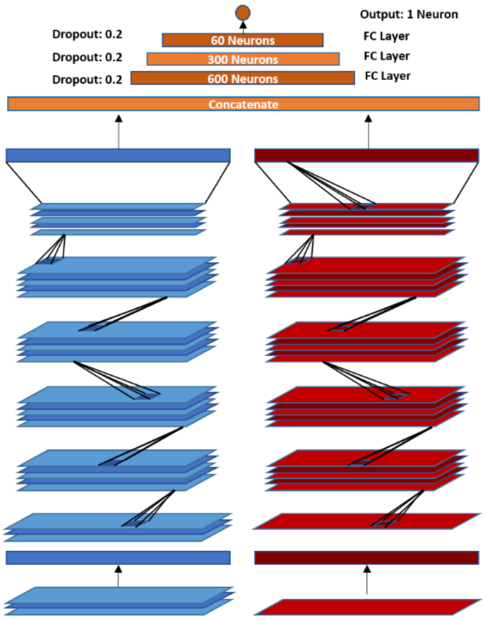


Figure 7. Model Architecture for RGB + Depth Multi-Channel CNN

4.3.7. Segmentation + Depth Multi-Channel CNN

Finally, a Segmentation + Depth Multi-Channel CNN was created by removing the RGB channel from the full model. This time the fully-connected network had less neurons: 80, 40, 10, and 1, respectively. The dropout rate was increased to 0.3 and all else was identical to the full Multi-Channel Model. The architecture can be seen in Figure 8.

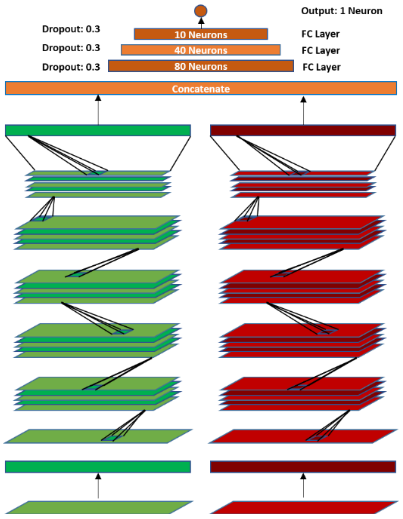


Figure 8. Model Architecture for Seg + Depth Multi-Channel CNN

4.4. Dataset & Data Pre-Processing

All networks trained using the Udacity self-driving dataset used in their self-driving steering angle prediction competitions consisting of 33,808 RGB images of size 640x480. Using this dataset provides a basis for comparison when evaluating model performance against available benchmark performances. The model used an 80/20 training/test dataset split. Each dataset for the individual models required various pre-processing steps given each model was trained on different size images and required their own normalization procedure.

4.4.1. RGB Images

The raw RGB images of the Udacity dataset were cropped the remove the top half to focus only on the road and remove background features that were not pertinent to all images. These images were then down-sampled by a factor of 16 to a final size of 160x60. The pixel values of the images were normalized to value between 0 and 1 by dividing each pixel by 255.

4.4.2. Segmentation Images

For the RoadSeg model described in detail in [18], the required image input size was 640x192 while the depth map model required an image size of 416x128. On top of this, the images needed to be prepared in a manner that resembled the dataset used in each model. Both the segmentation and depth map models were trained using the KITTI dataset [22]. Given this, the Udacity images were cropped to have a similar FOV and aspect ratio of the KITTI images to ensure the inputs were as close the original as possible. Once the cropping operation was complete on the RGB set, a new set was created for the segmentation and depth maps by resizing to the required input sizes. The output of the segmentation model was down-sampled to a final size of 160x48. The segmentation images were normalized such that the pixel values ranged between 0 and 1 representing the probability a pixel belonged to the class of “not road” or “road”, respectively.

4.4.3. Depth Maps

The output depth maps of the depth model were normalized in a similar manner used in by [17] to bound each value between 0 and 1 with values closer to 1 being nearer to the camera and values closer to 0 being farther away. This provides a uniform measurement of relative distances across all input images and provides the model with a synthesized depth signal that can be used to conjunction with the segmentations and RBG images to provide a full spatial representation of the driving scene.

4.5. Training Process

4.5.1. Framework

The Keras API with Tensorflow GPU-supported Backend was used to build, train, and evaluate all models in pursuit of this task.

4.5.2. Loss Function

The loss function chosen for optimization is the Mean Squared Error (MSE) in Equation 1. The MSE function is a measure of the average errors between the ground truth and model predictions. By minimizing this function, model performance is improved. This function also penalizes higher errors more severely due to the squaring term. This is advantageous when low errors are important and high errors are undesirable, exactly as required in the steering angle prediction task.

$$MSE = \frac{1}{N}\sum(y_i - \hat{y})^2 \tag{1}$$

4.5.3. Evaluation Metrics

While the model is optimized on the MSE function, other metrics are used to evaluate performance. Root Mean Squared Error (RMSE) as shown in Equation 2 is almost identical to MSE with the only difference being the square root being applied. RMSE provides the same advantages as MSE, however, it also provides a metric in the units we care about (steering angle) and serves as the standard deviation of the data. The spread of predictions indicates how closely the data surrounds the regression line. A lower value indicates a more accurate model.

$$RMSE = \sqrt{\frac{1}{N}\sum(y_i - \hat{y})^2} \tag{2}$$

Mean Absolute Error (MAE), shown in Equation 3, is also used as a performance metric. MAE provides the average error the model is predicting giving each prediction equal weight. Using the MAE together with RMSE, the model performance can be better understood. While MAE will always be smaller than RMSE, the magnitude of the difference informs how high or low the variance is in a set of predictions. Higher magnitudes indicate more variance of the errors, while lower magnitudes (MAE and RMSE are closer to each other) indicate less variance of the errors.

$$MAE = \frac{1}{N}\sum|y_i - \hat{y}| \tag{3}$$

4.5.4. Optimizer

For training, the Adam optimizer was used. Adam is commonly used for training neural networks due to its adaptive learning ability and computational efficiency. It was found that the default parameters of the Adam optimizer were sufficient for training with the exception of the learning rate which was set to 0.0001 for all models.

4.5.5. Epochs & Batch Size

The models were trained for a range of 20 to 50 epochs depending on the performance. Some models converged earlier while others needed more time and exposure to the data. Each model was trained using a batch size of 50 samples which worked well regardless of the input data or model architecture.

4.5.6. Callbacks

A checkpoint callback was implemented to save best model and weights as they were achieved. This was done by continuously looking at the validation loss after each epoch to determine if model performance improved. This allowed the best model and weights during a training run to always be obtained regardless if the training process began to plateau or degrade by the end. These saved models also served as the individual channel CNN models in the main network. It should be noted that while the best model and weights were always saved, the models were fine-tuned to ensure no underfitting or overfitting conditions occurred.

5. RESULTS

All seven CNN models were trained and evaluated with their performance documented in Table 1. For the individual models, the RGB Network performed the best while the RGB+Depth Network performed the best of the multi-channel networks and had the best performance overall. The worse performing individual network was the Segmentation network which also was involved in the worst performing multi-channel network in the Seg + Depth model. In order to understand the model results more, a plot of the real vs predicted steering angles as well as their errors was created for each network and can be seen in Figure 9.

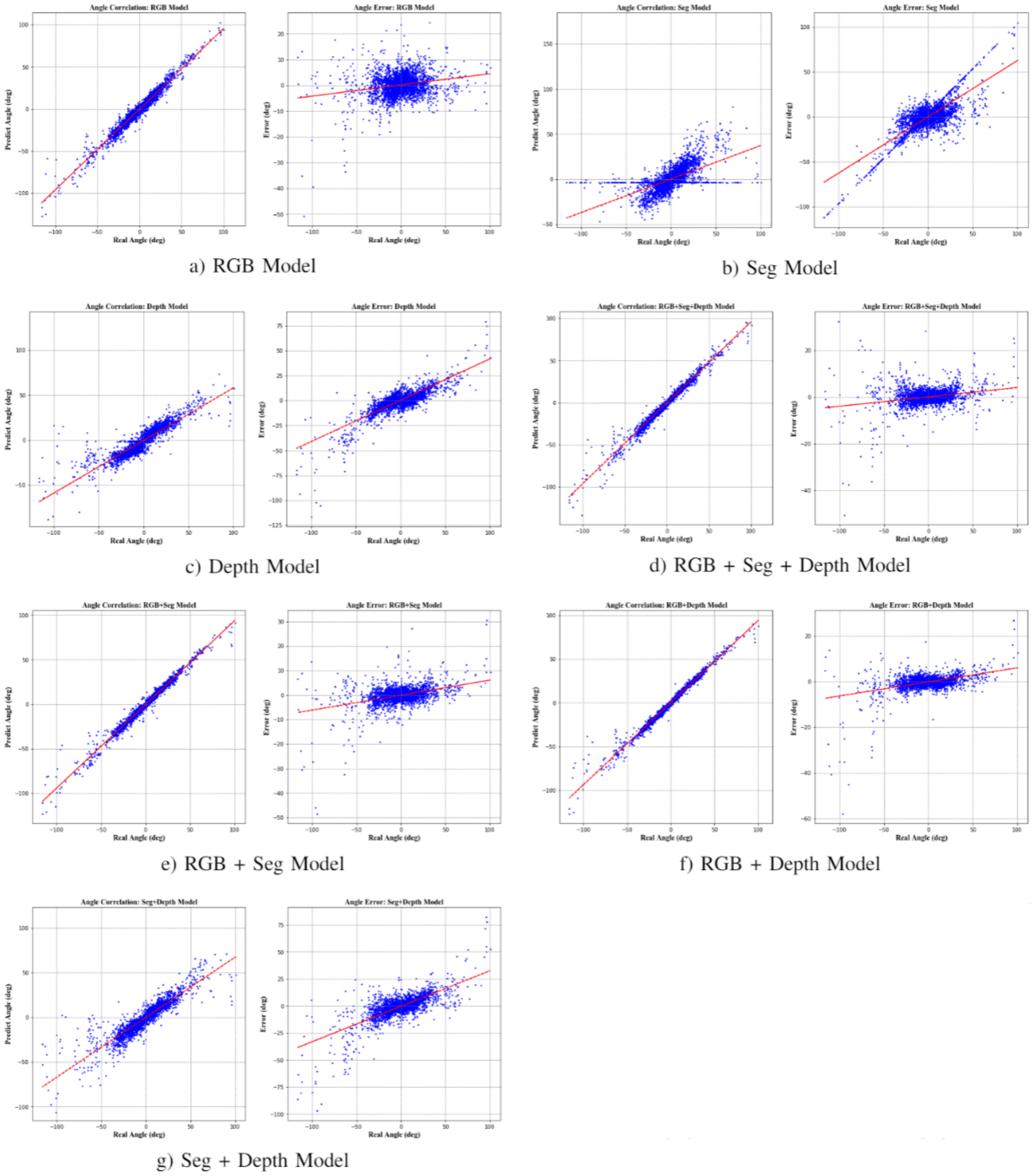
Table 1. Steering Angle Prediction Results

Network Architecture	**MAE (°)**	**RMSE (°)**
RGB	2.15	3.30
Seg	5.97	12.20
Depth	4.01	7.88
RGB + Seg_+ Depth	1.25	2.56
RGB + Seg	1.19	2.43
RGB + Depth	**1.02**	**2.32**
Seg_+ Depth	3.38	7.07

The bias and variance of the predictions can be seen in each model's plot. A perfect correlation plot would show a trendline with a slope of 1 and y-intercept of 0, and a perfect error plot would show a trendline with 0 slope and 0 y-intercept. The "good" performing models can be seen in Figure 9 a, d, e, and f and "bad" performing models can be seen in b, c, and g. These plots illustrate the relationship of the tabular performance values with how the data is distributed around the trendlines. For the "good" models, the predicted values are clustered tighter around the trendline and the dispersion of error magnitudes are smaller indicating lower variance. The RGB + Depth Model plots (f) demonstrate how the MAE and RMSE is the lowest of all with the predictions being the most tightly situated around the trendline as well as having the tightest error plot. The Seg Model on the other hand, shows why it performed the worst. Not only are the predictions severely off and widely dispersed, the trendline has less than half the slope of a well performing model indicating a problem with the data inputs. It was found the pre-trained RoadSeg Model was not performing well on the Udacity dataset even though the images were prepared as close to the KITTI dataset as possible. Upon further inspection of random samples, the segmentation of the road was spotty at best with some images having good results with the road fully segmented, while others had mere blotches of road, and many had no segmentation at all. This would explain the performance of this model as passing an array of zeros would not provide any useful information for steering prediction. In fact, it can be seen a steering angle of zero (or near zero) was predicted often over the entire range of possible steering angles. This indicates more work needs to be done to truly test the impact of the segmentation in multi-channel CNNs and should not be discounted as a means of improving steering performance.

One of the most interesting aspects of applying the multi-channel CNN models to the inputs, is the ability for the models to learn how to make use of this "bad" data and calculate a set of weights that provide better steering angle prediction performance than it could in the individual models. As long as the RGB image inputs were present in a multi-channel CNN, the model performed better as whole than the individual channels alone. This can be seen when the segmentation and depth models were paired together. The two worst performing individual models created the worst performing multi-channel model, which was expected. However, it was expected the full RGB + Seg + Depth model would provide the best overall performance, yet it was the RGB + Depth model that came out on top. This demonstrates that one, multi-channel CNNs can extract useful features from different types of input data, and two, the introduction of depth information provides another spatial component that assists in predicting steering angles.

Another model performance is evaluated from a different perspective by plotting the predicted steering angles over a trajectory. Predictions from each model were calculated and plotted over a subset of the Udacity test sample data and can be seen in Figure 10. In Figure 10a, a plot of all individual CNNs are overlaid on the ground truth data, b overlays all multi-channel CNN model predictions over ground truth, and c and d overlay the best (RGB+Depth) and worst (Seg) predictions over the ground truth, respectively. Significant performance increases can be seen between the single and multi-channel models again reinforcing the idea different image inputs are useful for the steering prediction task. The contrast between the best and worst model is stark. The best model is able to tightly follow the ground truth data over a wide variety of angles (-50 to +70 degrees) while the worst is unable to reliably follow the trajectory. It should be noted that all models struggled with large steering angles though obviously some are better than others. This may be attributable to a lack of data for larger angles. The dataset is oversampled with smaller angles as that is what most driving conditions require. More data containing larger steering angles will help this problem through upsampling of less frequent angles by adding flipped images of existing data or downsampling more frequent smaller angles to remove their bias. Downsampling may come at the expense of less accurate overall predictions which only emphasizes the need for more training data.



a) RGB Model

b) Seg Model

c) Depth Model

d) RGB + Seg + Depth Model

e) RGB + Seg Model

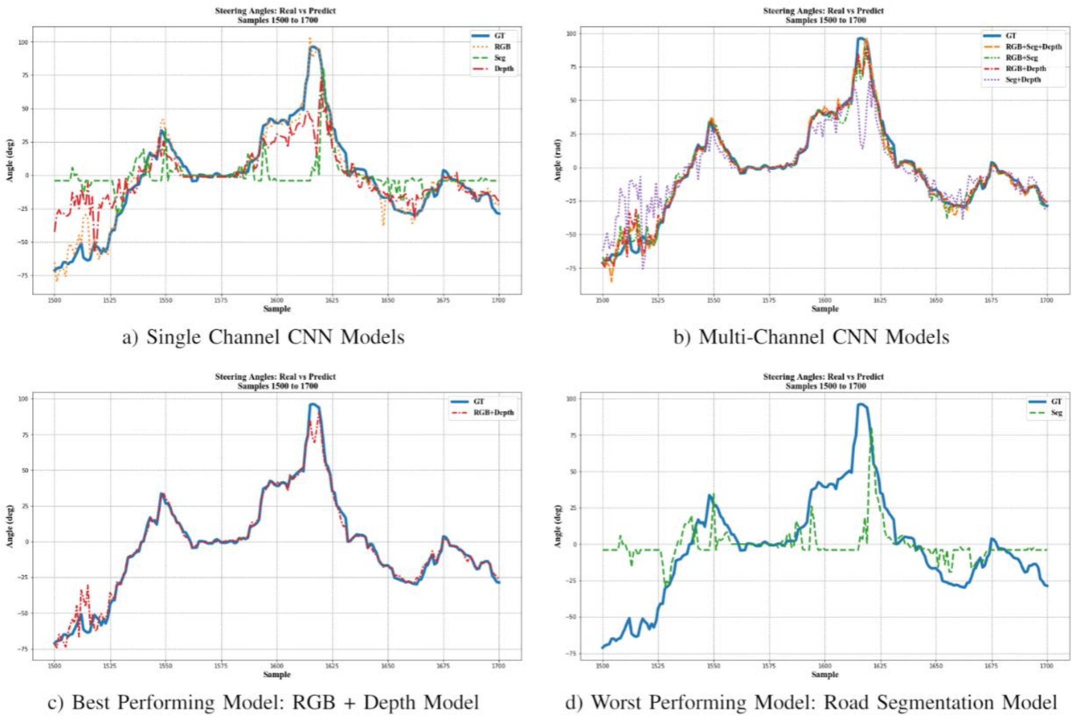
f) RGB + Depth Model

g) Seg + Depth Model

Figure 9. Steering Angle Correlation and Error Between Ground Truth and Predictions

To evaluate the efficacy of multi-channel CNNs for steering angle predictions, the performance of the best model (RGB + Depth) was compared to previous steering angle prediction models trained on the Udacity dataset. Table 2 shows the comparison. Previous models utilized CNN models though some implemented variants. These variants include standard structures as in the NVIDIA PilotNet architecture, transfer learning of pre-trained CNN feature extractors, a temporal aspect using 3D CNN or LSTM models, or a network utilizing auxiliary tasks which is most similar to this current work. The addition of a temporal component increased performance however, the FM-Net combined auxiliary networks to provide feature inputs of segmentation and optical flow using a 3D ResNet and LSTM architecture performed best among them [11]. It was shown the RGB + Depth model outperformed all of these models, many considered SOTA, relying only on single image inputs to a simpler CNN architecture. It is suspected that had the

Segmentation inputs were of higher quality, the full network (RGB + Seg + Depth) likely would have provided even better performance though this needs verification through training.



a) Single Channel CNN Models

b) Multi-Channel CNN Models

c) Best Performing Model: RGB + Depth Model

d) Worst Performing Model: Road Segmentation Model

Figure 10. Steering Angle Prediction Results on Udacity Dataset (Sampled from Test Set)

Table 2. Comparison of Methods on Udacity Dataset

Network Architecture	**MAE (°)**	**RMSE (°)**
CNN + FCN (NVIDIA) [19]	4.12	4.83
CNN + LSTM [19]	4.15	4.93
3D LSTM [9]	-	6.44
ResNet50 Transfer [9]	-	4.06
3D CNN [11]	2.56	3.66
MSINet [28]	-	2.81
3D CNN + LSTM [11]	1.86	2.72
Feudal Steering [24]	1.09	2.67
FM-Net [11]	1.62	2.35
RGB + Depth	**1.02**	**2.32**

The results of the models clearly indicate that not only can additional features besides the RGB mononuclear images enhance the signals to the model in the development of a robust and accurate end-to-end steering angle prediction model, but the additional features to enhance the spatial awareness of an autonomous vehicle can be synthesized from the RGB images of an onboard camera alone. This implies the potential for AVs to perceive, learn, and navigate the driving environment from minimal data input. It also demonstrates how individual deep learning models can be trained for separate specific tasks and combined for use in new applications and potentially increasingly complex tasks.

6. CONCLUSIONS

This paper demonstrated the effectiveness of Multi-Channel CNNs using different camera image representations to accurately and reliably make steering angle predictions. Compared to individual CNN models trained on separate inputs, multi-channel CNNs allow for improved performance without the introduction of feature signals other than the images provided from the onboard camera of an autonomous vehicle. It was shown that depth data computed from a pre-trained model in combination with an RGB image provide the best overall steering angle predictions. Networks involving road segmentation provided the worst performance, however, this is most likely due to the pre-trained model's inability to make predictions on different data than the one it was trained on, or the data required an alternate pre-processing procedure to predict effectively. More work is required to investigate segmentation as additional signals and should not be completely dismissed from the results of this paper. The best multi-channel CNN exceeded performance of previous models using various network architectures that included spatial and temporal elements, leveraged transfer learning techniques, and implemented parallel auxiliary networks to feed various levels of features to layers of a single CNN network.

Future work may include refining the multi-channel network to train on other datasets, increasing samples of larger steering angles, or implementing better performing pre-trained models to obtain accurate data inputs. A temporal aspect was not considered for this architecture; however, it is possible to implement data from a series of video frames rather than individual images and still provide a valid model that supports the goal of using camera image data alone to predict steering angles.

ACKNOWLEDGEMENTS

Thank you to WPI for providing a curriculum that enables the study of artificial intelligence and applications to real-world problems.

REFERENCES

[1] "'anyone relying on lidar is doomed,' elon musk says — TechCrunch,"https://techcrunch.com/2019/04/22/anyone-relying-on-lidar-is-doomed-elon-musk-says/, (Accessed on 04/25/2021).

[2] "Lidar vs. camera — which is the best for self-driving cars? —by Vincent Tabora—0xmachina—medium,"https://medium.com/0xmachina/lidar-vs-camera-which-is-the-best-for-self-driving-cars-9335b684f8d, (Accessed on 04/25/2021).

[3] R. Meganathan, A. A. Kasi, and S. Jagannath, "Computer vision based novel steering angle calculation for autonomous vehicles,"*2018 Second IEEE International Conference on Robotic Computing (IRC)*, pp. 143–146, 2018.

[4] U. Venkatasubramanian, S. Amarjyoti, T. Bakshi, and A. Singh, "Steering angle estimation for autonomous vehicle navigation usinghough and euclidean transform,"*2015 IEEE International Conference on Signal Processing, Informatics, Communication and Energy Systems, SPICES 2015,* 04 2015.

[5] D. Pomerleau, "Alvinn: An autonomous land vehicle in a neuralnetwork," in*NIPS*, 1988.

[6] M. Bojarski, D. D. Testa, D. Dworakowski, B. Firner, B. Flepp, P. Goyal, L. D. Jackel, M. Monfort, U. Muller, J. Zhang, X. Zhang,J. Zhao, and K. Zieba, "End to end learning for self-driving cars,"2016.

[7] V. Singhal, S. Gugale, R. Agarwal, P. Dhake, and U. Kalshetti, "Steering angle prediction in autonomous vehicles using deep learning," in*2019 5th International Conference On Computing, Communication, Control And Automation (ICCUBEA)*, 2019, pp. 1–6.

[8] "How a high school junior made a self-driving car — by sullychen — towards data science," https://towardsdatascience.com/how-a-high-school-junior-made-a-self-driving-car-705fa9b6e860, (Accessed on 03/14/2021).

[9] S. Du, H. Guo, and A. Simpson, “Self-driving car steering angle prediction based on image recognition,” 2019.

[10] Y. Chen, P. Palanisamy, P. Mudalige, K. Muelling, and J. M. Dolan, “Learning on-road visual control for self-driving vehicles with auxiliary tasks,” 2018.

[11] Y. Hou, Z. Ma, C. Liu, and C. C. Loy, “Learning to steer by mimicking features from heterogeneous auxiliary networks,” 2018.

[12] Y. Xiao, F. Codevilla, A. Gurram, O. Urfalioglu, and A. López, “Multimodal end-to-end autonomous driving,”ArXiv, vol. abs/1906.03199, 2019.

[13] N. Fernandez, “Two-stream convolutional networks for end-to-end learning of self-driving cars,” 2018.

[14] Z. Yang, Y. Zhang, J. Yu, J. Cai, and J. Luo, “End-to-end multi-modal multi-task vehicle control for self-driving cars with visual perceptions,”*2018 24th International Conference on Pattern Recognition (ICPR),* pp. 2289–2294, 2018.

[15] X.Chen, H.Ma, J.Wan, B.Li, and T.Xia, “Multi-view 3d object detection network for autonomous driving,”*CoRR,* vol. abs/1611.07759, 2016. [Online]. Available:http://arxiv.org/abs/1611.07759

[16] U. M. Gidado, H. Chiroma, N. Aljojo, S. Abubakar, S. I. Popoola, and M. A. Al-Garadi, “A survey on deep learning for steering angle prediction in autonomous vehicles,”*IEEE Access*, vol. 8, pp. 163 797–163 817, 2020.

[17] T. Zhou, M. Brown, N. Snavely, and D. G. Lowe, “Unsupervised learning of depth and ego-motion from video,” in *CVPR,* 2017.

[18] “Github - robertklee/kitti-roadseg: A course project for road segmentation using a u-net convolutional neural network on the kitti road2013 dataset,” https://github.com/robertklee/KITTI-RoadSeg, (Accessed on 04/25/2021).

[19] J. Kim and J. Canny, “Interpretable learning for self-driving cars by visualizing causal attention,” 2017.

[20] K. Cantrell., C. Miller., and C. Morato., “Practical depth estimation with image segmentation and serial u-nets,” in Proceedings of the 6th International Conference on Vehicle Technology and Intelligent Transport Systems - VEHITS, INSTICC. SciTePress, 2020, pp. 406–414.

[21] V. John, “Vision-based steering angle prediction by the fusion of depth and intensity deep features,” in *CVPR,* 2018.

[22] A. Geiger, P. Lenz, C. Stiller, and R. Urtasun, “Vision meets robotics: The kitti dataset,” International Journal of Robotics Research (IJRR), 2013.

[23] H. Saleem, F. Riaz, L. Mostarda, M. A. Niazi, A. Rafiq and S. Saeed, "Steering Angle Prediction Techniques for Autonomous Ground Vehicles: A Review," in IEEE Access, vol. 9, pp. 78567-78585, 2021, doi: 10.1109/ACCESS.2021.3083890.

[24] F. Johnson and K. Dana, "Feudal Steering: Hierarchical Learning for Steering Angle Prediction," 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), 2020, pp. 4316-4325, doi: 10.1109/CVPRW50498.2020.00509.

[25] Chi, L., & Mu, Y. (2017). Deep Steering: Learning End-to-End Driving Model from Spatial and Temporal Visual Cues. ArXiv, abs/1708.03798.

[26] A. I. Maqueda, A. Loquercio, G. Gallego, N. García, and D. Scaramuzza, “Event-based vision meets deep learning on steering prediction for self-driving cars,”2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 5419–5427, 2018

[27] R. Valiente, M. Zaman, S. Ozer and Y. P. Fallah, "Controlling Steering Angle for Cooperative Self-driving Vehicles utilizing CNN and LSTM-based Deep Networks," 2019 IEEE Intelligent Vehicles Symposium (IV), 2019, pp. 2423-2428, doi: 10.1109/IVS.2019.8814260.

[28] T. Wu, A. Luo, R. Huang, H. Cheng and Y. Zhao, "End-to-End Driving Model for Steering Control of Autonomous Vehicles with Future Spatiotemporal Features," 2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), 2019, pp. 950-955, doi: 10.1109/IROS40897.2019.8968453.

[29] F. Munir, S. Azam and M. Jeon, "Visuomotor Steering angle Prediction in Dynamic Perception Environment for Autonomous Vehicle," 2020 IEEE International Conference on Consumer Electronics - Asia (ICCE-Asia), 2020, pp. 1-6, doi: 10.1109/ICCE-Asia49877.2020.9276907.

AUTHORS

Jason Munger is a graduate student in the Department of Robotics Engineering at Worcester Polytechnic Institute. He has a B.Sc. in Mechanical Engineering from the Georgia Institute of Technology. He works full time at NASA's Jet Propulsion Laboratory in Pasadena, California as a Mechanical Engineer on a variety of space-bound instruments and mechanisms.

Carlos W. Morato is Professor in the Department of Robotics engineering at Worcester Polytechnic Institute. He is leading researcher in the fields of human robot interaction, collaborative robots self-driving vehicles, and meta-intelligence for augmented intelligent systems. He has a PhD from the Department of Mechanical Engineering, University of Maryland at College Park, USA, a M.Sc degree in Aerospace Engineering and a M.Sc degree in Computer Science both focused in intelligent robots.

DIVIDE-AND-CONQUER FEDERATED LEARNING UNDER DATA HETEROGENEITY

Pravin Chandran, Raghavendra Bhat,
Avinash Chakravarthy and Srikanth Chandar

Intel Technology India Pvt. Ltd, Bengaluru, India

ABSTRACT

Federated Learning allows training of data stored in distributed devices without the need for centralizing training-data, thereby maintaining data-privacy. Addressing the ability to handle data heterogeneity (non-identical and independent distribution or non-IID) is a key enabler for the wider deployment of Federated Learning. In this paper, we propose a novel Divide-and-Conquer training methodology that enables the use of the popular FedAvg aggregation algorithm by over-coming the acknowledged FedAvg limitations in non-IID environments. We propose a novel use of Cosine-distance based Weight Divergence metric to determine the exact point where a Deep Learning network can be divided into class-agnostic initial layers and class-specific deep layers for performing a Divide and Conquer training. We show that the methodology achieves trained-model accuracy at-par with (and in certain cases exceeding) the numbers achieved by state-of-the-art algorithms like FedProx, FedMA, etc. Also, we show that this methodology leads to compute and/or bandwidth optimizations under certain documented conditions.

KEYWORDS

Federated Learning, Divide and Conquer, Weight divergence.

1. INTRODUCTION

Federated Learning has been proposed as a new learning paradigm to overcome the privacy regulations and communication overheads associated with central training [1,2]. In Federated Learning, a central server shares a global model with participating client devices and the model is trained on the local datasets available at the client device. The local dataset is never shared with the server, instead, local updates to the global model are shared with the server. The server combines the local updates from the participating clients using an Optimization (or Aggregation) Algorithm and creates a new version of the global model. This process is repeated for the required number of communication rounds until the desired convergence criteria are achieved.

Federated Learning differs significantly from traditional learning approaches in terms of optimization in a distributed setting, privacy preserving learning, and communication latency during the learning process [3]. Optimization in Distributed setting differs from the traditional learning approach due to statistical and systems heterogeneity [1]. The statistical heterogeneity manifests itself in the form of non-independent and identical distribution (non-IID) of training data across participating clients. The non-IID condition arises due to a host of reasons that is specific to the local environment and usage patterns at the client. Causes for the skewed data distribution have been surveyed extensively and it has been proven that any real-world scale deployment of Federated Learning should address the challenges around non-IID data. A good example specific to the medical domain can be found in [4]. Several approaches have been

David C. Wyld et al. (Eds): AIMLA, DBDM, CCNET - 2021
pp. 21-33, 2021. CS & IT - CSCP 2021 DOI: 10.5121/csit.2021.111302

studied to address the non-IID heterogeneity. Data Distillation which involves sharing of client data with central server [5, 6], Client specific local models or Personalization layers to customize the last few layers of the global model specific to the client data [7, 8, 9, 10, 11], Novel optimization algorithms are some of these most researched approaches.

Data Distillation techniques violate the strict privacy requirements. Client specific model approach results in multiple models, which does not cater to any specific requirement for a single model for deployment. In this paper, we focus on the Optimization Algorithm approach to address the non-IID challenge. While there are numerous state-of-the-art algorithms like FedProx [12], FedMA [13], FedMAX [14] etc., these approaches are not productized in a large scale to the best of knowledge of the authors. Hence, we focus on the most widely deployed FedAvg algorithm [1] and investigate improving its ability to handle non-IID data to the same level as state-of-the-art algorithms like FedMA, FedProx, FedMAX, etc.

The primary contribution of this paper is proposing a novel Divide-and-Conquer training methodology which in combination with FedAvg is able to meet state-of-the-art performance in simulated environment. Another contribution of this paper is the novel use of the Cosine Distance based Weight Divergence metric to partition the global model into class agnostic initial layers and class-specific deep layers. The two parts of the global model are trained in a mutually exclusive manner while freezing the other part. Under certain documented conditions, this approach also leads to better compute and bandwidth optimization.

The rest of the paper is organized as follows. Section 2 discusses the limitation with vanilla FedAvg algorithm while Section 3 explains the Divide-and-Conquer methodology. We document the simulation environment, experiments, and results in the simulated environment in Section 4 establishing the state-of-the-art credentials of the approach. Finally, we conclude the paper and discuss possible future work in Section 5.

2. FEDAVG AND ITS CHALLENGES

Federated Learning (FL) methods are designed to train over multiple devices, each holding their own data, with a central server driving the global learning objective across the entire network. The standard formulation of FL aims to find the minimizer of the overall population loss [12] shown in EQ1 below.

$$\min_{w} f(w) = \sum p_k \ F_k(w) = E_k \ [F_k(w)], \qquad (EQ1)$$

$$where\ N\ is\ number\ of\ devices, p_k \geq 0\ and \sum_{k} p_k = 1$$

In general, the local objectives measure the local empirical risk over possibly differing data distributions with samples available at each device. In a non-IID environment, the assumption of a global minimizer being representative of the overall population is not valid as every client has its own data distribution which differs from other clients and the overall population. Hence, on each client, a local objective function based on the client's data is used as a surrogate for the global objective function. At each outer iteration, a subset of devices is selected, and local solvers are used to optimize the local objective functions of the selected client. Each client then communicates its local model updates to the central server, which aggregates them and updates the global model accordingly. In addition to the usual hyper-parameters of traditional learning like batch size, optimizer, etc., Federated Learning has additional hyper-parameters like epochs per round (E_p), number of communication rounds, number of participants in each round, and optimization algorithm which can be tweaked for optimal performance.

In FedAvg, the local objective function at client k is $F_k(.)$, and the local solver is the stochastic gradient descent (SGD), with the same learning rate η and number of local epochs used on each client. At each round, a subset K<N of the total clients are selected and run SGD locally for E_p number of epochs, and then the resulting model updates are averaged. The details are summarized below.

Algorithm 1: Federated Averaging Algorithm

```
Input: $K, T, \eta, E_p, w^o, N, p_k, k=1,\ldots,N$

for $t=1$ to $T-1$ do
        Server selects a subset $S_t$ of $K$ clients at random.
        Server sends $w_t$ to all chosen clients
        Each client $k \in S_t$ updates $w_t$ for $E_p$ epochs of SGD
        On $F_k$, with step size $\eta$ to obtain $w_k^{t+1}$
        Each client $k \in S_t$, sends $w_k^{t+1}$ back to server
        Server aggregates the $w$'s as $w_{t+1} = \frac{1}{K} \sum k \in S_t \ w_k^{t+1}$
end for
```

Tuning of the hyper-parameters is a critical requirement for optimal performance of FedAvg. The number of epochs plays a critical role in convergence as a greater number of epochs leads to faster convergence. This comes at the cost of higher compute on client devices but with the benefit of lower communication. However, the high number of epochs has diminishing returns on the speed of convergence in non-IID conditions. For FedAvg, there is a significant drop in reduction of accuracy due to weight divergence [5]. The trade-off between high number of epochs and convergence speed for FedAvg has been addressed in other optimization algorithms like FedProx, FedMA, FedMAX etc. FedProx is very similar to FedAvg but addresses the limitations of the latter by adding a proximal term to client cost functions to limit the impact of local updates within a particular range of global model. This approach allows the number of epochs to be tuned based on the non-IIDness of the client data. While it addresses the weight divergence issue with FedAvg, the convergence speed is slower at higher number of epochs when compared to other state-of-the-art algorithms [6,13,14,15]. FedMA offers the best accuracy and convergence speed in comparison to others but comes with significant compute cost on the client devices. The complexity of this algorithm is also high in comparison with FedAvg or FedProx leading to restrictions on its applicability on certain NN models.

An ideal optimization algorithm should come with the simplicity and elegance of FedAvg, allow for state-of-the-art accuracy in non-IID environments with comparable or better convergence speed. In this work, we present a novel Federated Training methodology that is well suited to handle non-IID challenges using the simple FedAvg algorithm. Our methodology eliminates performance overheads associated with methods like FedMA while achieving comparable accuracy. Since FedAvg is the de-facto standard in majority production deployments, the proposed method can be easily integrated to offer significant accuracy and convergence benefits with little performance overhead.

Note on the terminology: In the rest of the document, clients will be referred to as Collaborators and the server will be referred to as Aggregator, reflecting the role they play in the overall federation.

3. Divide-and-Conquer Training Methodology

The impact of non-IIDness of data in Federated Learning is well researched in literature. A non-IID data environment leads to over-fitting of local models to the skewed training data at individual collaborators resulting in distortion of previously aggregated feature detectors and descent of SGD optimizer to different local minima at different collaborators.

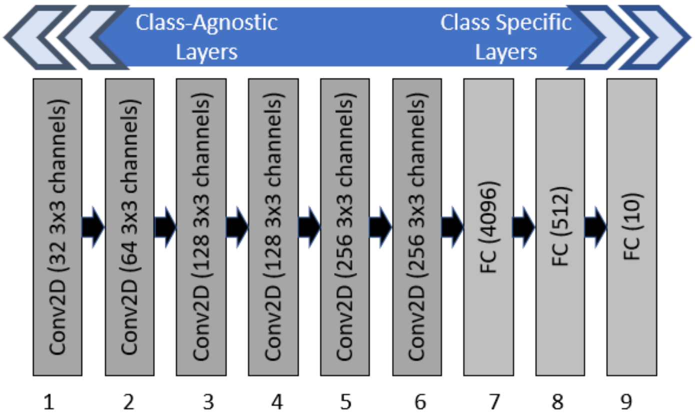


Figure 1. DNN Layer Significance - VGG9 Image Classification Topology

Typically, the initial layers of a Deep Neural Network (DNN) learn low level or class agnostic features and deeper layers are responsible for learning high level or class-specific features as illustrated for a vision architecture, VGG9 [16], in Figure 1. For training paradigms like Transfer Learning \cite{tlearn1}, data scarcity mandates the use of special training methods that learn class agnostic features from generic datasets and learn class specific features for any new tasks by freezing the initial layers. This process of decoupling feature-learning and task-learning has been successfully applied to multiple training tasks including recent advances like Few Shot Learning [17]. This work extends the idea to Federated Learning to address the challenges with non-IID. Our methodology involves splitting the given DNN into two parts, namely (a) Class Agnostic Layers and (b) Class Specific Layers.

The two parts are trained separately. Federated Learning is typically performed using several Communication Rounds (CR), where trained weights from individual collaborators are aggregated together in a central Aggregator. Our proposed method configures collaborators to perform feature-learning and task-learning or fine-tuning in alternate rounds as shown in Figure 2. Weights corresponding to relevant trained layers alone are transferred over to the Aggregator, which results in communication bandwidth reduction. Communication saving is realized during model transfers in both directions (i) Transfer of global models to Collaborators and (ii) Transfer of local trained models from Collaborator to the Aggregator.

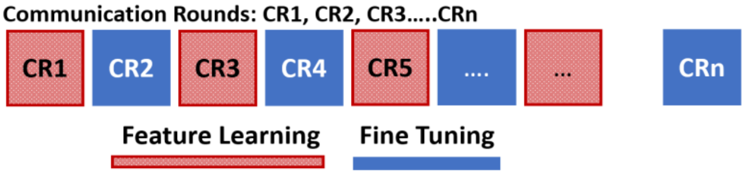


Figure 2. Divide and Conquer Training Methodology using alternate Feature-Learning and Fine-Tuning rounds. CR1, C2, represent the communication rounds.

Class Agnostic layers, comprised of initial layers of the DNN architecture, are trained more aggressively as compared to Class-Specific layers. Class Agnostic layer training is treated similar to feature-learning. Class Specific layers, consisting of deep layers are trained similar to fine-tuning. This ensures that weight divergence across different collaborators, due to non-IIDness of constituent data is minimal and features are insulated from distortion that would otherwise occur due to combined learning of all layers.

While methods like FedProx limit weight divergence, they penalize all layers of the network and hinder learning in Class Agnostic layers. Our approach addresses this by allowing different layers of a network to train differently after grouping initial layers separately from deep layers. Training rounds are configured to alternate between feature-learning and fine-tuning to facilitate learning under non-IID conditions by freezing relevant layers of DNN architecture. At the beginning of a communication round, the aggregator broadcasts the desired hyper-parameter configurations to collaborators, together with specifications for layers to be frozen. The exact point at which a DNN architecture must be broken into two parts is decided based on 'weight divergence' observed from the pre-pass round of training. The key contributions of our paper can be summarized to the following two key points:

- Novel methodology, called Divide-and-Conquer (D&C), to train topology in pairs of feature-learning and fine-tuning steps to handle non-IID conditions.
- Novel use of weight-divergence metric, observed from the pre-pass round of training, to split the given DNN topology into Class Agnostic and Class Specific layers. This metric provides a measure of non-IIDness across participating collaborators as a mapping of the layers of DNN architecture they impact the most.

Choice of layers that are chosen for base class feature-learning as against novel class fine-tuning is a hyper-parameter in Divide-and-Conquer training methodology. Few options for splitting the VGG9 topology is shown in Figure 3. For instance, Divide3, divide at layer3, assigns layers 1 to 3 for learning class agnostic features and remaining layers for learning class or task specific features. This hyper-parameter is dependent on the weight-divergence metric which in turn reflects the non-IIDness of data.

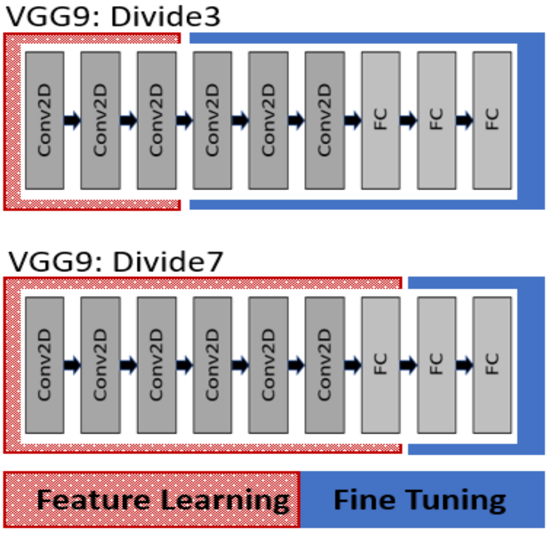


Figure 3. VGG9: Topology Division and assignment of layers to Feature-Learning and Fine-Tuning Groups

After determining an optimal split, feature-learning} and fine-tuning is achieved by control of other hyper-parameters like number of Epochs E_p and Learning rate η. Fine-Tuning round of learning is scheduled using lower E_p and η, which is aligned with the conditions under which FedAvg performs the best in non-IID conditions. Federated Learning at a faster pace is achieved by alternating low-level feature-learning and high-level fine-tuning along with appropriate hyper-parameters as described in the next section.

4. Divide-and-Conquer: Experiments, Results, and Discussion

This section describes the simulation environment, experiments done, and results. The comparison with other state-of-the-art approaches is also captured in the results section to establish the state-of-the-art credentials of our proposed approach.

4.1. Experimental Setup

We present observations from Divide-and-Conquer on VGG9 topology using 3 different non-IID conditions as in [13], which includes coverage for convolutional layers and LSTMs. Classification and NLP models used were also same as [13].

- Classification using Color Skewed CIFAR10 Dataset [19]: CIFAR10 dataset is split into two groups of 5 classes each, with each class assigned uniquely to the two collaborators. To skew the data further using a 95-5% skew pattern, 95% of images in the first group are converted to gray-scale and 5% of images in the second group are converted to gray-scale. This results in the first collaborator holding gray-scale dominant data and the second collaborator holding color dominant data.
- Classification using Class Imbalanced CIFAR10 Data: Data is distributed non-uniformly across different collaborators to create non-IID conditions from the perspective of total training data per collaborator as well as the number of records per class.
- Next Character prediction model on Shakespeare dataset [18] leveraging non-IIDness in speaking-roles: Data corresponding to each speaking-role in the play is grouped to create unique collaborators, to simulate natural non-IID condition. For the trial, we selected only clients with a minimum of 10k data points and sampled a random subset of 66 clients.

4.2. Hyper Parameter Tuning

4.2.1. Fine Tuning Epoch and Learning Rate

Divide-and-Conquer allows the use of variable hyper-parameters for different parts of the network. As discussed earlier, we train feature-learning group more aggressively than fine-tuning group by control of parameters like E_p and η. Use of lower E_p for fine-tuning rounds result in slightly better accuracy compared to higher epochs. This is because the local models are skewed by over-fitting to non-IID data at the individual collaborators. By using lower values for epoch and learning rate for fine-tuning rounds, we achieve better accuracy while simultaneously reducing compute requirements needed for fine-tuning rounds. Data from Color Skewed distribution is presented in Figure 4. This observation is in alignment with the behaviour of FedAvg where a high number of E_p leads to lower training accuracy due to weight divergence.

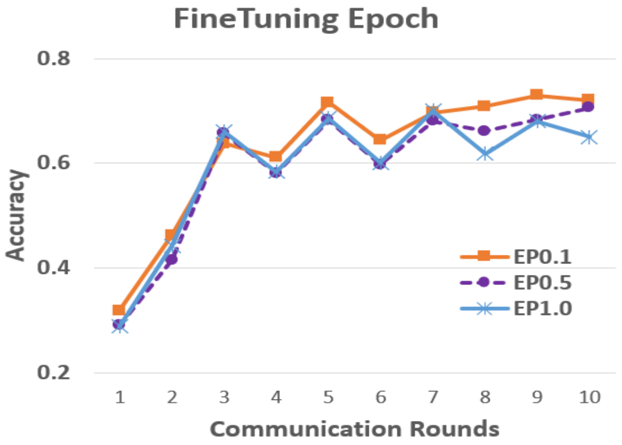


Figure 4. Effect of Training Epochs: Ep0.1 corresponds to fine-tuning epoch that is 10% of the value used for feature-learning

4.2.2. Fine Tuning Epoch and Learning Rate

Depending on the nature and magnitude of non-IIDness, the Class Agnostic and Class Specific layers in a given model will diverge across different collaborators. We explored, weight divergence in the learned model, to guide D&C methodology. The metric given below (EQ2) was explored in [5].

$$W_d = ||W_1 - W_2||/||W_1|| \qquad \text{(EQ2)}$$

$$W_d = CosineDist(W_1, W_2)/||W_1|| \qquad \text{(EQ3)}$$

We modified the divergence computation as shown in (EQ3), to capture direction aware divergence to guide our D&C methodology. Weight divergence from VGG9 model for Color Skewed non-IID simulation described in Section 4.1 is shown in Figure 5. A pre-pass training is initially performed for 5 rounds using entire model and layer-wise divergence strategy is devised using observations from the pre-pass round as reference. For notation, model at end of pre-pass comprising 5 rounds is denoted by M4. L1, L2 denotes different layers of VGG9 and M5, M6, etc., corresponding to models from future communication rounds post pre-pass. Observing pre-pass model M4, we find that the layer-wise divergence is low for the initial set of layers and starts to increase around Layer5. D&C can be applied around layer 5 to split the topology for creating feature-training and fine-tuning groups.

To validate the efficacy of W_d, Accuracy and convergence behaviour for VGG9 under different layer division schemes were checked using a brute force sweep across different splits. Accuracy for different division schemes is presented in Figure 6. As discussed in Section 2, Divide5 corresponds to division after layer5. From the figure, Divide5 offers the best accuracy and convergence speed under the given non-IID condition. All runs used 20 epochs for feature-learning and 4 epochs for fine-tuning. Likewise, learning rate for fine-tuning round was half that of feature-learning. Learning Rate Decay was also applied across the communication rounds starting from 0.001 and reducing by 10% for every round.

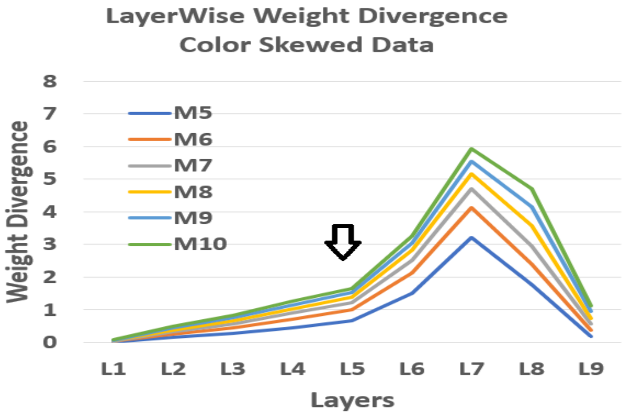


Figure 5. LayerWise Weight Divergence for Color Skewed distribution at different communication rounds. L1, L2 corresponds to layers and M5, M6 corresponds to model at end of successive communication rounds. Divergence is low for initial layers suggesting opportunities for Divide-and-Conquer

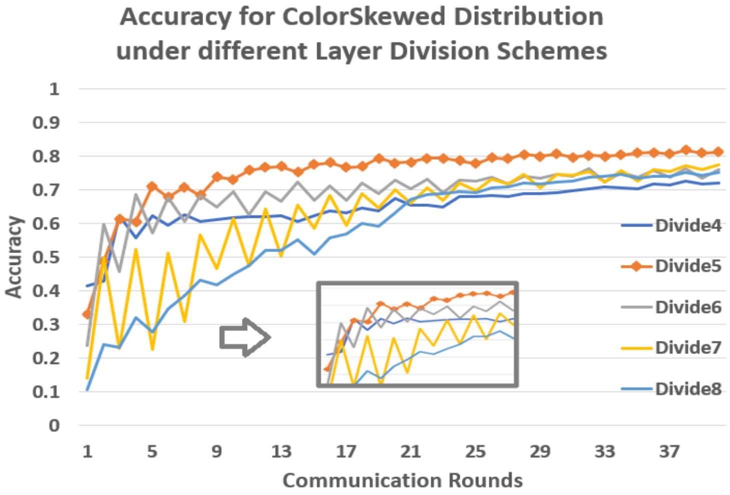


Figure 6. Training Accuracy under different Layer Division Schemes. Divide5 offers optimal results in-line with weight divergence

For certain division schemes (ex: Divide7), large spread is seen in accuracy between feature-learning and fine-tuning rounds suggesting that the layer assignment strategy for the two groups is sub-optimal. For Divide7, Divide6, etc., we find that accuracy is higher for fine-tuning rounds (Communication Round CR=2, 4, 6,…) and drops for feature-learning rounds (CR=1,3,5,…). The trend however reverses for Divide5 where the accuracy is higher for feature-learning rounds and marginally drops for fine-tuning group. The divergence between feature-learning and fine-tuning is also minimal in this split. When fewer layers are present in feature-learning group as in Divide4, we find that the rate of learning starts to fall, and accuracy spread between the two learning groups increases again. This suggests that Divide5 is an optimal split for this topology for this non-IID dataset thereby validating the usage of weight divergence metric to determine the point in a model where the layer split can be performed.

For the Class Imbalanced non-IID condition, the weight-divergence is high across all the layers of the topology Figure 7, suggesting that Divide-and-Conquer might not offer significant accuracy benefits. We chose to divide after layer8, in-line with traditional transfer learning strategies, where the last layer is used for fine-tuning for realizing bandwidth savings. The Next Character prediction model has 3 layers and we again use the last layer for fine-tuning.

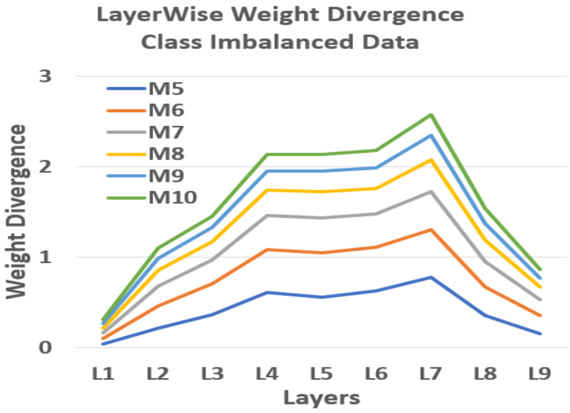


Figure 7. LayerWise Weight Divergence for class-imbalanced distribution at different communication rounds. L1, L2 corresponds to layers and M5, M6 corresponds to model at end of successive communication rounds. Divergence is high for all layers.

In the current work, layer division is determined using a pre-pass run and the division scheme is fixed for the entire duration of training. Future work will extend this to explore a dynamic scheme assignment where layers from a group can be reassigned to other group based on observed trend in feature-learning accuracy vs fine-tuning accuracy over few communication rounds.

5. Results

Results from the Divide-and-Conquer methodology under different non-IID scenario is presented in this section. For training we use 20 epochs for feature-learning and 4 epochs for fine-tuning. Learning rate was initialized to 0.001 and allowed to decay by 10% for every communication round. Learning rate for fine-tuning was 50% of learning-rate for feature-learning round.

Divide-and-Conquer uses half the network bandwidth for data transfers compared to FedAvg, as the full model is transferred for every two communication rounds. For FedMA, results from equivalent matched averaged round is presented based on equivalency established in [13]. Though FedMA uses much lower communication bandwidth, compute overhead for layer matching increases with model depth as well as width, making it less desirable for practical deployments. We show that our methodology yields similar accuracy levels as more complex algorithms like FedMA in acceptable rounds of communication.

Note: In the tables providing the comparison across different approaches, Divide-and-Conquer is captured under D&C.

5.1. Image Classification: Color Skewed Distribution

Training accuracy and convergence profile for different aggregation algorithms using Color Skewed 95-5% CIFAR10 data are shown in Figure 8. For this category of non-IIDness, the model reaches high accuracy with much smaller communication rounds compared to FedAvg. Divide5 was used for this analysis as described in the earlier section along with the same values for E_p and η. Results for additional levels of Color Skew is presented in Table 1. We chose 18 rounds of communication for the comparison to align with FedMA.

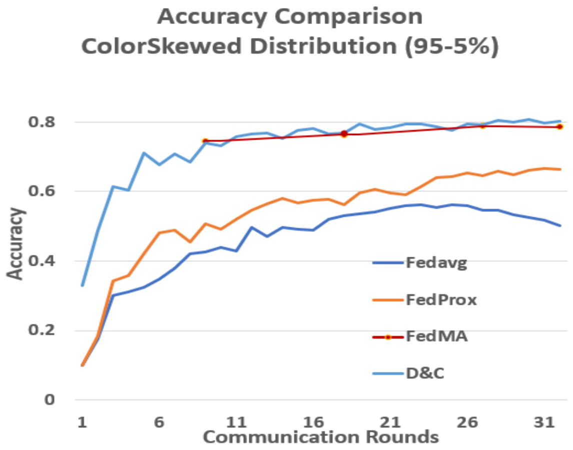


Figure 8. Accuracy Comparison for Color Skewed Distribution with 95-5% skew. Accuracy and Convergence rate for Divide-and-Conquer (using FedAvg) is higher than FedAvg

Table 1. Accuracy for Color Skewed Distribution for 18 communication rounds under different levels of skew for 2 collaborators. D&C (using FedAvg) delivers high classification accuracy under this non-iidness

#Col	SKEW	FedAvg	FedProx	FedMA	D&C
2	95-5%	53.1%	56.2%	81.0%	80.1%
2	75-25%	52.8%	74.6%	78.8%	79.2%
2	50-50%	49.1%	67.2%	79.9%	81.8%

5.2. Image Classification: Class Imbalance Distribution

For Class Imbalanced data, the observed weight divergence from the pre-pass run was high for most layers. This indicates that D&C does not offer opportunities for Layer Splitting for accuracy improvements, though it might still offer opportunity for bandwidth saving. As an experiment, we divided the topology at layer8, similar to traditional transfer learning. D&C yields slightly lower accuracy compared to FedAvg and FedMA for half the bandwidth requirement, as documented in Table 2. However, if bandwidth saving is not sacrificed and D&C is run for additional rounds to get a similar amount of model transfer as FedAvg, the performance of Divide-and-Conquer is marginally better. This is captured in the table under the column D&C. Though FedMA achieves its accuracy levels using much lower communication bandwidth, compute overhead for layer matching increases with model depth as well as width, as discussed earlier, making it less desirable for practical deployments.

Given the results, in cases where weight divergence suggests no clear split layer, it is recommended adopt D&C solely for bandwidth saving. As collaborator count increases in our experimental setup, training data per collaborator decreases, (as same data is divided across the collaborators). This could also lead to increased divergence, when feature-learning is done aggressively on sparse data. In a truly federated set up with a large training corpus across collaborators, we expect our methodology to offer better accuracy improvements.

Table 2. Accuracy for Class-Imbalanced Distribution for 18 communication rounds using 5 & 10 collaborators. Accuracy from D&C (using FedAvg) is in-line with other algorithms at half the bandwidth requirement.

#Col	**FedAvg**	**FedProx**	**FedMA**	**D&C**	**D&C'**
5	88.5%	87.5%	87.5%	87.1%	89.3%
10	83.5%	80.0%	82.5%	76.8%	82.2%

An extreme case of Class Imbalance based heterogeneity is when each collaborator exclusively holds data from one unique class. All the tested algorithms performed poorly (accuracy less than 15%) under this scenario, suggesting a need for more research in this area.

5.3. Next-Character Prediction: Speaker-Role based non-IID Distribution

Results from application of D&C to a character prediction model is shown in Table 3. At end of 9 communication rounds, the accuracy from D&C is comparable to other algorithms while only requiring half the amount of data transfer as FedAvg. 9 rounds of communication were chosen to align with FedMA.

Table 3. Accuracy for next-character prediction using lstm-model for 9 communication rounds using 66 collaborators. Accuracy from D&C (using FedAvg) is in-line with other algorithms at half the bandwidth requirement.

#Col	**FedAvg**	**FedProx**	**FedMA**	**D&C**
66	50.8%	44.6%	47.4%	49.6%

6. CONCLUSIONS

In this work, we presented a weight-divergence based, Divide-and-Conquer algorithm which builds on popular FedAvg algorithm to achieve state-of-the-art accuracy under non-IIDness. By training network in parts, our novel methodology is shown to a) Achieve faster convergence when low-level features are well-represented b) Reduce communication by half, because of training and weight exchange in parts, and c) Require less compute compared to state-of-the-art techniques like FedMA, which has performance overheads from weight matching. A static topology splitting strategy is adapted in this work, where the topology is divided at the beginning of training using a pre-pass run. Future work can explore a dynamic Divide-and-Conquer strategy where layers are moved between feature-learning and fine-tuning groups based on accuracy observations during training. Future work can also explore the application of Divide-and-Conquer methodology to learning paradigms like Few Shot Learning to identify Class Agnostic layers for the backbone network.

REFERENCES

[1] McMahan, H.B., Moore, E., Ramage, D., Hampson, S., & Arcas, B.A. (2017). Communication-Efficient Learning of Deep Networks from Decentralized Data. AISTATS.

[2] Li, T., Sahu, A.K., Talwalkar, A.S., & Smith, V. (2020). Federated Learning: Challenges, Methods, and Future Directions. IEEE Signal Processing Magazine, 37, 50-60.

[3] Bonawitz, K., Eichner, H., Grieskamp, W., Huba, D., Ingerman, A., Ivanov, V., Kiddon, C., Konecný, J., Mazzocchi, S., McMahan, H.B., Overveldt, T.V., Petrou, D., Ramage, D., & Roselander, J. (2019). Towards Federated Learning at Scale: System Design. ArXiv, abs/1902.01046.

[4] Xu, J., & Wang, F. (2021). Federated Learning for Healthcare Informatics. Journal of Healthcare Informatics Research, 1 - 19.

[5] Zhao, Y., Li, M., Lai, L., Suda, N., Civin, D., & Chandra, V. (2018). Federated Learning with Non-IID Data. ArXiv, abs/1806.00582.

[6] Lin, T., Kong, L., Stich, S.U., & Jaggi, M. (2020). Ensemble Distillation for Robust Model Fusion in Federated Learning. ArXiv, abs/2006.07242.

[7] Fallah, A., Mokhtari, A., & Ozdaglar, A. (2020). Personalized Federated Learning with Theoretical Guarantees: A Model-Agnostic Meta-Learning Approach. NeurIPS.

[8] Ghosh, A., Chung, J., Yin, D., & Ramchandran, K. (2020). An Efficient Framework for Clustered Federated Learning. ArXiv, abs/2006.04088.

[9] Hanzely, F., & Richtárik, P. (2020). Federated Learning of a Mixture of Global and Local Models. ArXiv, abs/2002.05516.

[10] Dinh, C.T., Tran, N.H., & Nguyen, T.D. (2020). Personalized Federated Learning with Moreau Envelopes. ArXiv, abs/2006.08848.

[11] Hanzely, F., Hanzely, S., Horvath, S., & Richtárik, P. (2020). Lower Bounds and Optimal Algorithms for Personalized Federated Learning. ArXiv, abs/2010.02372.

[12] Sahu, A.K., Li, T., Sanjabi, M., Zaheer, M., Talwalkar, A.S., & Smith, V. (2020). Federated Optimization in Heterogeneous Networks. arXiv: Learning.

[13] Wang, H., Yurochkin, M., Sun, Y., Papailiopoulos, D., & Khazaeni, Y. (2020). Federated Learning with Matched Averaging. ArXiv, abs/2002.06440.

[14] Chen, W., Bhardwaj, K., & Marculescu, R. (2020). FedMAX: Mitigating Activation Divergence for Accurate and Communication-Efficient Federated Learning. ECML/PKDD

[15] Shoham, N., Avidor, T., Keren, A., Israel, N., Benditkis, D., Mor-Yosef, L., & Zeitak, I. (2019). Overcoming Forgetting in Federated Learning on Non-IID Data. ArXiv, abs/1910.07796.

[16] Simonyan, K., & Zisserman, A. (2015). Very Deep Convolutional Networks for Large-Scale Image Recognition. CoRR, abs/1409.1556.

[17] Yan, X., Chen, Z., Xu, A., Wang, X., Liang, X., & Lin, L. (2019). Meta R-CNN: Towards General Solver for Instance-Level Low-Shot Learning. 2019 IEEE/CVF International Conference on Computer Vision (ICCV), 9576-9585.

[18] Caldas, S., Wu, P., Li, T., Konecný, J., McMahan, H.B., Smith, V., & Talwalkar, A.S. (2018). LEAF: A Benchmark for Federated Settings. ArXiv, abs/1812.01097.

[19] Krizhevsky, A. (2009). Learning Multiple Layers of Features from Tiny Images.

AUTHORS

Raghavendra Bhat (Raghu) is a Principal Engineer at Intel Technology India Pvt. Ltd. He has a BTech in Computer Science from NIT Warangal and PGDBA (Finance) from Symbiosis Pune. He has over 22 years of industry experience spread across domains like Network Management, Embedded platform development for Mobile, IoT and Biometrics solutions. In his current role at Intel, he leads exploration in Healthcare space as part of Vertical Solutions and Services Group. In his role, the primary focus is on design, development, and ecosystem adoption of optimized AI solutions on Intel AI portfolio for Speech, Language, Image and Video analytics use cases. Currently, a significant focus is on AI Training at the Edge where training methodologies like Federated Learning, Incremental Learning etc. He is on the ISO and BIS standardization panels of Blockchain and AI. His interests extend to SW architecture methodologies, Blockchain and Aadhaar ecosystem where he has contributed towards Iris biometric device development and specifying device security requirements. As part of IEEE standards organization, he is leading the pre-standardization study for low resourced Indian languages. He has several patents and publications to his credit.

Pravin Chandran works as Deep Learning R&D Engineer at Intel Technology India Pvt, Ltd. He holds a M.S in EE from Clemson University, USA and B.E from University of Madras, India. He has 13 years of professional experience in wide range of areas including ML/DL, Software Development, Statistical Design Analysis, Yield Estimation, VLSI EDA Methodology and SoC Design.

Avinash Chakravarthi (Avi) works as Deep learning scientist at Intel, he joined Intel as graduate intern after completing his bachelors from VIT University in Electronics, 2016. His interests and area of work include Federated learning, Bio inspired computing and Software development.

Srikanth Chandar graduated from PES University, and currently works an AI Engineer at Intel Corporation. Anything ML excites him, and he likes taking up a challenge that could shape a new paradigm in the same space, be it application-based or research. Voice cloning using GANs, and Communication optimization in FL are his recent research pursuits. Another thing that interests him other than talking about himself in the third person is Animal Welfare. He feels very strongly about animal abuse and runs an NGO (Dystopia-Animal Welfare) to fight the same through campaigns, volunteering activities, and also technical projects.

A FRAGMENTATION REGION-BASED SKYLINE COMPUTATION FRAMEWORK FOR A GROUP OF USERS

Ghoncheh Babanejad Dehaki[1], Hamidah Ibrahim[1],
Nur Izura Udzir[1], Fatimah Sidi[1] and Ali Amer Alwan[2]

[1]Faculty of Computer Science and Information Technology,
Universiti Putra Malaysia, Selangor, Malaysia
[2]Kulliyah of Information and Communication Technology,
International Islamic University Malaysia, Kuala Lumpur, Malaysia

ABSTRACT

Skyline processing, an established preference evaluation technique, aims at discovering the best, most preferred objects, i.e. those that are not dominated by other objects, in satisfying the user's preferences. In today's society, due to the advancement of technology, ad-hoc meetings or impromptu gathering are becoming more and more common. Deciding on a suitable meeting point (object)for a group of people (users) to meet is not a straightforward task especially when these users are located at different places with distinct preferences. A place which is close by to the users might not provide the facilities/services that meet all the users' preferences; while a place having the facilities/services that meet most of the users' preferences might be too distant from these users. Although the skyline operator can be utilised to filter the dominated objects among the objects that fall in the region of interest of these users, computing the skylines for various groups of users in similar region would mean rescanning the objects of the region and repeating the process of pair wise comparisons among the objects which are undoubtedly unwise. On this account, this study presents a region-based skyline computation framework which attempts to resolve the above issues by fragmenting the search region of a group of users and utilising the past computed skyline results of the fragments. The skylines, which are the objects recommended to be visited by a group of users, are derived by analysing both the locations of the users, i.e. spatial attributes, as well as the spatial and non-spatial attributes of the objects. Several experiments have been conducted and the results show that our proposed framework outperforms the previous works with respect to CPU time.

KEYWORDS

Skyline Queries, Preference Queries, Group Preferences, Fragmentation Strategy.

1. INTRODUCTION

The skyline operator introduced by [6] which is used to filter a set of interesting objects from a potentially large multi-dimensional set of objects by keeping only those objects that are not worse than any other; has been greatly explored in several studies in an attempt to accurately and efficiently solve problems of real-world applications that are related to decision support and decision making. It attempts to derive the best, most preferred set of objects known as skylines according to a set of established criteria. The process of computing skylines becomes more challenging when conflicting criteria are involved while the number of criteria to be considered is huge. A classic example is selecting a hotel for a holiday whereby hotels that are close to the beach are

David C. Wyld et al. (Eds): AIMLA, DBDM, CCNET - 2021
pp. 35-50, 2021. CS & IT - CSCP 2021 DOI: 10.5121/csit.2021.111303

known to be expensive. While other criteria like facilities, rating, service, etc are equally important, distance and price are examples of conflicting criteria.Although there are a considerable amount of works in processing skylines, however most of them are limited to the aim of satisfying the preferences of a single user [5, 6, 8, 11, 15].Today's advancement of technology shows that ad-hoc meetings or unplanned gathering are becoming more and more common. Determining the best, most preferred objects in which the preferences of several users are to be considered is more complex as opposed to a single user. The following scenario simulates a sample of situation considered in this paper.

Assume a group of users who are located at different locations; would like to gather and hence have to decide on a place to meet. Several criteria need to be considered such as the location of the place, i.e. how far it is from the location of each user (spatial attribute), the opening hour, food, ticket price, rating, facilities provided, etc (non-spatial attributes). Intuitively, deciding on the best meeting point for these users is not a straightforward task as many criteria need to be considered. A place which is near to the users might not be a place that meets all the users' preferences. While a place which provides facilities/services that meet most of the users' preferences might be located far away from these users. Thus, it is essential to have a method that could find an object(s) within a predetermined region that dominates other objects with respect to both the spatial (location) and non-spatial attributes of the objects that best suits the preferences of a group of users. Although this has been well tackled by [9, 19, 26, 27], there is no attempt made to resolve the following issue. Obviously, similar regions may have to be explored again in computing skylines for different groups of users (or even the same group of users at a different day/time). Attempting to rescanning the objects of previously visited regions and recomputing the skylines of the regions (i.e. repeating the process of pairwise comparisons among objects) are undoubtedly unwise and costly.

Motivated by the above example, we propose a region-based skyline computation framework, *RSGU*, an enhancement of our previous framework, *SGMU* [9], with the main aim at avoiding the process of rescanning the objects of a previously visited region by utilising a fragmentation strategy as well as re computing the skylines for a group of users by utilising the past computed skyline results of the fragments. The skylines, which are the objects recommended to be visited by the group of users, are derived by analysing both the locations of these users, i.e. spatial attributes, as well as the spatial and non-spatial attributes of the objects.

This paper is organised as follows: Section 2 presents the related works which are organised into two parts, namely: skyline algorithms for a single user and skyline algorithms for a group of users. This is followed by Section 3 which introduces the notations and deliberates the terms that are used throughout this paper. It also presents the problem tackled by this paper. Section 4 presents our proposed framework and the steps to be performed in order to achieve the main aim of the work. Section 5 discusses the initial results achieved by our proposed framework while the last section, Section 6, gives the summary of the paper.

2. RELATED WORK

The skyline operator proposed by [6] is a well-studied technique for filtering the best, most preferred objects from a multi-dimensional set of objects. Since then many variants of skyline algorithms have been proposed, each tackling a slightly different issue mainly due to the nature of data being handled. We categorised these skyline algorithms into two main categories, namely: skyline algorithms for a single user and skyline algorithms for a group of users.

Skyline algorithms for a single user – Generally, these skyline algorithms filter the best, most preferred objects from a potentially large multi-dimensional set of objects with the assumption that all users have the same property with the same objective function (preferences). Among the earlier and most cited skyline algorithms in the literature are *Block Nested Loop* (*BNL*) [6], *Divide-and-Conquer* (*D&C*) [6], *Linear Elimination Sort for Skyline* (*LESS*) [11], *Branch and Bound Skyline* (*BBS*) [24], *SkyCube* [6], and *Sort and Limit Skyline algorithm* (SaLSa) [5]. These algorithms attempt to resolve the optimisation problem which is proven through the reduction of the processing time. Later due to advancement in technology that produces gigantic amount of various forms of data, several skyline algorithms have been proposed. These algorithms attempt not only to resolve the optimisation problem but also issues related to the uncertainty of data which is defined as the degree to which data are inaccurate, imprecise, untrusted, unknown or incomplete. These include among others *ISkyline* [14], *sorting-based bucket skyline* [18], *Incoskyline* [2], *Jincoskyline* [1], and *OIS* [12] that handle the issues of incompleteness of data; *probabilistic skyline model* [25], *τ-Skyline* [29], *SkyQUD* [20, 21, 22, 23] and *SkyQuiD* [17] focus on the challenges in computing skyline queries for uncertain database; the works by [4] and [10] attempt to solve the issues related to uncertain data in a data stream; while the work by [3] focuses on dynamic database. Nonetheless, these algorithms are specifically designed to cater only a single user query.

Skyline algorithms for a group of users – These skyline algorithms keep those objects that are not worse than any other from a potentially large multi-dimensional set of objects in which the preferences of multiple users are taken into account. As we assume that the objects are static, hence we further elaborate only those works that are similar to our intention. To the best of our knowledge the only works that contribute to skyline queries for a group of users are the works done by [19], [26], and[27]. In processing spatial skyline query for a group of users, [26] have proposed two algorithms, namely: B^2S^2 and VS^2. The B^2S^2 algorithm utilises the R-tree while the VS^2 algorithm utilises the Voronoi diagram. Both algorithms are performed on static user points. Later, [27] proposed VCS^2, an enhancement to B^2S^2 and VS^2, which aims at processing skyline query by taking into consideration the movements of the users. However, VCS^2 only calculates the last location of the users and does not consider the changes of locations to prevent recalculation of the skylines. In [19], the authors proposed an algorithm, *VR* (Voronoi and R-tree), to find spatial skylines for a group of user points. In their work the user points and objects are considered static. The two data structures, R-tree and Voronoi, used in [27] are combined in this work. Both the spatial and non-spatial attributes of the objects are analysed to find the skylines. Meanwhile, our previous solution, *SGMU* [9], is designed with the main aim to continuously derive skylines for a group of mobile users. In *SGMU*, while the users decide on a place to visit, the skylines are continuously updated since a place that was initially in the top list based on the users' locations at time $t_a$might no longer be the place of interest at time t_b where $t_a < t_b$ since the users' locations at time t_a might be different at time t_b.

Although [9, 19, 26, 27] considered the spatial attributes of the group of users in determining the skylines, but there is no attempt made to avoid rescanning of objects of previously visited regions and simultaneously avoid repeating the process of pairwise comparisons among the objects.

3. PRELIMINARIES

This section elaborates the concepts that are related to the work presented in this paper. It also defines the terms and introduces the notations used throughout the paper. Towards the end of this section, we formulate the problem tackled in this paper. To clarify the concepts and steps proposed in this work, the following sample of data will be used. Table 1(a) and Table 1(b) present the spatial attribute (*Location*) of the users of group a, G_a, and group b, G_b, respectively. Here, we assume that the request submitted by G_a is at time t_a, while the request submitted by G_b is at time t_b where $t_a < t_b$. Table 2 presents the spatial (*Location*) and non-spatial (*Rate*, *Fee*) attributes of the objects. For the non-spatial attributes, we assume higher rate and lower fee are preferable.

Table 1. The spatial attribute of the users.

ID	Location
u_1	(8, 8)
u_2	(14, 16)
u_3	(2, 5)

(a) Group a, G_a

ID	Location
u_1	(5, 8)
u_2	(10, 10)
u_3	(18, 10)

(b) Group b, G_b

Table 2. The spatial and non-spatial attributes of the objects.

Restaurant	Location	Rate	Price	Restaurant	Location	Rate	Price
o_1	(2, 3)	3	70	o_{24}	(4, 13.3)	3	75
o_2	(3, 4)	4	65	o_{25}	(7, 13)	1	90
o_3	(3, 1)	5	80	o_{26}	(16, 15)	2	86
o_4	(7, 1.7)	2	75	o_{27}	(20, 14)	5	80
o_5	(6, 5)	3	65	o_{28}	(23, 20)	3	60
o_6	(7, 7)	5	70	o_{29}	(21, 21)	5	62
o_7	(9, 8)	1	80	o_{30}	(17, 23)	4	95
o_8	(8, 9.7)	2	85	o_{31}	(14, 20)	2	65
o_9	(7, 11)	4	73	o_{32}	(13, 18)	2	55
o_{10}	(10, 5)	3	50	o_{33}	(10, 19)	3	70
o_{11}	(10.7, 6)	1	65	o_{34}	(1, 16)	4	62
o_{12}	(15, 2)	2	80	o_{35}	(3, 22)	4	81
o_{13}	(17, 1)	5	105	o_{36}	(7, 20)	3	90
o_{14}	(22, 4.7)	4	90	o_{37}	(24, 15)	2	66
o_{15}	(17, 5.7)	3	85	o_{38}	(-3, -1)	1	57
o_{16}	(20, 7)	4	90	o_{39}	(-1, 7)	1	61
o_{17}	(23, 9)	1	55	o_{40}	(10.3,13)	4	71
o_{18}	(16, 8)	2	54	o_{41}	(-4, 4)	3	98
o_{19}	(14, 10)	4	80	o_{42}	(8, -2)	2	58
o_{20}	(11, 9.7)	5	56	o_{43}	(8, 18)	2	85
o_{21}	(4, 10)	3	67	o_{44}	(-2, 10)	4	70
o_{22}	(2, 12)	5	100	o_{45}	(3, -1)	5	80
o_{23}	(3, 13)	4	74				

Given a dataset $D = < A, U, O >$, where $U = \{u_1, u_2, \dots, u_n\}$ is a list of $n$users, $O = \{o_1, o_2, \dots, o_m\}$ is a list of $m$objects, and $A = < A_S, A_N >$ are the criteria (dimensions) to be considered in the skyline computation where A_S is a spatial attribute while $A_N = \{d_1, d_2, \dots, d_l\}$is a set of non-spatial attributes. Based on Table 2, $A_S = Location$ and $A_N = \{Rate, Price\}$.

The following definitions defined the properties of a user and an object as used in ourwork.

Definition 1 *Property of a User*: Each user, $u_i \in U$, is associated with a spatial attribute which represents the location of the user at time, t. This is denoted by $u_i(x_i, y_i)$ where x_i and y_i represent the latitude and longitude coordinates, respectively. For instance, $u_1(8, 8)$ of Table 1(a) denotes the location of user u_1 where $x_i = 8$ and $y_i = 8$.

Definition 2 *Properties of an Object*: Each object $o_j \in O$ has two main elements denoted by $o_j = (s_j, ns_j)$ where s_j is the value of spatial attribute (location), A_S, and

$ns_j = \{o_j.d_1, o_j.d_2, \ldots, o_j.d_l\}$ is a set of values of non-spatial attributes, A_N, associated to o_j. The location of an object $o_j \in O$ is denoted by $o_j(x_j, y_j)$. As we assume that each object $o_j \in O$is static, thus the location of the object is fixed regardless the changes in time. Hence, $o_j = (s_j, ns_j)$ can be written as $o_j = ((x_j, y_j), \{o_j.d_1, o_j.d_2, \ldots, o_j.d_l\})$. For instance, the object $o_1$of Table 2 can be written as $o_1 = ((2,3), \{3,70\})$.

The following definitions defined the notion of dominance in ourwork.

Definition 3 *Dominance*: Given two objects $o_i = (s_i, ns_i) \in O$ and $o_j = (s_j, ns_j) \in O$ where $i \neq j$, o_i is said to dominate o_j (denoted by $o_i \prec o_j$) if and only if both of the following conditions hold:

(1) o_i non-spatially dominates $o_j(o_i \prec_{ns} o_j)$ and
(2) o_i spatially dominates $o_j(o_i \prec_s o_j)$.

Without loss of generality, this definition is applicable for a given bounded space, S, i.e. O is a set of objects in the space S. Similar note applies for *Definition* 4 and *Definition* 5.

Definition 4*Non-spatial Dominance*: Given two objects $o_i = (s_i, ns_i) \in O$and $o_j = (s_j, ns_j) \in O$ where $i \neq j$, o_i is said to non-spatially dominate o_j(denoted by $o_i \prec_{ns} o_j$) if and only if $o_i$is no worse than (in this definition, greater value is preferable) o_j in all the non-spatial attributes, A_N. This is formally written as follows: $o_i \prec_{ns} o_j$ if and only if $\forall d_k \in A_N, o_i.d_k \geq o_j.d_k \wedge \exists d_l \in A_N, o_i.d_l > o_j.d_l$. For instance, given $o_6 = ((7,7), \{5,70\})$ and$o_{12} = ((15,2), \{2,80\})$, $o_6 \prec_{ns} o_{12}$ since o_6 is better than o_{12} in both the dimensions *Rate* and *Price*.

Definition 5 *Spatial Dominance*: Given two objects $o_i = (s_i, ns_i) \in O$ and $o_j = (s_j, ns_j) \in O$ where $i \neq j$, o_i is said to spatially dominate o_j (denoted by $o_i \prec_s o_j$) if and only if for every user $u_k \in U$, the distance between o_i and $u_k$is no worse than the distance between $o_j$and u_k. This is formally written as follows: $o_i \prec_s o_j$ if and only if $\forall u_k \in U, dist(o_i, u_k) \leq dist(o_j, u_k) \wedge \exists u_l \in U, dist(o_i, u_l) < dist(o_j, u_l)$. For instance, the distances between $o_1 = ((2,3), \{3,70\})$ and u_1, u_2, and u_3 of group G_a are 7.81, 17.69, and 2, respectively; while the distances between $o_2 = ((3,4), \{4,65\})$ and u_1, u_2, and u_3 of group G_a are 6.4, 16.27, and 1.41, respectively. Thus, $o_2 \prec_s o_1$.

Definition 6 is an extension of *Definition* 3 in which the dominance testing is performed over a predefined space. The list of objects considered in *Definition* 3 is the $m$objects as defined in the system while in *Definition* 6 the list of objects is confined to those objects that fall within a certain space.

Definition 6 *Dominance in a Space*: Given a bounded space, S (region, MBR, fragment, area, polygon, etc), and two objects $o_i = (s_i, ns_i) \in O$and $o_j = (s_j, ns_j) \in O$ where $i \neq j$ in S, o_i is said to dominate o_j (denoted by $o_i \prec o_j$) in S if and only if

(1) o_i non-spatially dominates $o_j(o_i \prec_{ns} o_j)$ in S and
(2) o_i spatially dominates $o_j(o_i \prec_s o_j)$ in S.

Definition 7 *Skylines of a Space*: An object $o_i \in O$ in a space S is a skyline of S if there are no other objects $o_j \in O$ in the space S where $i \neq j$ that dominates o_i. In this paper, $SkyG_p$ is used to denote the skyline set for the group $G_p$of a given space S.

Based on the above definitions, we now formulate the problem that is tackled by this paper.

Problem Formulation

Given a group of users, $G_p = \{u_1, u_2, \ldots, u_p\}$, where $G_p \subset U$, and the list of candidate skylines of G_p, CS_{G_p}, in region S_{G_p}. Find the skylines of a group of users $G_q = \{u_1, u_2, \ldots, u_q\}$ in region S_{G_q} where $G_q \subset U$, $G_q \neq G_p$, and$S_{G_p} \cap S_{G_q} \neq \{\}$ by utilising CS_{G_p} that has been derived for G_p. This is depicted in Figure 1 where the area covered to compute the skylines for $G_q$that falls in the region S_{G_q} can be reduced to the area defined by $S_{G_q} - S_{G_p}$, while the skyline computation that has been performed earlier over the area $S_{G_p} \cap S_{G_q}$ for $G_p$can be avoided by simply utilising the obtained results derived earlier for G_p, i.e. CS_{G_p}.

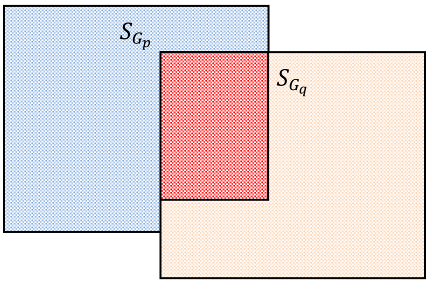
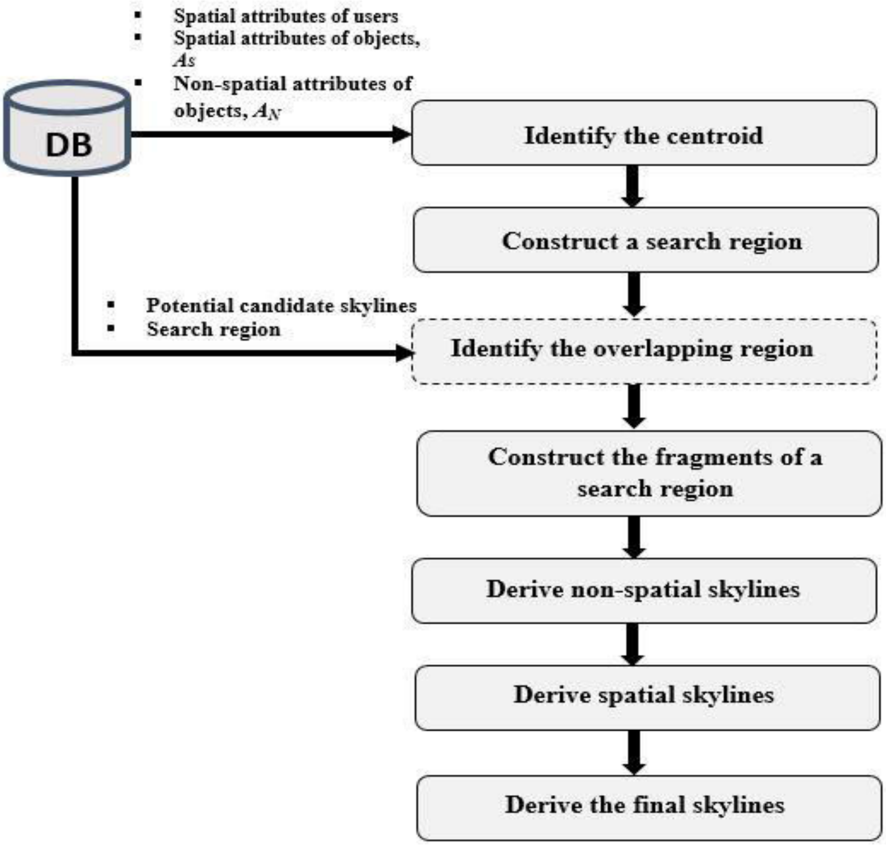


Figure 1. The reduction area in deriving skylines for a group of users

4. THE PROPOSED FRAMEWORK

This section elaborates the extended framework, *RSGU* [9], which we have proposed in order to solve the problem defined in Section 3. The framework is presented in Figure 2. It consists of seven main steps that are: (1) Identify the centroid, (2) Construct a search region, (3) Identify the overlapping region, (4) Construct the fragments of a search region, (5) Derive non-spatial skylines, (6) Derive spatial skylines, and (7) Derive the final skylines. Step (3) is conducted only when past computed skyline results of the fragments are available. Step (3) and Step (4) are the new steps incorporated into *RSGU* [9]. Each of these steps is elaborated in the following subsections.



Spatial attributes of users
Spatial attributes of objects, A_S
Non-spatial attributes of objects, A_N
DB
Identify the centroid
Construct a search region
Potential candidate skylines
Search region
Identify the overlapping region
Construct the fragments of a search region
Derive non-spatial skylines
Derive spatial skylines
Derive the final skylines

Figure 2. The proposed framework

4.1. Identify the Centroid

When a group of users, $G_p = \{u_1, u_2, \dots, u_p\}$, decided to meet, there must be a point to guide the direction of their movements. In our work, we assume that the group of users will move towards a point that has the tendency to be a center based on the users' locations. This point is called centroid and is denoted by$C_{G_p}(x_{G_p}, y_{G_p})$. The centroid of a given group of users, C_{G_p}, is determined using the following formula [13]:

$$C_{G_p}\ (x_{G_p} = \frac{\sum_{i=1}^{n} xi}{n}, y_{G_p} = \frac{\sum_{i=1}^{n} yi}{n}) \qquad (1)$$

where x_i is the x coordinate of user u_i location, $y_i$is the y coordinate of user u_i location, x_{G_p} is the average of the x coordinates of all users in the group G_p, and $y_{G_p}$is the average of the y coordinates of all users in the group G_p. Based on the example given in Table 1(a), the centroid of G_a is $C_{G_a}(8, 9.6)$.

4.2. Construct a Search Region

The aim of constructing a search region is to limit the searching space to those spaces in which potential candidate skylines (objects) are derived. Since we are interested with a group of users, thus the searching space should include the regions of interest of all users in the group. This is achieved by: (1) identifying the search region for each user, S_{u_i} and (ii) identifying the search region given a group of users, S_{G_p}.

4.2.1. Identify the search region for each user, S_{u_i}

Since the centroid of a given group of users, say C_{G_p}, which is identified in the previous step does not necessarily contain an object, therefore the nearest object, o_n, to the centroid C_{G_p}will have to be determined. The nearest object is an object with the shortest Euclidean distance from the centroid, i.e. $\{o_n | o_n \in O \wedge \forall o_i \in O - \{o_n\}: Ed(C_{G_p}, o_n) < Ed(C_{G_p}, o_i)\}$ where $Ed$is the Euclidean distance function. Based on the example given in Table 1(a), the nearest object to the centroid of G_a, i.e. $C_{G_a}(8, 9.6)$, is $o_8(8, 9.7)$. The search region for a user, u_i, denoted as S_{u_i}, is the area bounded by a rectangle also known as the minimum bounding rectangle, MBR_{u_i}.We use the notation S_{u_i} to denote the search region of u_i while MBR_{u_i} is used in forming the S_{u_i}. The distance between a user, u_i, and the nearest object, o_n, denoted by $R_{u_i o_n}$, is calculated by the following equation:

$$R_{u_i o_n} = \sqrt{(x_{o_n} - x_i)^2 + (y_{o_n} - y_i)^2} (2)$$

where x_i is the x coordinate of user u_i location, y_i is the y coordinate of user u_i location, x_{o_n} is the x coordinate of object o_n location, and y_{o_n} is the $y$coordinate of object o_n location. AMBR is formed based on four vertices as explained in the following: the vertex at the bottom left of the MBR is denoted by $bl = (x_{bl}, y_{bl})$; the vertex at the bottom right of the MBR is denoted by $br = (x_{br}, y_{br})$; the vertex at the top left of the MBR is denoted by $tl = (x_{tl}, y_{tl})$; and the vertex at the top right of the MBR is denoted by$tr = (x_{tr}, y_{tr})$. Figure 3 depicts these notations. These vertices are calculated as follows:

$$\begin{aligned} bl &= (x_i - R_{u_i o_n}, y_i - R_{u_i o_n}) \\ br &= (x_i + R_{u_i o_n}, y_i - R_{u_i o_n}) \\ tl &= (x_i - R_{u_i o_n}, y_i + R_{u_i o_n}) \\ tr &= \left(x_i + R_{u_i o_n}, y_i + R_{u_i o_n}\right) \end{aligned}$$

4.2.2. Identify the search region given a group of users, S_{G_p}

This step is simply achieved by performing union on the search region of each user in the group, i.e. $S_{G_p} = \bigcup_{i=1}^{p} S_{u_i}$. An example of a search region $S_{G_a} = \bigcup_{i=1}^{3} S_{u_i}$ can be seen in Figure 4.

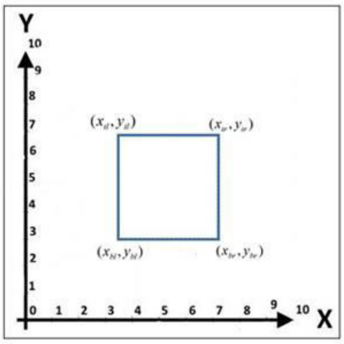


Figure 3. Minimum Bounding Rectangle (MBR)

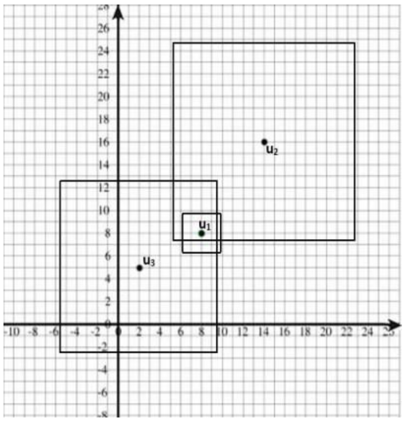


Figure4. The search region for a group of users

4.3. Construct the Fragments of a Search Region

This step partitions the search region of a group of users, S_{G_p}, into m fragments (subspaces). Here, the vertices of the MBRassociated to each $S_{u_i}$are analysed and sorted according to the x- and y-axes.The search region (space) is vertically fragmented based on the x-coordinates, while it is horizontally fragmented based on the y-coordinates. The$MRBs$ formed within the $S_{G_p}$are the fragments of the region.

Objects that fall within each fragment are then identified. Given an object, $o_j(x_j, y_j)$, and a fragment, F_i, with $bl(x_l, y_b)$, $br(x_r, y_b)$, $tl(x_l, y_t)$, and $tr(x_r, y_t)$, the following cases are identified:

(a) If $x_l < x_j < x_r$ and $y_b < y_j < y_t$, then the object $o_j(x_j, y_j)$ is said to fall within the boundary of fragment, F_i.

(b) If $x_j = x_l$or $x_j = x_r$or $y_j = y_b$ or $y_j = y_t$, then the object $o_j(x_j, y_j)$ is said to intersect with the boundary of fragment, F_i.

(c) Objects that do not meet the above two cases are objects that are outside the boundary of fragment, F_i.

Further, utilising the non-spatial dominance testing given in *Definition* 8, an extension to the *Definition 4*,over the objects that satisfy the cases (a) or (b) above, denoted by O_{F_i}, the non-spatial candidate skylines of a fragment are determined, $CS_{ns_{F_i}}$, as defined by *Definition* 9.

Definition 8 *Non-spatial Dominance of the Fragment* F_i: Given two objects $o_i = (s_i, ns_i) \in O_{F_i}$and $o_j = (s_j, ns_j) \in O_{F_i}$where $i \neq j$, $o_i$is said to non-spatially dominate o_j(denoted by $o_i \prec_{ns} o_j$) if and only if $o_i$is no worse than (in this definition, greater value is preferable) $o_j$in all the non-spatial attributes, A_N. This is formally written as follows: $o_i \prec_{ns} o_j$ if and only if $\forall d_k \in A_N, o_i.d_k \geq o_j.d_k \land \exists d_l \in A_N, o_i.d_l > o_j.d_l$.

Definition 9 *Candidate Skylines of the Fragment* F_i: An object $o_i \in O_{F_i}$ in a space $F_i$is a non-spatialcandidate skyline of F_i if there are no other objects $o_j \in O_{F_i}$ in the space $F_i$where $i \neq j$ that non-spatially dominates o_i.

This will avoid rescanning the objects of the region and repeating the process of pairwise comparisons among the objects during the computation of subsequent skyline queries. Figure5 presents the fragments constructed based on theS_{G_a}given in Figure4. The x-coordinates = {-5.6, 0, 5.3, 6.3, 9.6, 9.7, 22.7} and y-coordinates = {-2.6, 0, 6.3, 7.3, 9.7, 12.6, 24.7}. Altogether there are 28 fragments; some samples are given in Table 3.

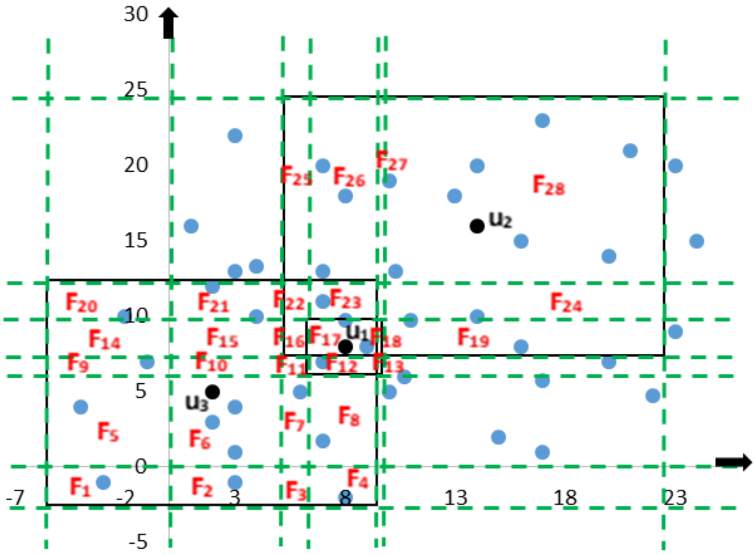


Figure5. The fragments derived based on the S_{G_a}given in Figure4

Table 3. Sample of fragments and their associated candidate skylines

x Coordinate (x_l, x_r)	y Coordinate (y_b, y_t)	$bl(x_l, y_b)$	$br(x_r, y_b$	$tl(x_l, y_t$	$tr(x_r, y_t)$	Fragment, F_i	Objects	Candidate skylines, $CS_{ns_{F_i}}$
-5.6, 0	-2.6, 0	-5.6, -2.6	0, -2.6	-5.6, 0	0, 0	F_1	o_{38}	o_{38}
0, 5.3	-2.6, 0	0, -2.6	5.3, -2.6	0, 0	5.3, 0	F_2	o_{45}	o_{45}
5.3, 6.3	-2.6, 0	5.3, -2.6	6.3, -2.6	5.3, 0	6.3, 0	F_3	-	-
6.3, 9.6	-2.6, 0	6.3, -2.6	9.6, -2.6	6.3, 0	9.6, 0	F_4	o_{42}	o_{42}
…	…	…	…	…	…	…	…	…
0, 5.3	0, 6.3	0, 0	5.3, 0	0, 6.3	5.3, 6.3	F_6	o_1, o_2, o_3	o_2, o_3
…	…	…	…	…	…	…	…	…
9.7, 22.7	7.3, 24.7	9.7, 7.3	22.7, 7.3	9.7,24.7	22.7, 24.7	F_{28}	$o_{26}, o_{27}, o_{29}, o_{30}, o_{31}, o_{32}, o_{33}, o_{40}$	o_{29}, o_{32}

4.4. Derive Non-Spatial Skylines

This step performs the non-spatial dominance testing given in *Definition* 8 towards the $CS_{ns_{F_i}}$lists derived in the previous step to generate the non-spatial skylines of a given group of users. In other words, the pair wise comparisons are only performed between objects that are the candidate skylines of a fragment. The objects that non-spatially dominate the other objects, given the $CS_{ns_{F_i}}$ lists where i ={1, 2, …, 28} in Table 3 areo_{18} and o_{20}, thus $Sky_{ns_{G_a}}$ = {o_{18}, o_{20}}.

4.5. Derive Spatial Skylines

This step applies the spatial dominance testing given in *Definition* 5 towards the $CS_{ns_{F_i}}$lists. It calculates the distance between each object and each user as well as the sum of the distances (*Sum Distance*). The sequence of comparisons between these objects is based on the lowest value of *Sum Distance*. An example is shown in Table 4; $o_8$will be the first object selected which is then followed by o_7. The objects that spatially dominate the other objects, given the $CS_{ns_{F_i}}$listsin Table 3 are as listed in $Sky_{S_{G_a}} = \{o_2, o_5, o_6, o_7, o_8, o_9, o_{20}, o_{25}, o_{26}, o_{32}, o_{40}\}$.

Table 4. The distance and sum distance of each object in $CS_{ns_{F_i}}$

Restaurant	u_1	u_2	u_3	Sum Distance	Restaurant	u_1	u_2	u_3	Sum Distance
o_1	7.81	17.69	2	27.5	o_{27}	13.41	6.32	20.12	39.85
o_2	6.4	16.27	1.41	24.08	o_{29}	18.38	8.6	24.83	51.81
o_3	8.6	18.6	4.12	31.32	o_{30}	18.6	7.61	23.43	49.64
o_4	6.37	15.92	5.99	28.28	o_{31}	13.41	4	19.2	36.61
o_5	3.6	13.6	4	21.2	o_{32}	11.18	2.23	17.02	30.43
o_6	1.41	11.4	5.38	18.19	o_{33}	11.18	5	16.12	32.3
o_7	1	9.43	7.61	18.04	o_{36}	12.04	8.06	15.81	35.91
o_8	1.7	8.7	7.62	18.02	o_{38}	14.21	24.04	7.81	46.06
o_9	3.16	8.6	7.81	19.57	o_{39}	9.05	17.49	3.6	30.14
o_{18}	8	8.24	14.31	30.55	o_{40}	5.5	4.76	11.52	21.78
o_{19}	6.32	6	13	25.32	o_{41}	12.64	21.63	6.08	40.35
o_{20}	3.44	6.97	10.15	20.56	o_{42}	10	18.97	9.21	38.18
o_{21}	4.47	11.66	5.38	21.51	o_{43}	10	6.32	14.31	30.63
o_{22}	7.21	12.64	7	26.85	o_{44}	10.19	17.08	6.4	33.67
o_{25}	5.09	7.61	9.43	22.13	o_{45}	8.6	20.24	6.08	34.92
o_{26}	10.63	2.23	17.2	30.06					

4.6. Derive the Final Skylines

This is the final step that combines the results produced in the steps presented in subsections 4.4 and 4.5 above. Based on *Definition* 7, the final skylines for a given group G_i is given by, $Sky_{G_i} = Sky_{ns_{G_i}} \cup Sky_{S_{G_i}}$. Thus, the final skylines for the group G_a, $Sky_{G_a} = \{o_2, o_5, o_6, o_7, o_8, o_9, o_{18}, o_{20}, o_{25}, o_{26}, o_{32}, o_{40}\}$.

4.7. Identify the Overlapping Region

This step constructs the overlapping region, O_R, between the search regions of two groups of users, say S_{G_i} and S_{G_j}. We assume that the results of the skyline queries of a group of users, say G_i, have been derived. Thus, the overlapping region indicates that the region has been scanned and it is unwise to scan it again. Figure 6 shows two search regions, S_{G_a}, the polygon with black border line and S_{G_b}, the polygon with red border line which represent the search region of group G_a and group G_b, respectively.

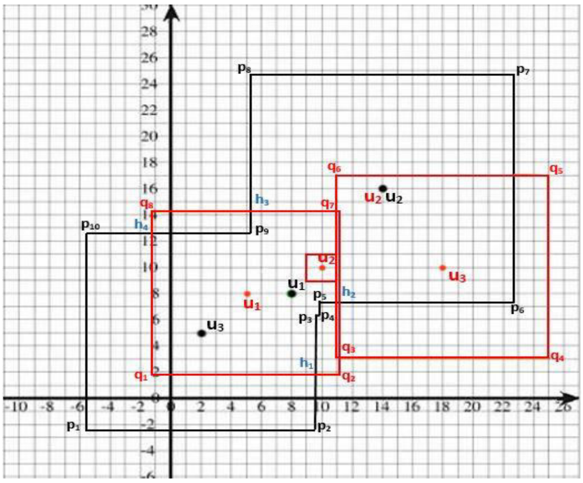


Figure6. The overlapping region between S_{G_a} and S_{G_b}

To identify the overlapping region, the following steps are performed:

(1) Get the polygon's vertices of S_{G_i}. Based on our example, $S_{G_a} = \{p_1, p_2, p_3, p_4, p_5, p_6, p_7, p_8, p_9, p_{10}\}$. Note that for simplicity, we omit the coordinates of the vertices.
(2) Get the polygon's vertices of S_{G_j}. Based on our example, $S_{G_b} = \{q_1, q_2, q_3, q_4, q_5, q_6, q_7, q_8\}$.
(3) Get the vertices of S_{G_i}that are also in S_{G_j}. Based on our example, $I_{G_a-G_b} = \{p_3, p_4, p_5, p_6, p_9\}$.
(4) Get the vertices of S_{G_j}that are also in S_{G_i}. Based on our example, $I_{G_b-G_a} = \{q_1, q_6, q_7\}$.
(5) Get the coordinates where the edges of S_{G_i} and S_{G_j} meet. Based on our example, $H = \{h_1(9.6, 1.8), h_2(11.2, 7.3), h_3(22.7, 17), h_4(5.3, 14.2), h_5(-1.2. 12.6)\}$.
(6) The overlapping region, O_R, is defined as a polygon derived based on the following ces:$I_{G_a-G_b} \cup I_{G_b-G_a} \cup H$. Based on our example, $O_R = \{h_1, p_3, p_4, p_5, h_2, p_6, h_3, q_6, q_7, h_4, p_9, h_5, q_1\}$.

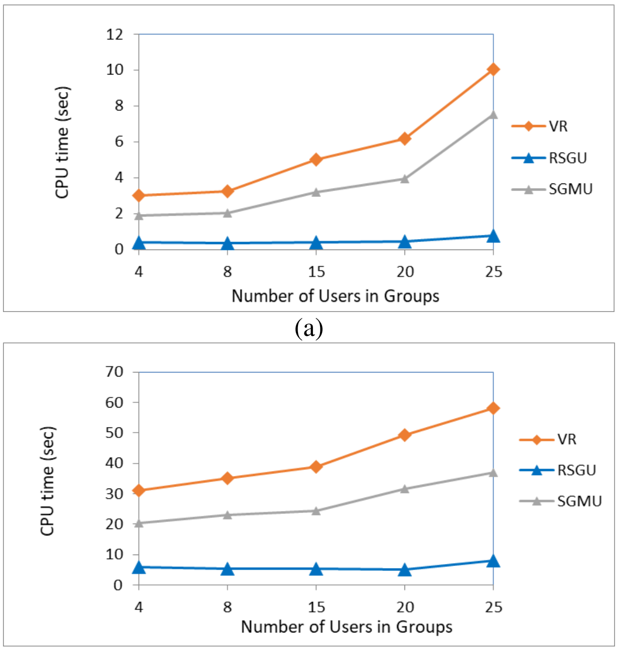
Once the O_R has been defined, the fragments derived in the earlier step are analysed. Those fragments that fall within the O_R; are retrieved together with their candidate skylines, CS_{OR}. Hence, scanning this area is no longer necessary. While for the non-overlapping area, denoted as $\neg O_R$, the following steps as discussed above will be conducted: (4) Construct the fragments of the non-overlapping region, i.e. $\neg O_R = S_{G_j} - O_R$ (5 and 6) Derive non-spatial skylines and spatial skylines, respectively by considering both the lists CS_{OR} and $CS_{\neg OR}$, and (7) Derive the final skylines.

5. Performance Evaluation

In this section, we present the initial results of the experiments that we have conducted. The experiments are conducted on a PC with Intel core™ i7 processor, 2.50 GHz CPU, 16GB main memory, and 900GB hard disk. We used both real and synthetic datasets. The real dataset is obtained from median of each road line fragment data of Long Beach from the TIGER database [28]. The dataset contains 50,747 points standardised in [0, 1000] [0, 1000] space. We used synthetic dataset consisting of 100 points with different densities standardised in [0, 1000] [0, 1000] space. The density in synthetic dataset and real dataset of TIGER database is based on the number of point's falls into one square unit in normal. We ran our proposed framework, *RSGU*, *SGMU*

[9], and *VR* algorithm [19] using randomly selected user points and objects in each dataset. Each dataset contains a spatial attribute and two non-spatial attributes. Three experiments have been conducted as elaborated below.

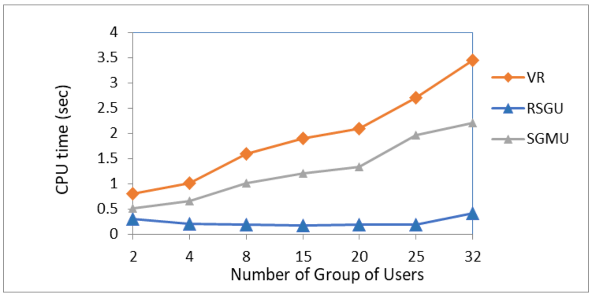
Effect of number of users in a group – Figure 7 presents the experimental results of *RSGU*, *SGMU* [9], and *VR* algorithm [19] for both the (a) syntactic dataset and (b) real dataset with respect to the CPU time when the number of users in a group is varied. In this experiment we varied the number of users in a group from 4 to25whilethenumberofobjectsisfixedto100and50000, respectively with 32 number of groups of users, and 50% of overlapping region. It is obvious, when the number of users in a group increases, the CPU time also increases. Nonetheless, our proposed framework, *RSGU*, outperforms both the *SGMU* and *VR* algorithm; while *SGMU* is better than *VR* algorithm.



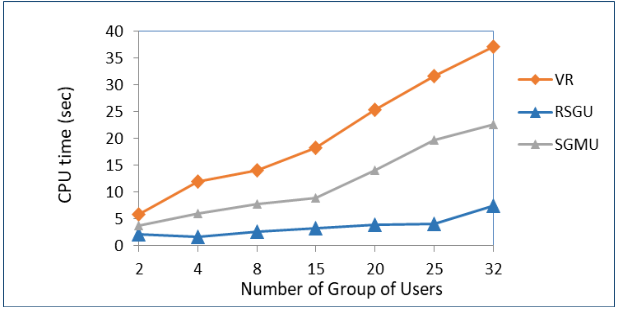
(b)

Figure 7. CPU time with varying number of users in a group over (a) synthetic dataset and (b) real dataset

Effect of number of groups – Figure 8 presents the experimental results of *RSGU*, *SGMU* [9], and *VR* algorithm [19] for both the (a) syntactic dataset and (b) real dataset with respect to the CPU time when the number of groups is varied. In this experiment, the number of users in a group is fixed to 10, the number of objects is fixed to100and50000, respectively with 50% of overlapping region while the number of groups of users is varied from 2 to 32. It is obvious that when the number of groups increases, the CPU time also increases. Nevertheless, our proposed framework, *RSGU*, outperforms both the *SGMU* and *VR* algorithm; while *SGMU* is better than *VR* algorithm.



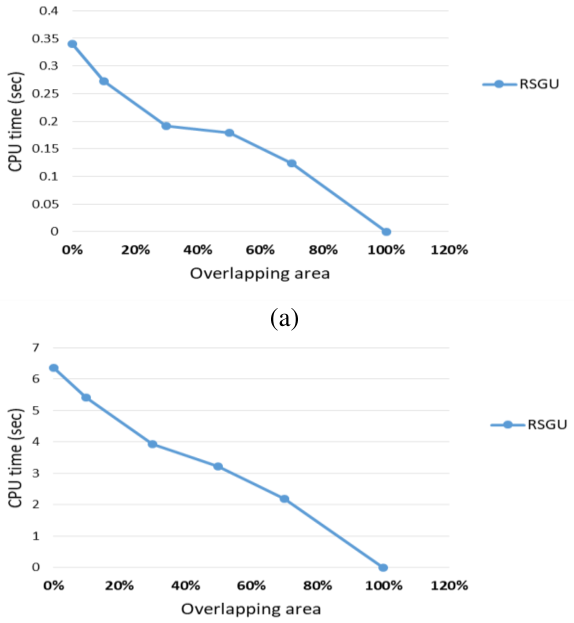
(a)



(b)

Figure8. CPU time with varying number of groups of users over (a) synthetic dataset (b) real database

Effect of percentage of overlapping area – Figure 9 presents the experimental results of *RSGU* for both the (a) syntactic dataset and (b) real dataset with respect to the CPU time when the percentage of overlapping area is varied. In this experiment, we did not compare with the *SGMU* and *VR* algorithms as determining the overlapping area is not considered in these algorithms. Here, in this experiment, the number of users is fixed to 10, the number of group of users is fixed to 16, the number of objects is fixed to 100 and 50000, respectively, while the percentage of overlapping region is varied from 0 to 100 for both datasets. It is obvious that when the percentage of overlapping region increases, the CPU time decreases for both the syntactic and real datasets.



(a)

(b)

Figure9.CPU time with varying percentage of overlapping region over (a) synthetic dataset and (b) real dataset

From these experiments, we can conclude that our proposed framework, *RSGU*, achieved lower CPU time as compared to our previous solution, *SGMU*, and the *VR* algorithm, even when the number of users in a group and the number of groups increase, while the CPU time decreases when the percentage of overlapping region increases. This shows that the fragmentation strategy proposed in our solution, *RSGU*, has significantly reduced the CPU time needed in the computation of subsequent skyline queries of a group of users.

6. CONCLUSION

This paper presents our proposed framework, *RSGU*, which aims at deriving skylines for groups of users by avoiding the unnecessary computation of skylines. Our initial results show that the performance of our proposed framework with respect to CPU time is better compared to [9] and [19]. As future works, we attempt to (1) organise the fragments in a hierarchy [7] so that the scanning time taken to search for the fragments that overlap with the region of a group of users under consideration can be reduced and (2) enhance our framework to identify skylines not only based on the spatial and non-spatial attributes of an object but also the closeness of the object to other interesting objects in the area. This would give more benefit to the users, since they might want to visit a place where there are several other interesting places nearby.

ACKNOWLEDGEMENTS

This work was supported by the Ministry of Higher Education Malaysia under the Fundamental Research Grant Scheme (FRGS/1/2016/ICT04/UPM/01/2) and the Universiti Putra Malaysia. All opinions, findings, conclusions and recommendations in this paper are those of the authors and do not necessarily reflect the views of the funding agencies.

REFERENCES

[1] Alwan, A.A., Ibrahim, H., Udzir, N.I., and Sidi, F. Processing skyline queries in incomplete distributed databases, *Journal of Intelligent Information Systems (JIIS)*, 2017, 48(2017): 399-420.

[2] Alwan, A.A., Ibrahim, H., and Udzir, N.I. An efficient approach for processing skyline queries in incomplete multidimensional database, *Arabian Journal for Science and Engineering*, 2016, 41(8): 2927-2943.

[3] Babanejad, G., Ibrahim, H., Udzir, N.I., Sidi, F., and Alwan, A.A. Efficient computation of skyline queries over a dynamic and incomplete database, *Journal of IEEE Access*, 2020, 8(2020):141523–141546.

[4] Bai, M., Xin, J., and Wang, G. Probabilistic reverse skyline query processing over uncertain data stream, *International Conference on Database Systems for Advanced Applications (DASFAA)*, 2012, pp. 17-32.

[5] Bartolini, I., Ciaccia, P., and Patella, M. SaLSa: computing the skyline without scanning the whole sky, *Proceedings of the International Conference on Information and Knowledge Management,* 2006, pp.405-414.

[6] Börzsönyi, S., Kossman, D., and Stocker, K. The skyline operator, *Proceedings of the International Conference on Data Engineering*, 2001, pp.421-430.

[7] Brown, R.A. Building a balanced KD tree in O (kn log n) time, *arXiv preprint arXiv:1410.5420*,2014.

[8] Chomicki, J., Godfrey, P., Gryz, J., and Liang, D. Skyline with presorting: theory and optimizations, *Proceedings of the Intelligent Information Processing and Web Mining*, 2005, pp.595-604.

[9] Dehaki, G.B., Ibrahim, H., Udzir, N.I., Sidi, F., and Alwan, A.A. Framework for processing skyline queries for a group of mobile users, *Proceedings of the 20th International Conference on Information Integration and Web-based Applications & Services (iiWAS2018)*, 2018, pp. 331-337.

[10] Dzolkhifli, Z.,Ibrahim, H., Sidi, F., Affendey, L.S., Mohd Rum, S.N., and Alwan, A.A. A skyline query processing approach over interval uncertain data stream with k-means clustering technique, *Proceedings of the Eleventh International Conference on Advances in Databases, Knowledge and Data Applications (DBKDA 2019)*, 2019, pp. 51-56.

[11] Godfrey, P., Shipley, R., and Gryz, J. Maximal vector computation in large data sets, *Proceedings of the International Conference on Very Large Database*, 2005, pp.229-240.

[12] Gulzar, Y., Alwan, A.A., and Turaev, S. Optimizing skyline query processing in incomplete data, *Journal of IEEE Access*, 2019, Vol. 7, pp. 178121–178138.

[13] Halliday, R. Walker. Fundamentals of physics electricity and magnetism,2011.

[14] Khalefa, M.E., Mokbel, M.F., and Levandoski, J.J. Skyline query processing for uncertain data, *Proceedings of the Conference on Information and Knowledge Management*, 2010, pp. 1293–1296.

[15] Kossmann, D., Ramsak, F., and Rost, S. Shooting stars in the sky: an online algorithm for skyline

queries, *Proceedings of the 28th International Conference on Very Large Databases (VLDB'02)*, 2002, pp. 275-286.

[16] Kung, H.T., Luccio, F., and Preparata, F. P. On finding the maxima of a set of vectors, *Journal of the ACM (JACM)*, 1975, 22(4): 469-476.

[17] Lawal, M.M., Ibrahim, H., Mohd Sani, N.F., and Yaakob, R.An indexed non-probability skyline query processing framework for uncertain data, *Proceedings of the 5th International Conference on Advanced Machine Learning Technologies and Applications (AMLTA-2020)*, 2020, pp. 289-302.

[18] Lee, J, Im, H., and You G-W. Optimizing skyline queries over incomplete data, *Information Sciences*, 2016, 361-362: pp. 14-28.

[19] Mohammad, S.A., Ma, G., and Morimoto, Y. A spatial skyline query for a group of users, *JSW*, 2014, 9(11): 2938-2947.

[20] Mohd Saad, N.H., Ibrahim, H., Sidi, F., and Yaakob, R. Skyline probabilities with range query on uncertain dimensions, *Advances in Computer Communication and Computational Sciences, part of the Advances in Intelligent Systems and Computing Book Series (AISC)*, 2018, pp. 225-242.

[21] Mohd Saad, N.H., Ibrahim, H., Sidi, F., and Yaakob, R. Non-Index based skyline analysis on high dimensional data with uncertain dimensions, *Proceedings of the 13th International Baltic Conference on Databases and Information Systems*, 2018, pp. 272-286.

[22] Mohd Saad, N.H., Ibrahim, H., Sidi, F., Yaakob, Y., and Alwan, A.A. Computing range skyline query on uncertain dimensions, *Proceedings of the 27th International Conference on Database and Expert Systems Applications (DEXA 2016)*, 2016, pp. 377-388.

[23] Mohd Saad, N.H., Ibrahim, H., Sidi, F., Yaakob, R., and Alwan, A.A. A framework for evaluating skyline query over uncertain autonomous databases, *Proceedings of the International Conference of Computational Science (ICCS 2014)*, 2014, pp. 1546-1556.

[24] Papadias, D., Tao, Y., Fu, G., and Seeger, B. An optimal and progressive algorithm for skyline queries, *Proceedings of the ACM SIGMOD International Conference on Management of Data*, 2003,pp. 467-47.

[25] Pei, J., Jiang, B., Lin, X., and Yuan, Y. Probabilistic skylines on uncertain data, *Proceedings of the International Conference on Very Large Database*, 2007, pp. 15–26.

[26] Sharifzadeh, M. and Shahabi, C. The spatial skyline queries, *Proceedings of the 32nd International Conference on Very Large Data Bases*, 2006, pp.751-762.

[27] Sharifzadeh, M., Shahabi, C., and Kazemi, L. Processing spatial skyline queries in both vector spaces and spatial network databases, *Journal of ACM Transactions on Database Systems (TODS)*, 2009, 34(3): 14.

[28] Tiger. Available at: http://tiger.census.gov/

[29] Wang, H. and Zhang, W. The τ-skyline for uncertain data, *Proceedings of the 26th. Canadian Conference on Computational Geometry*, 2014.

RESEARCH ON TASK SCHEDULING STRATEGY BASED ON THE TRUSTWORTHINESS OF MAPREDUCE

QIN Jun[1,2], SONG Yanyan[3] and ZONG Ping[4,5]

[1]Communication University of China, Nanjing, China
[2]Nanjing University of Posts and Telecommunications, China
[3]Communication University of China, Nanjing, China
[4]Nanjing University of Science and Technology Zijin College, China
[5]Nanjing University of Posts and Telecommunications, China

ABSTRACT

With the rapid development and popularization of information technology, cloud computing technology provides a good environment for solving massive data processing. Hadoop is an open-source implementation of MapReduce and has the ability to process large amounts of data. Aiming at the shortcomings of the fault-tolerant technology in the MapReduce programming model, this paper proposes a reliability task scheduling strategy that introduces a failure recovery mechanism, evaluates the trustworthiness of resource nodes in the cloud environment, establishes a trustworthiness model, and avoids task allocation to low reliability node, causing the task to be re-executed, wasting time and resources. Finally, the simulation platform CloudSim verifies the validity and stability of the task scheduling algorithm and scheduling model proposed in this paper.

KEYWORDS

Cloud Environment, Failure Recovery Mechanism, Task Scheduling Algorithm .

1. INTRODUCTION

With the extensive use of Internet, e-commerce and other application technologies, how to use data analysis technology to extract resources with commercial and application value from massive data resources is a problem worthy of in-depth study. MapReduce programming model is distinguished by its advantages of good scalability, fault tolerance and large-scale parallel processing, and has become the key technology of big data processing and analysis [1]. Google company proposed to use MapReduce programming model to deal with parallel computing of large-scale data sets. MapReduce uses the design ideas of functional programming language and vector programming language for reference, and proposes a simple parallel programming method, which realizes basic parallel computing tasks by using map and reduce function programming. MapReduce distributes the task of large-scale dataset operation to each sub node under the management of the master node to complete together, and then obtains the final result by integrating the intermediate results of each sub node, so as to realize the reliable execution and fault tolerance mechanism of the task [2].

One of the purposes of MapReduce programming model design [3,4] is to use a large number of working nodes to process massive amounts of data, so MapReduce must be able to quickly

David C. Wyld et al. (Eds): AIMLA, DBDM, CCNET - 2021
pp. 51-59, 2021. CS & IT - CSCP 2021 DOI: 10.5121/csit.2021.111304

handle failed machines [5]. In the MapReduce programming model, the Job Tracker node periodically pings each Task Tracker node. If the Task Tracker node does not respond within a specified time, the node will be marked as invalid. All tasks completed on the failed node will be set to an unexecuted state and assigned to be executed again on other Task Tracker nodes.

This paper makes the following assumptions about the cloud environment platform:

(1) The cloud environment platform is heterogeneous.
(2) All nodes in a heterogeneous platform have and only have two working states: normal operation state and failure state. When a node fails, it is in a failed state.
(3) The failure state of any node has nothing to do with other nodes and will not affect other nodes in a normal state.
(4) If a node is in an idle state when it fails, the node will be replaced by another node in the standby state, which does not affect the schedulability of the task. If there are tasks being executed on the failed node, the failure recovery mechanism and replacement mechanism will be used to ensure that the node is restored to a normal operating state.

2. Reliability Task Scheduling Strategy with Introduction of Failure Recovery Mechanism

Trustworthiness is a parameter to evaluate the reliability of a system or product [6,7]. Here reliability means that a certain system or product is trustworthy or trustworthy. Usually, trustworthiness metrics are as follows.

(1) Trustworthiness

Trustworthiness refers to the probability that the system completes the specified task within a specified time t according to user requirements. It is expressed as a time function:

$$R(t) = P(T > t) 0 < t < \infty$$

T represents the working time before the failure. Untrustworthiness and trustworthiness have a complementary relationship. Untrustworthiness is also called failure probability. Failure probability F is also a function of time.

$$F(t) = P(t < T) 0 < t < \infty$$
$$F(t) = 1 - R(t)$$

(2) Failure rate

The failure rate refers to the probability that the product or system has not failed at the moment of work until the time t, and the system will fail in the next unit time after the time t.

Assuming that the failure rate of a system or product obeys an exponential distribution, the relationship between trustworthiness and failure rate is:

$$R(t) = e^{-\lambda t}$$

(3) Mean Time to Failure (MTTF)

Suppose that under the same test conditions, there are N irreparable systems or products whose failure time is:

$$MTTF = \frac{1}{N}\sum_{i=1}^{N} ti$$

For some systems or products that fail to be repaired, the average time before failure is their average lifespan. If the probability of system failure obeys an exponential distribution, there is

$$MTTF = \int_{0}^{\infty} e^{-\lambda t}\, dt = \frac{1}{\lambda}$$

(4) Mean Time Between Failures, MTBF

The mean time between failures refers to how long the system or product runs on average before a failure occurs. The continuous time of each work is calculated as t_1, t_2, ... t_N, and the mean time between failures is:

$$MTBF = \frac{1}{N}\sum_{i=1}^{N} ti$$

(5) Mean Time To Repair (MTTR)

The average repair time refers to the average repair time of a system or product that can be repaired. The repair time of a system or product is not a certain time. Assuming the repair rate is μ, and obeys the exponential distribution, the average repair time can be expressed as:

$$MTTR(t) = \int_{0}^{\infty} tue^{-ut}\, dt = \frac{1}{u}$$

The task scheduling strategy that aims at maximizing the reliability of task scheduling and minimizing the total response time of the task is a common task scheduling model in the cloud environment [8].

Definition 1: Node model. Assuming that the nodes in the cloud environment are heterogeneous, there are differences in the performance of each node Nj in N. The node model to be studied is described as an undirected graph G(T, E), the node set: N={N_1,N_2……,Nm}, m is the number of nodes, E represents the edge set of the graph, and mainly represents the number of nodes Interrelationships. Expressed by an n*m matrix RT, RTij represents the running time of task i on node j. A matrix TI of order m*m is used to represent the communication volume between nodes, where the communication volume between node a and node b is represented as TIab. In the undirected graph G(T, E), the virtual machines are independent of each other and have no dependencies.

Definition 2: Task response time T_{ij}. It mainly includes task waiting time WT_{ij}, task transmission time CT_{ij} and task execution time RT_{ij}.

$$T_{ij} = WT_{ij} + CT_{ij} + RT_{ij} \qquad (1)$$

The time when all tasks are completed, we use ST to indicate.

$$ST = max \quad T_{ij} \qquad (2)$$

Definition 3: Trustworthiness. The trustworthiness of the node is mainly considered from the two aspects of node failure rate and failure repair rate [9]. The failure of the node is mainly the communication link between the node and the node and whether the node itself fails. Assuming that the probability of node failure and the probability of failure of the communication link between nodes obey the Poisson distribution with parameters σ and ξ respectively, the probability of node failure p times within the time interval [0, t] is the same. It can be seen that the communication link failure probability $\lambda_p(t) = e^{-\sigma t}/p!$ within the time interval [0, t]. Assume that the probability of a node failure and recoverable obeys the Poisson distribution with parameter $\theta_p(t) = e^{-\xi t}/p!$, where the communication link cannot be recovered when the communication link fails. Assume the probability $\mu_p(t) = e^{-\varepsilon t}/p!$ that the node is repaired p times in the time interval [0, t].

Define $P_j(t)$ as the probability that there is no failure (ie p=0) on node j in the time interval [0,t], and its probability value is

$$P_j(t) = e^{-\sigma_j t} = e^{-\sigma_j RT_{ij}} \qquad (3)$$

Define $C_j(t)$ as the probability of no failure (ie p=0) on the communication link between node a and node b in the time interval [0, t], and its probability value is

$$P_j(t) = e^{-\xi t} = e^{-\xi_j TI_{ab}/Net_{ab}} \qquad (4)$$

In equation 4, TI_{ab}/Net_{ab} represents the communication time between nodes on the communication link.

Define the probability that $R_j(t)$ does not repair on node j in the time interval [0, t] (ie p=0), and its probability value is

$$R_j(t) = e^{-\varepsilon t} = e^{-(1-\mathrm{I}_j)\varepsilon_j RT_{ij}} \qquad (5)$$

According to the above introduction, the probability of node j completing task i is K_{ij}.

$$K_{ij} = P_{ij} \times C_{ij} \times R_{ij} = e^{-\sigma_j RT_{ij}} \times e^{-\xi_j TI_{ab}/Net_{ab}} \times e^{-\varepsilon_j RT_{ij}} = e^{-\sigma_j RT_{ij} - \xi_j TI_{ab}/Net_{ab} - \varepsilon_j RT_{ij}} \qquad (6)$$

In order to improve the parallelism of the application, a job is often divided into multiple tasks and executed in parallel on multiple nodes at the same time. A task is executed on only one node. When all tasks return the task execution results, it indicates that the job is successfully completed. Assuming that N(j) is the set of nodes performing the task, the reliability of the task can be expressed as

$$Trust(N) = \prod_{i=1}^{n}\prod_{j=1}^{m} K_{ij} = \prod_{i=1}^{n}\prod_{j=1}^{m} e^{-(\sigma_j RT_{ij} + \xi_j TI_{ab}/Net_{ab} + \varepsilon_j RT_{ij})} = \prod_{i=1}^{n}\prod_{j=1}^{m} e^{trust_{ij}} \qquad (7)$$

According to formula 2 and formula 7, in order to maximize the reliability of the scheduling strategy and minimize the total response time of the task, the objective function is set as formula 8.

$$MinF = -x\ ln(Trust(N)) + ST \qquad (8)$$

Since the value of the total response time of the task is larger than the value range of the reliability, x can be used as a scale factor to coordinate the proportion of reliability and time cost to prevent the time cost from controlling the objective function value.

3. Reliability Task Scheduling Algorithm Introducing Failure Recovery Mechanism

Aiming at the task scheduling problem in the cloud environment, the trust evaluation model considering the failure recovery mechanism is introduced into the ant colony simulated annealing algorithm, and the ant colony simulated annealing algorithm considering the failure recovery mechanism is proposed.

This paper proposes an ant colony algorithm based on SA (Ant Clony Optimization Simulated Anealling, ACOSA) [10]. Its principle is to find the local optimal task scheduling solution for tasks T_i and node N_j through ACO, and then use SA for local optimization. Thereby assigning tasks to appropriate heterogeneous resource nodes for execution. Among them, ACO sets the initial pheromone concentration to a constant before the ant searches, which will increase the ant's search space:

$$\tau_j(0) = c$$

With the increase of time, the concentration of pheromone becomes higher and higher, and the ant can choose the appropriate path according to the concentration of the pheromone on the path that the ant walked last time.

When the task is scheduled to be executed on the resource node, the trustworthiness reflects the reliability of the service provided by the target resource node. The ACOSA algorithm fully considers the heterogeneous characteristics of resource nodes but does not consider the impact of the trustworthiness of the resource nodes on the task scheduling results. For this reason, we take the maximization of credibility and the minimization of time cost as the objective function, and the Function as a heuristic function of ACO.

$$\eta_{ij} = \frac{1}{-x\ln(K_{ij}) + RT_{ij}}$$

In the actual cloud environment, when a node fails, the tasks assigned to the node will often be re-executed. The advantage of introducing a failure recovery mechanism is that for those failures that can be recovered, the node can recover the tasks that have stopped executing by running the failure recovery program.

Execution steps based on reliability task scheduling strategy:

(1) Initialization parameters. Set the initial temperature T_{max}, the maximum number of iterations Ite_{rmax}, and the initial pheromone τ_{ij}.

(2) Construct a feasible solution. Ant j selects the appropriate node according to the selection migration rule, adds the selected node to the taboo table, and instructs all tasks to be allocated to appropriate node resources.
(3) Update pheromone. When all ants complete the path search, a local optimal solution is generated, and the local optimal solution is used to update the local pheromone.
(4) SA performs partial optimization. According to the local optimal solution obtained by ACO, use SA to optimize the local optimal solution to obtain a new solution.
(5) Metropolis guidelines. According to Metropolis criterion, judge whether the new solution constructed by SA will be accepted.
(6) Termination criteria. Cool down the current temperature and determine whether the termination criterion is met. If it is met, execute (7), otherwise return to (4).
(7) Global pheromone update. Update the global pheromone according to the candidate solution generated by SA, the number of iterations is increased by 1. If the number of iterations is greater than $Iter_{max}$, all steps are terminated, otherwise, return to (2).

4. SIMULATION

In the simulation experiment, the influence of the failure recovery mechanism and its parameters on the evaluation of the trustworthiness of the resource node is discussed, and the performance of the algorithms is compared in the case of different nodes and the number of tasks.

The experimental environment parameters are set as follows: task Rcc is 0.1, 1, 10 respectively; the number of tasks is 100, the number of nodes is 20, and the number of communication links is 20. Set up two kinds of nodes with low trustworthiness, which account for 20% and 30% of the total number of nodes respectively, and the probability of execution failure of the two types of nodes is 80% and 50% respectively. The experimental results are shown in Figure 1.

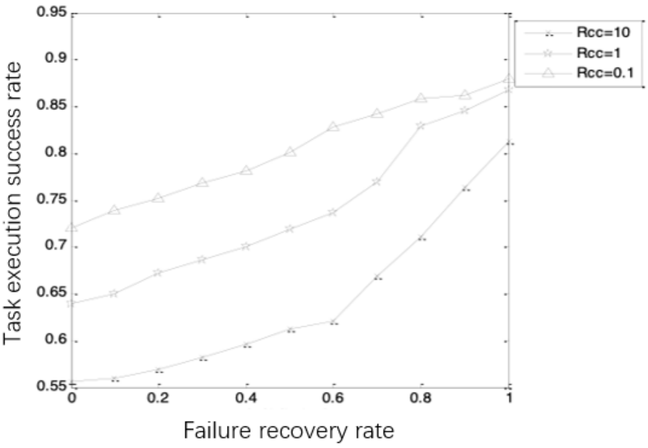


Figure 1. Task execution success rate under different failure recovery rates (without limiting the maximum number of recovery times)

It can be seen from Figure 1 that when there is no failure recovery mechanism (that is, uk=0), the probability of successful completion of the three types of tasks is relatively low. As the failure recovery rate u_k increases, the probability of successful execution of the three types of tasks slowly increases. When u_k is 1, the probability of successful task execution does not reach 100%. This is because although recoverable failures can be recovered through the application, some failures are not recoverable. The two curves with Rcc of 0.1 and 10 in Figure 1 represent calculation-intensive tasks and communication-intensive tasks, respectively. The communication-intensive tasks use the communication link for a relatively long time, and the probability of communication link failure is calculated. Intensive tasks are high. Since communication link failures are unrecoverable, the probability of successful completion of communication-intensive tasks is greater than that of computationally-intensive tasks.

In order to study the effect of failure recovery rate on task execution time, this article conducted experiments on three different types of tasks, and the experimental results are shown in Figure 2.

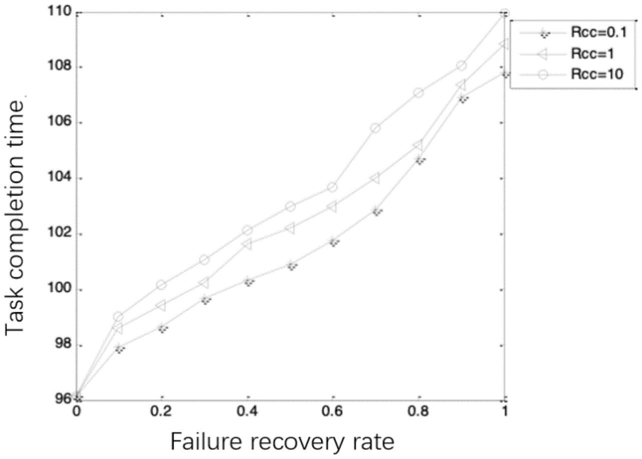


Figure 2. Task completion time corresponding to different failure rates

It can be seen from Figure 2 that different failure recovery rates correspond to different task execution times. As the failure recovery rate increases, the task completion time also gradually increases. This is because the service overhead generated during the failure recovery process causes the task execution time to increase. In an ideal situation, the node has no failure u_k=0, all tasks can be successfully executed according to the assigned node, and the task completion time is the shortest. In Figure 2, when the failure recovery rate is less than 0.6, the task completion time increases slowly, but when the failure recovery rate exceeds 0.6, the task completion time increases sharply. This is because the failure recovery time will increase when the node fails to recover. Frequent failure recovery will increase the task completion time. It can be seen that choosing an appropriate failure recovery rate has an important impact on the task completion time.

Experiments have proved that the introduction of a failure recovery mechanism does increase the probability of successful task execution, indicating that the failure recovery mechanism is effective, but time and resource overhead will also be incurred during the failure recovery process, so we have to choose an appropriate failure recovery probability. In the case of comparing the number of different tasks, compare the task execution success rate and the value of the objective function. Compare the FCFS algorithm (First Come First Service, FCFS) and ACOSA algorithm, where the failure recovery rate u_k is set to 0.6.

Figures 3 and Figures 4 respectively show different task execution success rates corresponding to different task numbers and the objective function values corresponding to different tasks.

It can be seen from Figure 3 that when the number of tasks gradually increases, the task execution success rate of the FCFS algorithm and the ACOSA algorithm is gradually decreasing, but the task execution success rate of the ACOSA algorithm with the introduction of a failure recovery mechanism is significantly higher than that of the FCFS algorithm.

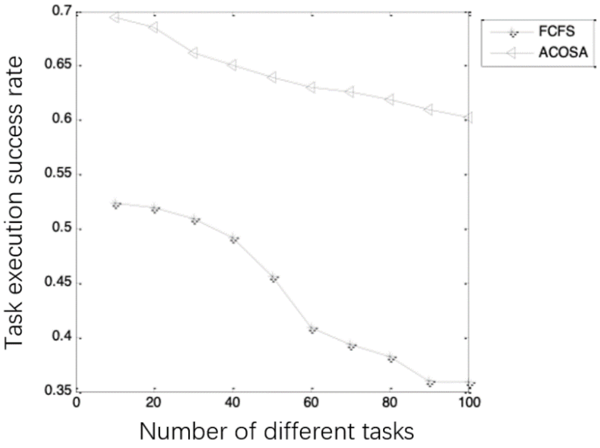


Figure 3. Different task numbers correspond to different task execution success rates

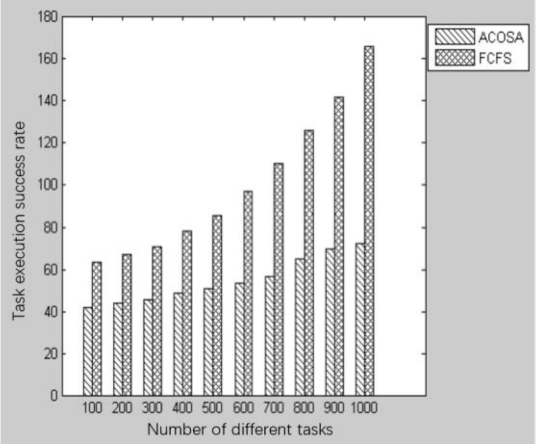


Figure 4. Objective function values corresponding to different tasks

It can be seen from Figure 4 that the task completion time of the ACOSA algorithm with the introduction of the failure recovery mechanism is shorter than the task completion time of the FCFS algorithm, which proves that the introduction of the failure recovery mechanism can not only improve the execution success rate of the task, but also select the appropriate failure recovery The performance of the ACOSA algorithm with the introduction of a failure recovery mechanism is better than that of the FCFS algorithm in the case of low rate.

5. CONCLUSION

A cluster composed of a large number of resource nodes in a cloud environment has the characteristics of heterogeneity, dynamics, and uncertainty. Reliability refers to the probability that a task submitted by a user is successfully completed. In a cloud environment with a large number of resource nodes, node failure events are inevitable. Aiming at the defects of the MapReduce fault tolerance mechanism, this paper proposes a reliability task scheduling strategy that introduces a failure recovery mechanism to evaluate the trustworthiness of resource nodes in the cloud environment. Establish a trustworthiness model to avoid assigning tasks to nodes with low reliability, causing tasks to be re-executed, and wasting time and resources. Finally, the simulation platform CloudSim verifies the validity and stability of the task scheduling algorithm and scheduling model proposed in this paper.

ACKNOWLEDGEMENTS

The authors would like to thank the support from General project of natural science research in universities of Jiangsu Provincial Department of Education (No. 19KJD520007).

REFERENCES

[1] CHENG Yan, ZHANG Yun. (2018) "Improvement and research on Apriori algorithm based on MapReduce -HBase". Journal of Nanjing University of Posts and Telecommunications. Vol.38, No. 5, pp91-99.

[2] LI Jun.（2021）"Design of online aggregation optimization of big data based on MapReduce". Journal of Hebei University, Vol. 41, No. 2, pp212-217.

[3] Jianjiang Li, Yajun Liu, Peng Zhang, Wei Chen, Lizhe Wan. (2020) "Map-Balance-Reduce: An improved parallel programming model for load balancing of MapReduce". Future Generation Computer Systems. Vol. 4, No. 1. pp150-157.

[4] Zhao Xincan, Zhu Yun, Mao Yimin. (2020) "Incremental mining algorithm of Apriori based on MapReduce". Application Research of Computers. Vol. 37, No. S2, pp73-75+79.

[5] WU Lizhen, KONG Chun, CHEN Wei. (2021) "Short-term load forecasting based on linear regression under MapReduce framework". Journal of Lanzhou University of Technology. Vol. 47, No.1, pp97-104.

[6] ZHANG Jie. (2016) "The Trust of Networked Software Measurement Model to Optimize the Simulation Analysis". Computer Simulation. Vol. 33, No. 10, pp278-281+299.

[7] JIANG Jing, YU Yonghong, ZHAO Weibin. (2020) "Research on A Trust Model Based on the QoS and Malicious Node Deletion". Computer & Digital Engineering. Vol. 48, No. 1, pp98-105.

[8] JIANG Qiang, YI Chun-lin, ZHANG Wei, GAO Sheng. (2021) "The Multi-objective Path Planning for Mobile Robot Based on Ant Colony Algorithm". Computer Simulation. Vol. 38, No. 2, pp318-325.

[9] WU Chun, YOU Xiaojian, LYU Tao. (2018) "Research on multipath secure routing based on confidence degree". Journal of Northeast Normal University (Natural Science Edition). Vol. 50, No. 4. pp66-72.

[10] LIU Pengfei, MAO Yingchi, WANG Longbao. (2019) "Task assignment method based on cloud-fog cooperative model". Journal of Computer Applications. Vol. 39, No.1, pp8-15.

AUTHORS

Qin Jun, Professor, graduated from Nanjing University of Posts and telecommunications. He has presided over and participated in more than 30 scientific research projects, won more than 10 awards, published more than 80 academic papers in academic journals and international conferences, and published 5 teaching materials.

COVID-19 TWITTER SENTIMENTS ACROSS THE UNITED STATES IN AUGUST 2020

Umesh R. Hodeghatta[1] , Ph.D and Sanath V. Haritsa[2]

[1]Northeastern University, Boston, MA, USA;
[2]NU-Sigma U^2 Analytics Lab, India

ABSTRACT

COVID-19 has drastically affected the entire nation. This study involved collecting tweets and analyzing the COVID tweets for August 2020. The aim was to understand whether people have expressed sentiments related to COVID-19 across all the states of the United States and find any correlation between the sentiment tweets and the number of actual cases reported. Around 400000 COVID-19 Twitter data was collected for August 2020 from the primary Twitter database. A simple NLP-based unigram sentiment analyser, a novel approach different from the traditional machine learning approach, was adopted to identify twitter sentiments. The results indicate that tweets related to COVID demonstrate the two types of sentiments, one related to the deaths and the other about the COVID symptoms.

Furthermore, the results show that the sentiments for each category vary from State to State. For example, states of New York, California, Texas are higher tweets sentiments regarding expressing death sentiment, and states of New York, California, Nevada, are higher regarding sentiments of expressing COVID-19 symptoms with an accuracy of 83%. As a part of the research, a new sentiment scorecard was created to provide a sentiment score based on the sentiments of the tweets expressed to the actual reported death cases. The sentiment scores for the 'symptoms' class are higher for Maryland, New Jersey, and Oregon, whereas sentiment scores for the 'death' class are higher for Virginia, Delaware, and Hawaii. These sentiment scores indicate that the Twitter users of these states are actively tweeting about symptoms and deaths even though the actual reported cases are less in these states. The analysis results also found no or little correlation between the COVID Tweets and the number of COVID death cases reported across all the states.

KEYWORDS

COVID 19, Twitter Behaviour, Twitter Analytics, Sentiment Analysis, Big Data Analytics, Business Analytics.

1. INTRODUCTION

Social media has changed the way people communicate and propagate information. Social media promotes individuals to connect with their friends and families and share pictures, videos, and their thoughts about the day-to-day activities, including experiences of using a product, services, latest news, or any other important information within their social network [5]. Over the last decade, social media also played an essential role in creating awareness and knowledge about public events, science & technology, world politics, and even public health.

On December 31, 2019, an outbreak of a new virus, severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2), coronavirus disease 2019 (COVID-2019), was announced by the Wuhan Municipal Health Commission. As of October 25, 2020, the virus resulted in total

David C. Wyld et al. (Eds): AIMLA, DBDM, CCNET, NLTM - 2021
pp. 61-73, 2021. CS & IT - CSCP 2021 DOI: 10.5121/csit.2021.111305

confirmed cases of 42.6 Million and more than 1.15 million deaths. According to the available statistics, 80% of COVID-19 infections are mild, 15% are severe, and 5% are critical (Coronavirus Resource Center, 2020). Within weeks of the announcement of COVID-19, information started spreading on social media platforms. People started talking about the diseases, precautions to be taken care of, number of cases, number of deaths, symptoms, isolations, quarantine, travel information, and other relevant data about the diseases, which created awareness among people also created panic. Further, COVID-19 itself caused several fake claims about the disease, as well as spreading racism instead of making use of social media for the good cause [6, 2]. Therefore, social media played an important role during the health crisis communications and changed the tourists risks perception [11].

Researchers have considered Twitter as an informational media and a popular microblogging platform among many social network platforms because of its ease and instantaneous reach to millions. Messages can be written to the point, crisp, and within a limited number of words and convey your points across to millions almost immediately [5]. Twitter is a micro-blogging site founded in 2006. It allows people to post their thoughts in a text from using just 140 characters. These posts are popularly known as tweets, which may include texts, and URLs. Today, Twitter has more than 600 million users [8]. Every three days, a billion tweets are being sent which reaches millions of people [8]. People use Twitter to track news, find out what others are talking about, the latest in politics, technology, events around your city, the latest phone or gadget, jobs in your area, and find out what people are talking about the latest movies. Twitter allows sharing the latest information, news, and ideas and solicits suggestions or ideas instantaneously across the globe unlike ever before. Some users may be active listeners, and some may be actively participating and exchanging information. Tweets may also contain people's opinions, views, or experiences and these tweets are available to the public [7]. Though the authenticity and meaning of the message may sometimes create confusion and conflicts due to the messages' unfriendly and short characters, it is still a popular medium.

Sentiment analysis is the process of detecting sentiments expressed in a given text [12]. The sentiments can be found in customer feedback, reviews, or critiques in different forums. The sentiment analysis aims to determine an author's attitude concerning a given topic. Under conventional circumstances, it is challenging to find out why a consumer did not buy a product, but with the help of sentiment analysis tools, it becomes easier to find out the reasons and logic behind a customer not purchasing the product [12, 3]. Apart from product and marketing, sentiment analysis is useful in areas such as politics, sociology, and psychology.

In this paper, we analyze how people used Twitter, as a social media, to convey messages related to COVID during the pandemic time. The tweets are extracted across all states within the United States during August 2020. Analysis was carried out to see whether these messages carry any sentiments related to COVID, such as symptoms, deaths, or just information. Further, a correlation with the actual number of cases was also carried out. The following sections describe the research questions, research methodology, data collection, and data analysis.

2. RESEARCH QUESTIONS

This research aimed to study how people use Twitter to communicate about COVID-19 during the pandemic time. In this study, "COVID-19" was chosen as a search topic of Twitter message, and the study was carried out to answer the following:

1. How frequently people are communicating and tweeting COVID-19 information.
2. Type of behavioral sentiments and messages users are tweeting.
3. Relation between the tweets and number of actual death cases across all the states.

4. Predicting COVID-19 sentiments.

3. METHODOLOGY

Based on the limitations of the twitter.com free open access to the database, nearly 400,000 tweets were collected for all the 52 USA states. Not all the 400,000 tweets had COVID related tweets. The flow structure and the sentiment analyser model is as shown in Figure 1.

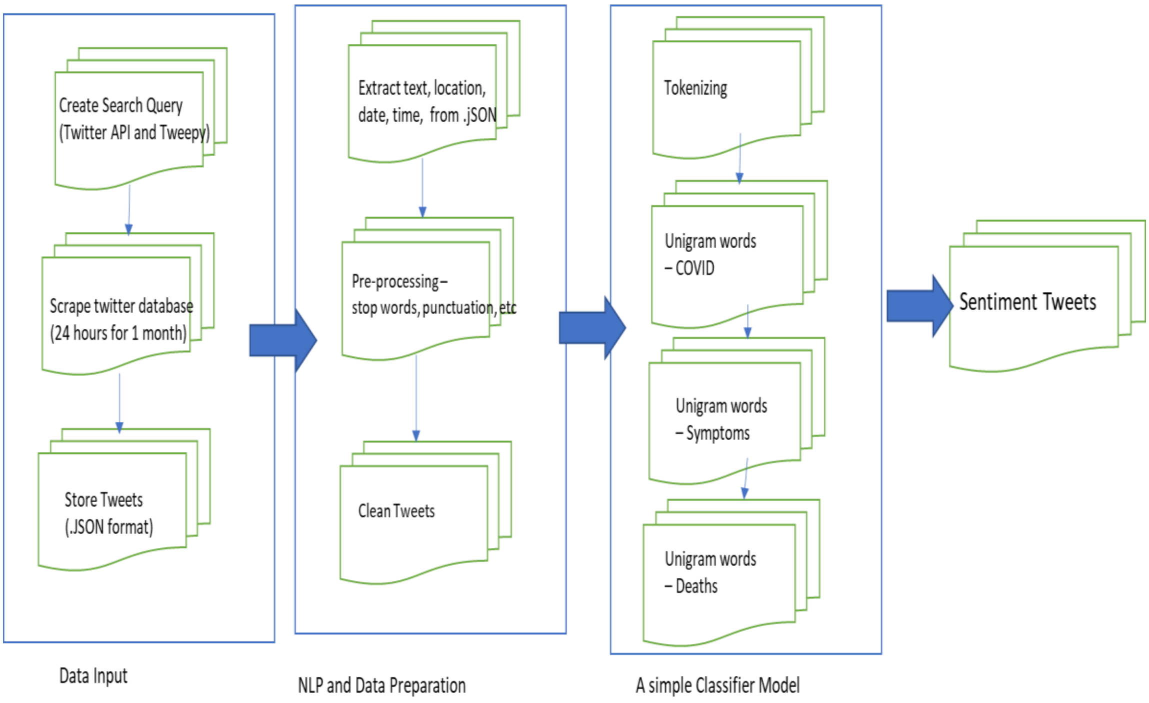
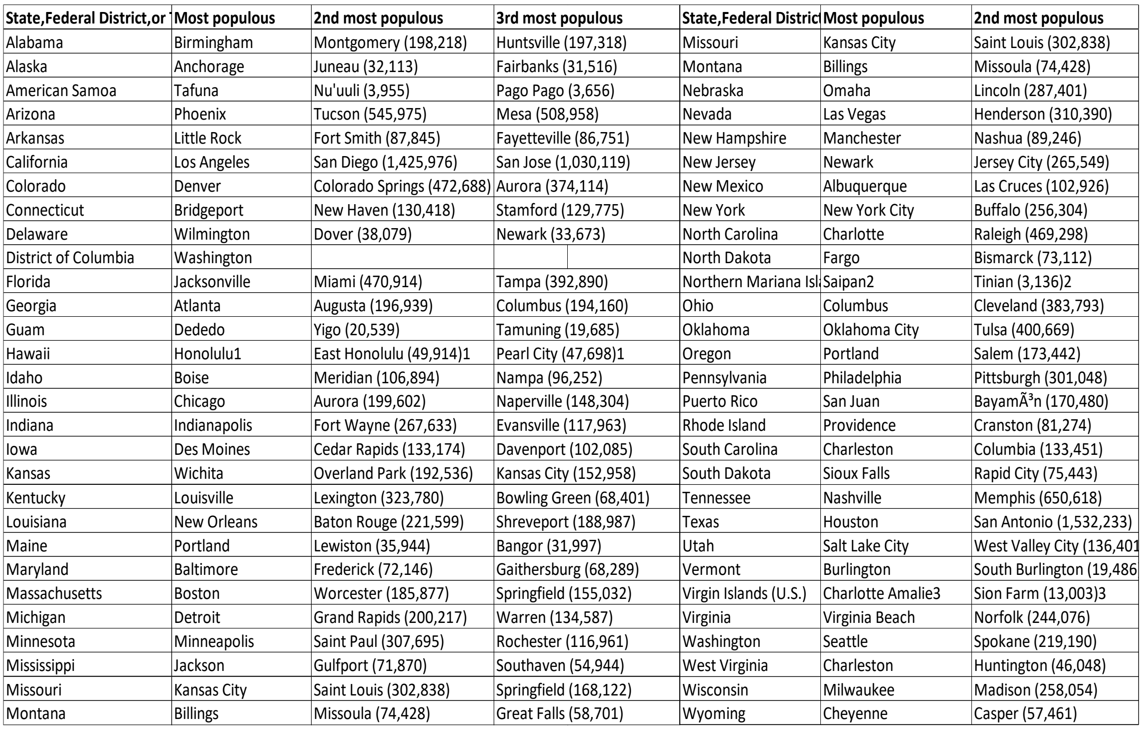

Create Search Query (Twitter API and Tweepy)
Scrape twitter database (24 hours for 1 month)
Store Tweets (.JSON format)
Extract text, location, date, time, from .jSON
Pre-processing– stop words, punctuation, etc
Clean Tweets
Tokenizing
Unigram words – COVID
Unigram words – Symptoms
Unigram words – Deaths
Sentiment Tweets
Data Input
NLP and Data Preparation
A simple Classifier Model

Figure 1. Architecture Framework of COVID Tweets sentiment analyser

3.1. Data Source (Data Input)

In this research, the primary data was directly from the Twitter database (http://www.Twitter.com). The Twitter database was accessed using Twitter search API. We selected the top three most populous cities in each State of the US and collected tweets every day for one month (of August 2020). Table 1 provides the details of states and cities used for search API geo-locations to collect the tweets.

Table 1. Twitter users cities and states used in the study



State,Federal District,or T	Most populous	2nd most populous	3rd most populous	State,Federal Distric	Most populous	2nd most populous
Alabama	Birmingham	Montgomery (198,218)	Huntsville (197,318)	Missouri	Kansas City	Saint Louis (302,838)
Alaska	Anchorage	Juneau (32,113)	Fairbanks (31,516)	Montana	Billings	Missoula (74,428)
American Samoa	Tafuna	Nu'uuli (3,955)	Pago Pago (3,656)	Nebraska	Omaha	Lincoln (287,401)
Arizona	Phoenix	Tucson (545,975)	Mesa (508,958)	Nevada	Las Vegas	Henderson (310,390)
Arkansas	Little Rock	Fort Smith (87,845)	Fayetteville (86,751)	New Hampshire	Manchester	Nashua (89,246)
California	Los Angeles	San Diego (1,425,976)	San Jose (1,030,119)	New Jersey	Newark	Jersey City (265,549)
Colorado	Denver	Colorado Springs (472,688)	Aurora (374,114)	New Mexico	Albuquerque	Las Cruces (102,926)
Connecticut	Bridgeport	New Haven (130,418)	Stamford (129,775)	New York	New York City	Buffalo (256,304)
Delaware	Wilmington	Dover (38,079)	Newark (33,673)	North Carolina	Charlotte	Raleigh (469,298)
District of Columbia	Washington			North Dakota	Fargo	Bismarck (73,112)
Florida	Jacksonville	Miami (470,914)	Tampa (392,890)	Northern Mariana Isl	Saipan2	Tinian (3,136)2
Georgia	Atlanta	Augusta (196,939)	Columbus (194,160)	Ohio	Columbus	Cleveland (383,793)
Guam	Dededo	Yigo (20,539)	Tamuning (19,685)	Oklahoma	Oklahoma City	Tulsa (400,669)
Hawaii	Honolulu1	East Honolulu (49,914)1	Pearl City (47,698)1	Oregon	Portland	Salem (173,442)
Idaho	Boise	Meridian (106,894)	Nampa (96,252)	Pennsylvania	Philadelphia	Pittsburgh (301,048)
Illinois	Chicago	Aurora (199,602)	Naperville (148,304)	Puerto Rico	San Juan	BayamÃ³n (170,480)
Indiana	Indianapolis	Fort Wayne (267,633)	Evansville (117,963)	Rhode Island	Providence	Cranston (81,274)
Iowa	Des Moines	Cedar Rapids (133,174)	Davenport (102,085)	South Carolina	Charleston	Columbia (133,451)
Kansas	Wichita	Overland Park (192,536)	Kansas City (152,958)	South Dakota	Sioux Falls	Rapid City (75,443)
Kentucky	Louisville	Lexington (323,780)	Bowling Green (68,401)	Tennessee	Nashville	Memphis (650,618)
Louisiana	New Orleans	Baton Rouge (221,599)	Shreveport (188,987)	Texas	Houston	San Antonio (1,532,233)
Maine	Portland	Lewiston (35,944)	Bangor (31,997)	Utah	Salt Lake City	West Valley City (136,401
Maryland	Baltimore	Frederick (72,146)	Gaithersburg (68,289)	Vermont	Burlington	South Burlington (19,486
Massachusetts	Boston	Worcester (185,877)	Springfield (155,032)	Virgin Islands (U.S.)	Charlotte Amalie3	Sion Farm (13,003)3
Michigan	Detroit	Grand Rapids (200,217)	Warren (134,587)	Virginia	Virginia Beach	Norfolk (244,076)
Minnesota	Minneapolis	Saint Paul (307,695)	Rochester (116,961)	Washington	Seattle	Spokane (219,190)
Mississippi	Jackson	Gulfport (71,870)	Southaven (54,944)	West Virginia	Charleston	Huntington (46,048)
Missouri	Kansas City	Saint Louis (302,838)	Springfield (168,122)	Wisconsin	Milwaukee	Madison (258,054)
Montana	Billings	Missoula (74,428)	Great Falls (58,701)	Wyoming	Cheyenne	Casper (57,461)

We developed the Twitter crawler application using Python and Java, and Tweepy library to collect tweets from the Twitter database. The Twitter API enables to access the Twitter database and interact with tweets from users. Twitter data allows developers, researchers, and others to study the twitter conversation using the search API, which gives access to the Twitter database for the last seven days. We used Tweepy library, an open source wrapper library that enables your program to communicate with Twitter platform via OAuth authentication protocol. Tweepy supports accessing Twitter via OAuth. It has an in-built method to handle OAuth. Tweepy provides access to the well documented Twitter API. With tweepy, it is possible to get any object and use any method that the official Twitter API offers.

Twitter REST API allows you to retrieve tweets and related information from Twitter. The Twitter standard REST APIs utilize a technique called 'cursoring' to paginate large result sets. Cursoring separates results into pages and provides a means to move backwards and forwards through these pages. Tweepy has a Cursor method provided which takes care of pagination for us. Method API.search is called within the cursor method to mine for tweets based on search parameters. The Cursor coupled with wait_on_rate_limit ensures we can get as many tweets as we need by specifying the count request parameter.

The search API criteria used to collect the tweets are as follows:

1) Keyword- Matches a keyword within the body of a Tweet. Multiple keywords can be specified by using OR operator and placing the words/phrases within parentheses () and wrapping them in double quotes “".
2) Geocode- Search for tweets made from particular locations around the globe by providing the following information [latitude, longitude, radius(mi)]. In order to obtain latitude, longitude coordinates for multiple cities in an efficient manner, Geopy library was used.

3) Geopy is a Python client for geocoding web services. Geopy makes it easy to locate the coordinates of addresses, cities, countries, and landmarks across the globe using third-party geocoders and other data sources.
4) Language- Only tweets made exclusively in the English language were retrieved for this project.
5) Tweet mode- Extended tweet mode was used in order to capture complete tweets. The character limit was 140 for publishing tweets at the start, but on September 26, 2017, Twitter started testing 280 characters for certain languages.
6) Filter: retweets- In order to obtain only the original tweets and leave out the retweets.

Twitter has a very stringent rate limit that prohibits requests from being made continuously. Tweepy provides a way to overcome this limit by passing wait_on_rate_limit=True into the api object. This removes the need for any complicated automation scripts being used to make sure the tweet gathering process goes on uninterrupted. Since there is a limitation to the number of tweets that can be retrieved from the Twitter database, a scheduler was introduced in our application. The scheduler process calls the crawler process automatically to retrieve the tweets over a period of time. Both the search keywords and the location are passed on as parameters to the crawler. In the Twitter search API's location 'geocode' parameter, the value is specified by the latitude, the longitude, and the radius. Search API returns the tweets by users located within a given radius of that particular latitude and longitude. For example, to retrieve tweets from Bangalore, the 'geocode' parameter is: 12.9715987, 77.5945627, 300mi. The crawler connects to the Twitter database, Twitter authenticates the user, and then pulls the relevant data based on the parameters. Finally, the tweets were dumped into a .json file for further analysis.

3.2. Data Preparation and NLP

After collecting tweets, the next step was to use python, numpy, pandas, NLTK, sklearn, ggplot(), and other libraries to process the tweets and perform sentiment analysis to extract the sentiments of the tweets.

In Twitter, users express their views in one or two sentences only. Hence most of the tweets may not contain full words and many tweets may not carry any meaning. Also, Twitter users are likely to make spelling and grammatical errors. Therefore, extracting information from Twitter messages is challenging, and a lot of time was spent cleaning the tweets before further processing and analysis.

Pre-processing was performed using python NLTK library. Using NLTK built-in libraries and regular expression techniques, we performed pre-processing of tweets. Following are some of the pre-processing techniques that was performed for cleaning tweets before feeding them into the analytics model and sentiment classifier:

(a) Converting to lower case: This has been the very first step in the pre-processing techniques. The tweets were first converted to lower case before applying other methods.
(b) Punctuation: Since the classifier is based on 'bag-of-words' techniques, removing punctuation brings more sanity to input. This involved identifying and removing coma, full stop, semicolon, colon, question marks, exclamation marks, any combination of punctuations, and removing more than one punctuation mark.
(c) Hash Tags and URLs: Many micro-blog messages contain URLs and hash tags. People use the hash-tag symbol # before a relevant keyword or phrase (no spaces) in their tweet to show more easily in Twitter Search. Clicking on a hash-tagged word in any message shows all other Tweets marked with that keyword. Hash-tags can occur anywhere in the tweet – at the beginning, middle, or end. Hash-tagged words are very popular in trending topics. However,

URLs and hash-tags do not contribute to the actual meaning of the message. Hence, both hash-tags and URLs were removed from the tweets.

(d) User name: The raw tweets contain the user name with @ sign in the beginning to identify which user has tweeted that particular message. Since the user name does not contribute to this analysis, the pre-processing technique required the removal of the user name as well as @ symbol.

(e) Stop words: Common words such as 'a', ‘an’, ‘the’ etc., do not provide useful information in classifying documents, and thus, it is worthwhile to remove these words. It is commonly known in information retrieval, as ‘stop words’. The application used a file which contained these ‘stop words’ and every tweet was searched against this file to remove the stop words.

(f) Emoticons: In order to convey emotions, such as ‘happy’, ‘sad’, ‘angry’, etc., in the text messages, it is very common for people to use symbols called emoticons. There are 30 emoticons including□, :D, □, =], =[, =(, etc. Since the analysis is based on bag-of-words, all the emoticons were removed from the tweets.

The crawler application programs for the experiments were ran on a quad-core Windows 7 operating systems machine. Tweets from different locations stored as a .json file database. The tweets are further processed and cleaned before the sentiment analysis.

After pre-processing, the tweets were meaningful to understand the pattern and find sentiments associated with the tweets.

3.3. Sentiment Analyser Model

None of the tweets retrieved from the Twitter database had any labels to perform any supervised machine learning sentiment or classification method. Therefore, as an initial step, we developed a unique but powerful NLP unigram method to look for the specific terms and categorize and label the sentiments.

After the initial cleanup of the tweets, using unigram approach to classify tweets to find out the sentimets. For this study, we identified three sentiments. Table 2 describes the sentiments and respective unigram terms to categorize the sentiments. For example, ‘Symptoms’ sentiment category is the twitter messages with terms such as ‘isolation, quarantine, symptoms’ and the COVID virus terms. Similarly, the ‘Death’ sentiment category is the twitter messages with terms such as ‘death, dying, died’ and the COVID virus terms.

Table 2. Sentiment Category

Sentiment Category	Unigram Terms
COVID virus	COIVD, Virus, corona, coronavirus
Symptoms	Isolation, Quarantine, Symptoms
Death	Died, death, dying.

3.4. Secondary Data Source

Secondary data comprises of the number of COVID cases from John Hopkin's University. The secondary data sources used for this research study are listed in Table 3.

Table 3. Dataset used for the case study

Data	Time Span	Source
Tweets	01 August, 2020 to 31st August, 2020	**http://www.twitter.com**
COVID Reported Cases	01 August, 2020 to 31st August, 2020	**https://coronavirus.jhu.edu/data**

3.5. Sentiment Score

We also developed a new score card to score the sentiments and behaviour of Twitter by comparing with the actual cases reported. The sentiment score is a ratio based on the total number of actual cases reported (data taken from John Hopkins University) and total number of tweets tweeted by users. Sentiment score provides a picture of how people are expressing sentiments in Twitter as a social media. Higher score indicates higher involvement, and lower score indicates lower involvement.

4. Results and Discussion

Figure 2 shows the tweets distribution across different states of United States. New York state had the highest number ~ 40% of tweets compared with all other states, followed by New Jersey, Maryland and Washington DC. However, when compared with the number of tweets per unit population of each State, the Z-test which indicates whether the value is higher than the mean value showed only 5 states above the mean value. District of Columbia, followed by Delaware, New Jersey, Maryland, and New York exhibited the highest tweets when compared to the mean value (Table 4). New York, which ranked 1st in the total tweets, ranked 5th compared to the number of tweets per unit population. Of the 52 states in US, only 5 states had the number of tweets per unit population above the mean value. Further the tweets were more in the northern states when compared with the southern States.

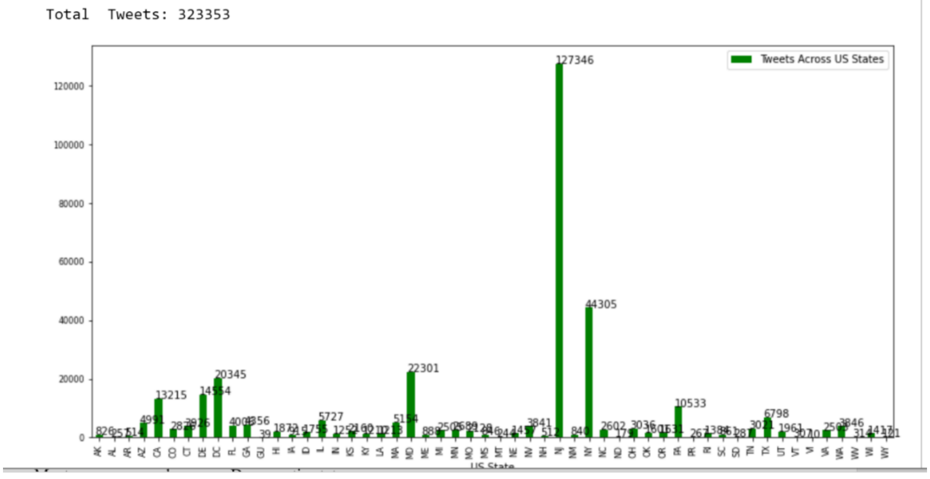


Figure 2. Tweets across different US States

Figure 3, shows that the sentiment analysis of the 'Symptoms' category. This graph shows the people who tweeted when they showed symptoms of COVID-19 or were isolated and quarantined or quarantined due to COVID-19 virus symptoms.

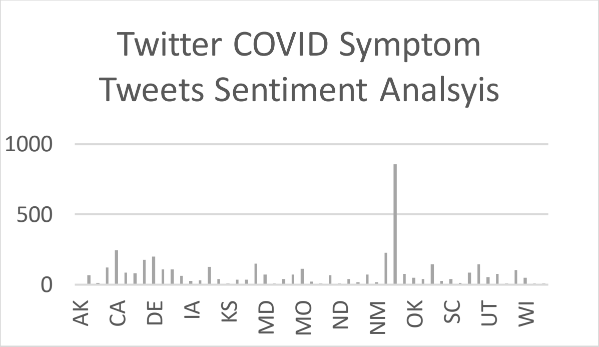


Figure 3. Twitter COVID Symptoms Sentiment Analysis

As we can see from the graph and Table 4, lists the highest number of people who tweeted about their symptoms during August 2020. Top 5 states are from New York, California, Nevada, Washighton DC, Massachusetts.

Table 4. Top 10 Twitter Symptoms Sentiment Tweets State

State	COVIDTweets	SymptomTweets
NY	27426	858
CA	10818	244
NV	3268	226
DE	3023	200
DC	6689	176
MA	4515	152
PA	5408	147
TX	6798	147
IL	4774	127
AZ	4579	122

Figure 4 shows the sentiment score, States of Maryland, New Jersey, Oregon shows the higher scores that means there is a higher correlation of the number of cases reported and people who are tweeting to bring the awareness to the community or to inform their friends and family about their situation.

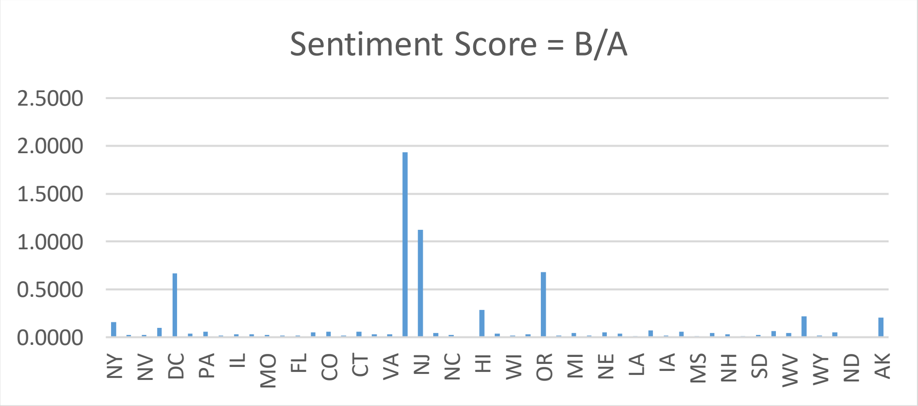
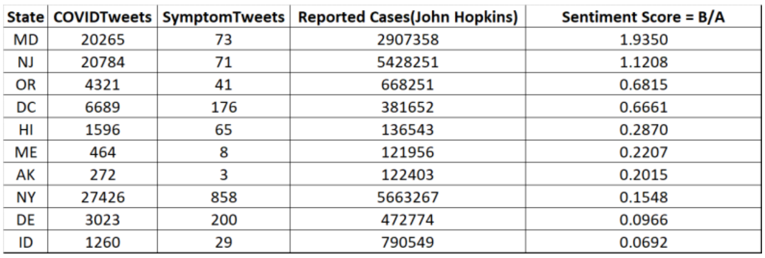


Figure 4. Twitter COVID-19 Symptoms Sentiment Score

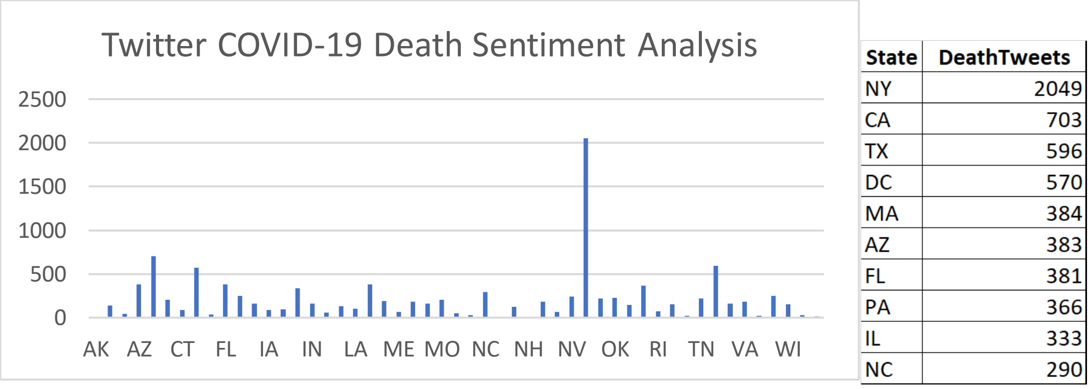
Table 4 lists the sentiment scores. States of Maryland, New Jersey, Oregon shows the higher scores compare to other States. This means there is a higher correlation of the number of cases reported and people who are tweeting.

Table 5. Twitter COVID-19 Symptoms Sentiment Score



State	COVIDTweets	SymptomTweets	Reported Cases(John Hopkins)	Sentiment Score = B/A
MD	20265	73	2907358	1.9350
NJ	20784	71	5428251	1.1208
OR	4321	41	668251	0.6815
DC	6689	176	381652	0.6661
HI	1596	65	136543	0.2870
ME	464	8	121956	0.2207
AK	272	3	122403	0.2015
NY	27426	858	5663267	0.1548
DE	3023	200	472774	0.0966
ID	1260	29	790549	0.0692

Similar analysis was performed for the case of death sentiment. This sentiment explains the number of twitter users expressed COVID death sentiment on twitter. These messages contain the news sentiment of someone they know in the family or friends died due to COVID.

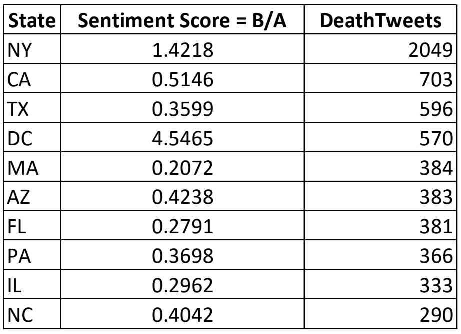


State	DeathTweets
NY	2049
CA	703
TX	596
DC	570
MA	384
AZ	383
FL	381
PA	366
IL	333
NC	290

Figure 5. Twitter COVID Death Sentiment Analysis

Figure 5 and Table 6 shows that the states of New York, California, Texas tops the list of tweets related to death sentiment.

Table 6. Twitter COVID-19 Symptoms Sentiment Score



State	Sentiment Score = B/A	DeathTweets
NY	1.4218	2049
CA	0.5146	703
TX	0.3599	596
DC	4.5465	570
MA	0.2072	384
AZ	0.4238	383
FL	0.2791	381
PA	0.3698	366
IL	0.2962	333
NC	0.4042	290

We also developed a similar score card to score the death sentiments. In this case, we took the reported deaths data from John Hopkins University and the total number of tweets tweeted. Sentiment score provides a picture of how people are expressing sentiments in Twitter social media. Higher score indicates higher involvement and lower score indicates lower involvement.

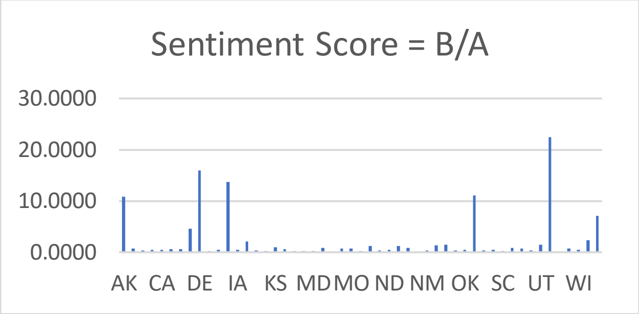


Figure 6. Twitter COVID-19 Symptoms Sentiment Score

Figure 6 and Table 7 shows the sentiment score for the twitter death sentiments, States of Varginia, Delaware, Hawai, Oregon, Arkansa shows the higher scores that means there is a higher correlation of the number of death cases reported and people who are tweeting to inform their friends and family about their loss.

Table 7. Twitter COVID-19 Death Sentiment Score

State	Sentiment Score = B/A	DeathTweets	ReportedDeaths(JOHN)
VA	22.3913	182	1678
DE	16.0045	33	17303
HI	13.6568	157	1188
OR	11.0805	149	11309
AK	10.8672	8	851
WY	7.1485	17	929
DC	4.5465	570	17265
WV	2.3443	30	4636
ID	2.1154	94	7984
UT	1.5030	160	10579

When the total cases were compared with the total COVID tweets, little or no correlation was found and when the number of reported deaths to the total COVID death tweets were compared there was a correlation of 0.40 as shown in figure 7.

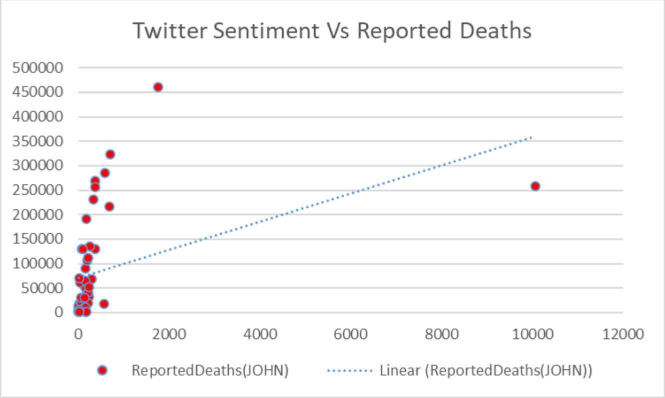
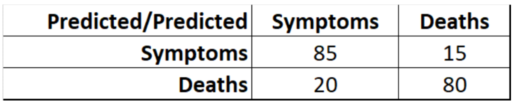


Figure 7. Correlation of Twitter Sentiments to the Reported Cases

4.1. Performance of the Sentiment Analyser Model

Though this is not a machine learning prediction or classifier, we still tested our unique sentiment analyser performance on the sample test data. The following truth Table 8 provides our testing performance with an accuracy of 83%.

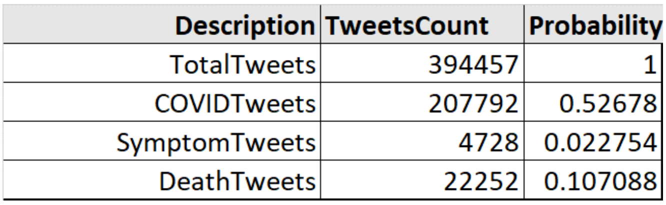
Table 8. Performance of the Unigram Sentiment Analyser



Predicted/Predicted	Symptoms	Deaths
Symptoms	85	15
Deaths	20	80

Table 9 summarizes our experiments and tweets sentiment statistics.

Table 9. Tweets Summary



Description	TweetsCount	Probability
TotalTweets	394457	1
COVIDTweets	207792	0.52678
SymptomTweets	4728	0.022754
DeathTweets	22252	0.107088

5. Future Work

The method used in this work is not the traditional machine learning classification or deep learning technique (Naïve bayes or Neural Network or NER). These conventional supervised machine learning techniques require training data with the classes labeled. Such data is not readily available unless someone manually labels the sentiment classes by looking at every tweet. This process is laborious and time consuming. If the sample data is small, this is possible. However, when we have lots of data, then this process can take lots of time. As a result, we spend more time on data preparation than solving research problems and finding some patterns in the data.

Hence, we used some of the classical NLP and exploratory data analysis model techniques to identify the sentiments using keywords to classify the sentiments. This work has several limitations. All these limitations are part of future work. Here are some limitations and future work under consideration:

1. We have considered only two classes of sentiments. Tweets may carry more categories, and this has to be explored.
2. Our tweets data collection during August 2021 month is based on several limitations such as language, region, cities, etc. Therefore, it just represents a small sample, not the entire Twitter population.
3. Applying supervised machine learning technique to predict twitter sentiments on the same or larger population of tweets. With our current method, data is already classified and this labeled data will be used as training data to create the new AI model.
4. Twitter has lots of slangs, spelling mistakes, grammatical errors. We can still improve data processing methods beyond what we have achieved in this work to enhance the model's efficiency.

6. Conclusion

In conclusion, tweets related to COVID demonstrates the two sentiments, one related to the deaths and the other related to the symptoms across all the USA states. The sentiments for each category vary from State to State. States of New York, California, Texas are higher tweets sentiments regarding expressing death sentiment, and states of New York, California, Nevada, are higher regarding sentiments of expressing COVID-19 symptoms. Our unique sentiment analyzer performed this prediction with an accuracy of 83%, and it can be further improved by overcoming certain limitations mentioned.

When we analysed the tweets sentiments expressed with actual reported cases and deaths, the sentiment scores are higher for the states of Maryland, New Jersey and Oregon for the Symptoms sentiment where as for the death sentiments, the states of Virginia, Delaware, and Hawai, showed higher sentiment scores indicating that the twitter users of these states are actively tweeting about symptoms and deaths even though the reported cases are less.

Twitter messages also indicate that people have tweeted more about the deaths than actual symptoms or isolations. The majority of the tweets are just related to COVID-19 and informational.

We also found no or little correlation between the Tweets and the number of cases across all the states. This is merely a case study and does not reflect the overall twitter behavior across the US states during the pandemic. Although some states had the higher number of tweets, it is in no way representative of the tweets over the entire period of the pandemic. Finally, the case study was carried out about the Twitter behaviour in the US for August 2020 only and is not a reflection of the entire pandemic.

References

1. Coronavirus Resource Center, August 2020, https://coronavirus.jhu.edu/map.html
2. Depoux, A., Martin, S., Karafillakis, E., Preet, R., Wilder-Smith, A., & Larson, H., "The pandemic of social media panic travels faster than the COVID-19 outbreak", *Journal of Travel Medicine*, Vol. 27, No.3, 2020
3. Kadam, Abhay B., and Sachin R. Atre. "Negative impact of social media panic during the COVID-19 outbreak in India.", *Journal of travel medicine, Vol.* 27. No.3, 2020.
4. Lee, D., Kim, H. S., & Kim, J. K., "The impact of online brand community type on consumer's community engagement behaviors: Consumer-created vs. marketer-created online brand community in online social-networking web sites", *Cyberpsychology, Behavior, and Social Networking*, Vol. 14 No. 1-2, pp. 59-63, 2011.
5. Liu, I.L., Cheung, C.M. and Lee, M.K., "Understanding Twitter usage: What drive people continue to tweet", *Pacific ASia Conference on Information Systems.* Taipei, Taiwan, 2010
6. Radwan, E. and Radwan A., "The spread of pandemic of Social Media Panic during the COVID-19 outbreak", *European Journal of Environment and Public Health, Vol. 4, No.* 2, pp 20-26, 2020.
7. Smith, A. N., Fischer, E., and Yongjian, C., "How does Brand-Related User-Generated Content Differ across YouTube, Facebook, and Twitter?", *Journal of Interactive Marketing*, Vol. 26, No. 2, pp. 102-113, 2912.
8. Twitter Data source, Retrieved from Twitter, *http://www.twitter.com*
9. Ye, Q., Zhang, Z. and Law, R., "Sentiment Classification of Online Reviews to Travel Destinations by Supervised Machine Learning Approaches", *Expert Systems with Applications*, Vol.36, No.3, pp.6527-6535, 2009.
10. Yu M, Li Z, Yu Z, He J, Zhou J., "Communication related health crisis on social media: a case of COVID-19 outbreak", *Current issues in tourism*, Vol.16, pp1-7, April 2020.

11. Zhao, Dejin, and Mary Beth Rosson, "How and why people Twitter: the role that micro-blogging plays in informal communication at work.", In *Proceedings of the ACM 2009 international conference on Supporting group work*, pp. 243-252. 2009.
12. Prabowo, R. and Thelwall, M., “Sentiment analysis: A combined approach”, *Journal of Informetrics*, Vol. 3, No. 2, pp.143-157, 2009.

AUTHORS

Prof. Umesh R. Hodeghatta, Ph.D

Dr. Umesh Hodeghatta Rao is an Engineer, a Scientist, and an Educator. He is a faculty member at Northeastern University, MA, USA, specializing in Analytics, AI, Machine Learning, Deep Learning, Natural Language Processing (NLP), Big Data Analytics and Cyber Security. He has more than 25 years of work experience in technical and senior management positions at AT&T Bell Laboratories, Cisco Systems, McAfee, and Wipro. He was also a faculty member at Kent State University, Kent, Ohio, and Xavier Institute of Management, Bhubaneswar, India. He has his master’s degree in Electrical and Computer Engineering (ECE) from Oklahoma State University, USA and Ph.D. from the Indian Institute of Technology (IIT), Kharagpur. His research interest is applying AI Machine Learning to strengthen an organization’s information security based on his expertise on Information Security and Machine Learning. As a Chief Data Scientist, he is helping business leaders to make decisions and recommendations linked to the organization's strategy and financial goals, reflecting an awareness of external dynamics based on data driven approach.

Dr. Hodeghatta has published many journal articles in international journals and conference proceedings. In addition, he has authored books titled "Business Analytics Using R: A Practical Approach" and “The InfoSec Handbook: An Introduction to Information Security” published by Springer Apress, USA. Dr. Hodeghatta has contributed his services to many professional organizations and regulatory bodies.

Sanath V Haritsa

Despite having a bachelor's degree in mechanical engineering, my keen interest for statistics and programming helped me change my career path. I have been studying machine learning and data science for the last 2 years through online as well as offline courses. Currently working for Nichesoft and Yokogawa Technology Solutions has helped me gain experience in fields of computer vision, performance optimization and fault detection in industries.

IMPACT OF BLOCKCHAIN TECHNOLOGY IN HEALTHCARE SECTOR DURING COVID-19 PANDEMIC

Zeba Mahmood

Software Engineering Department
Kaunas Technology University, Lithuania

ABSTRACT

Globally, the pandemic has affected management of risks. Progressively Blockchain is being applicable over the management of healthcare, as an imperative method for improving organizationalprotocols and for providing the convenient support for a productive and efficient decision-making process hinge on facts. In healthcare, different approaches to emergency preparedness can be recognized; indeed, each emergency is distinguished by different stages. In healthcare, we intend to role: explicitly, it will be responsible to enhance COVID19-safe clinical proceeding. The primary approaches obtainable from various blockchain-based models, and distinctly those associated by clinical individuals in the future throughout the current COVID-19 pandemic either on the would be capable to perform an outstanding assumption of furthermore infectious conditions. We believe that in real infectious disease outbreaks, blockchain technology undertaking, have been documented here and part in the future.

KEYWORDS

COVID–19, Blockchain, pandemic, healthcare.

1. INTRODUCTION

Blockchain is basically related to the broader group of Distributed Ledger technologies, the critically discussed. We have explored that blockchain can overcome the limitations of the existing system and thereby assist.substantiate blockchain and recommend a trace route on a side of a COVID19-safe clinical proceeding. In alliance along artificial intelligence systems, the adoption ofblockchain enables the development of a generalized predictive framework which can be conducive to pandemic risk constraints in the national territory. In future digital healthcare, blockchain may play a strategic. nodes. In the current sense of epidemic management, Blockchain has been emergingas an essential technical solution to provide aconsistent, reliable, and reduced solution to promote effective decision that might contribute substantially to faster interference during the same conflict. Blockchain is already exhibiting ample opportunity at becoming an essential part of the fight against COVID-19 because it would enable efficient monitoring and tracking solutions, make sure

operation of which is primarily based on a register organized in network-connected blocks; each transaction carried out in a network block is verified by a consensus- based mechanism distributed through all a consistent supply chain with vital donations and products, and safe payments. This really is feasible since this blockchain, a completely secured ledger systems

David C. Wyld et al. (Eds): AIMLA, DBDM, CCNET, NLTM, IOTBS - 2021
pp. 75-88, 2021. CS & IT - CSCP 2021 DOI: 10.5121/csit.2021.111306

transaction data provided among all network members, includes a sequentially orderly arrangement of cryptographic signatures. Furthermore, the adoption of blockchains and ledgers maximizes cost savings by withdrawal symptoms which mostly handle manual transaction records.

Innovative technologies such as blockchain, can help fight the critical situations. Blockchain technology, unusually, has thecapability to revolutionize differentindustries, in conjunction with supply chain, finance and the health division. Blockchain isdecentralized software along with the separate built-in features. This technology isa distributed ledger that consists of a block chain. Due to an inherent elementalcryptographic technology, which is typicallyused for participant's network authentication,the decentralized platform of blockchain is changeless. In addition, several resources are required to be capable of modifying transactions which are added facing the blockchain network, for the reason that evena single time transaction is verified and validated, it is restrained with an exclusive hash to previous transactions. In addition, all members of the network are made available with data stored on the blockchain, ensuring transparency between participants.

2. Background

The technology of blockchain is already extensively deployed to healthcare in recent years to improve operational protocols as well as to establish the appropriate base for a compelling decision-making process basedon facts. In the secure sharing of data between groups of people, Blockchain plays a strategic role, regardless of the cross- checking and reliability of these groups. It typically functions with the assistance of distributed tools, and it could be used with special attention to risk control in an advanced workflow or in enhanced protocols.In healthcare, we intent to verify blockchain along with the recommendation of a trace route for a COVID19-safe clinical practice inmore detail.

Blockchain is now a recently designedtechnology that allows transaction developers to allow financial transactions viapeer-to-peer (P2P) networks, withoutintermediate step entities, and to store transaction data in an organized ledger. By reason of the blockchain which stores data coming out of multiple individualssimultaneously, it is important to adjust data that is divided among the individuals simultaneously in order to modify the data. This makes it virtually impossible for the data to be forged or manipulated and to ensure its authenticity and accountability. The data which is stored in a blockchain is not lost, thus it is easy to monitor. Moreover, since therole of intermediaries is reduced, both financial and temporary expenditure savingscan be made. In different sectors, including banking, distribution, and manufacturing, internally has attempts to utilize blockchain, and its use in the medical sector is also beingstudied [1].Implemented to the healthcare sector, Blockchain will provide advanced andsuccessful ways to reinforce a range of pathology prevention and control activities and thus improve health risk management inthe context of a pandemic disaster, includingthe COVID-19 [2] The sudden arrival and uncontrolled spread of coronavirus worldwide clearly demonstrate not just the incompetence of current monitoring and surveillance to handle health care emergencies in a timely manner, as well as the apparent lack of predictive analytics systems for large-scale clinical data sharing capable of preventing but at least reducing emergency situations of this severity. In the health sector, various reports indicate all useof blockchain primarily for the exchange andimprovement of patient records, electronichealth records (EHR) as well as, while less common, supply chain management of medical systems or even medicines, drug prescription management, the improvement of clinical practice and the dissemination of scientific information, and also the development of clinical research. The emergence of modern as well as intelligent medical strategies had already initiated up latest possibilities for creative processes which had already been shown to work effectively or even safely [3]. Smart contracts based on blockchain technologies can also beused to automate auditing processes, boost supply chain management of pharmaceuticalproducts and monitor their safety or evencomply to existing regulations [4].

Moreover, latest IT infrastructural facilities don't really promote the facilitation of findings from scientific studies as well as the continuous exchange of clinical trials does not promote the creation and distribution of medical research capital. So Blockchain could be a viable tool for knowledge management that promotes the dissemination of improve medical practices and medicine based on evidence. By integrating blockchainand machine learning systems, which can beused to build predictive models that are useful in risk management, Blockchain isbelieved to be able to generate data: blockchain is built on innovations that have the tangible benefit of a distributed, intrinsic,and stable ledger, and security of patient privacy. Researchers have recently developed medical applications demanding its use of the internet: such applications have focused on artificial intelligence that has already been capable of facilitating prolonged machine learning in order to enhance crucial steps in the care and prognosis of multiple diseases [5].

Disintermediation, aimed as the nonappearance of a central authority which just collects, processes and validates the generated and exchanged data or models, allows the cost, time, and errors of process performance to be minimized in order to construct and update a predictive model that supports clinical practice and risk management of process performance. The blockchain is an automated framework that automates the processes involved in it and standardizes them. Furthermore, research into therapies that can combine early healingwith lower biological and economic costs has led researchers to experiment with smart materials and nanotechnologies, although the main challenges remain with regard to the safe application to human patients of such technologies [6].

Use of such technology as blockchain and its combined effect along with artificial intelligence systems allows a generalized predictive framework to be developed whichcould make a decisive contribution to just therestraint of pandemic threat in the territorial boundaries, including in the wider phase of risk management [7]. Furthermore, when the system responds to a cyberattack throughouta crisis, the need for a remote database couldultimately lead to greater damage, making it harder to identify modified data after justhacking. Though if blockchain have beenusing for epidemic monitoring systems, the data may be dynamically identified to that same final authority at a certain time as they are being processed with in blockchain, even without intermediate processing being carried out, improving the efficiency of the transmission of infectious disease outbreak data. In addition, because arbitrary editing of the results would be impossible, the situations of the pandemic would be clear and fully accessible to the target audience without interference [8].

The blockchain can prevent the dissemination of false information about infectious diseases. False data confuses people and can cause psychological anxiety and economic loss. Not only does the storingof reports and factual information on a blockchain network protect its alterations, it also helps make it identifiable, making itpossible to prevent the creation anddissemination of false information. Throughdiscarding the methods of printing anddistribution of a statement of diagnosis to both the actual clinic or hospital, blockchain could even help to mitigate the chances of infection through face-to-face interaction. When an insurance subscriber receives a premium, the identity of the subscriber can be determined, and the payment can be made regarding the hospital records recorded on the blockchain network [9].

Nevertheless, for frequent users of this type of technology [10] there are several certain things to remember related to the concept of privacy. Therefore, if, on the one hand, decentralization, and un-traceability, which are the common characteristics of the blockchain, enable the exact traceability andprotection of transactions, but at the other hand, an argument of conflict may arise withthe applicable legislation.

Moreover, the cryptographic method, the data immutability transmitted over the network as well as the unavailability of a central authority give rise to wider confidence in the system, since some need topreserve it disappears amongst its parties to the process. The parties' dedication to participating in the transmission and updatingof the temporary models is exemplified by a shared importance in achieving an ever moreprecise, practical and effective predictive model [11].

- **Identified Trust Conflicts with EstablishedInstitutions:**

Mediator foundations should give genuinely necessary, dependable, and solid administrations to society; in any case, the COVID-19 emergency has uncovered the restrictions of these organizations with regardto medical services. In this season of emergency, both public and private foundations just as conventional data frameworks have generally neglected to tackle issues identified with routine medical care conveyance, including accessibility of opportune information for projections of casenumbers, recognizable proof of high-hazard populaces, following contacts of people with COVID-19, and supply of individual defensive hardware or inventories of lifesaving drugs. Truth be told, it has been contended that the number of deaths because of COVID-19 might have been diminished with better admittance to solid information.

- **The United States and WorldwideCOVID-19 Pandemic:**

Before the end of July 2020, COVID-19 hadtainted around 19 million individualsworldwide and more than 700,000 deaths hasbeen identified. The United States is the mostextravagant nation on the planet, with a healthcare budget of US $3.5 trillion every year it has announced the most elevated number of individuals tainted with COVID- 19 (roughly 5 million) just as the most noteworthy number of deaths (>150,000) [12]. Through an absence of dependable information, powerlessness of medical services and general healthcare frameworks to perform dynamic observation, lacking administration of required clinical hardware, clashing data from various sources, and restricted innovation for commitment with patients, the COVID-19 pandemic has unmistakably shown the disappointment of existing establishments to secure human well-being and to dodge inescapable affliction.

3. METHODOLOGY

Transactions reflect the outcome of the activities that take place within the network between the topics. Via a cryptographic scheme, each block retains a connection to the previous one, hence the blockchain definition. Blockchain isn't really hosted on such a centralized server as with conventional web applications, yet are dispersed on network devices, maybe oneholding a copy of the blockchain. Two important elements that define this form of technology for our research are also useful to highlight. Due to the decentralization of consensus, the existence of trustworthiness and reliability between the researchers interested in any form of transaction as well as a centralized system would no longer be relevant. Similarly, in the second point, the persistence and storage through network nodes of various copies of various exchanges ensures greater system security and equity among users who can access the very same data efficiently, and thus the immutability and traceability of the verified transactions stored in the blocks. Thus, Blockchain is a peer-to-peer network in which all network members can trust the system without trusting each other. Blockchain adapted to the health sector will give advanced and successful ways to strengthen a range of pathology prevention and control activities and thus to improve the management of clinical risk in the sense of a pandemic crisis including the existing one. The abrupt presence as well as rapid and unregulated spread of Corona virus worldwide has shownus not only the inability of current health surveillance systems to manage public healthemergencies promptly, but also the obvious inadequacy of new predictive systems focused on the large-scale distribution of clinical data capable of preventing or at least minimizing emergencies of this magnitude. Blockchain can help with building a reliable and efficient framework (e.g., medical services) to battle the COVID'19 pandemic through verified, tested, circulated, and improve strong record innovation [13]. It canmake the first barrier to protect via a range ofinterconnected devices. In this segment, we discuss the expected strategies and use cases that blockchain innovation can offer to manage the COVID'19 pandemic.

Source: Blockchain for Covid19 895–9911(2020)

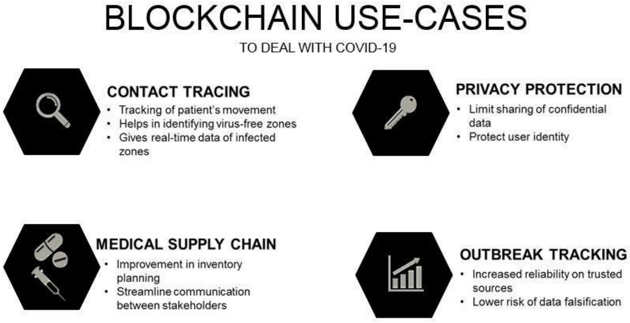

BLOCKCHAIN USE-CASES
TO DEAL WITH COVID-19
CONTACT TRACING
• Tracking of patient's movement
• Helps in identifying virus-free zones
• Gives real-time data of infected zones
PRIVACY PROTECTION
• Limit sharing of confidential data
• Protect user identity
MEDICAL SUPPLY CHAIN
• Improvement in inventory planning
• Streamline communication between stakeholders
OUTBREAK TRACKING
• Increased reliability on trusted sources
• Lower risk of data falsification

Figure 1. Block-chain Use cases with Benefits illustrate four major points During Covid-19 Pandemic (a) Contact tracing (b)Privacy Protection (c)Medical supply chain (d) Outbreak Tracking

To check the spread of COVID'19 requires fruitful and important inoculation of people against the infection through the organization of a functioning antibody. At the hour ofcomposing of this paper, many exploration foundations and research facilities are currently leading clinical preliminaries of a few immunization competitors. The adequacy, well-being, and authenticity of thevaccine are of extraordinary concern to the specialists, governments, and research foundations as the recently controlled immunization would antagonistically influence the health of an individual. Theexisting unified immunization the executive's structures face a couple oftroubles related to the risk of being failed to viably ensure about and scatter vaccines and breaking the coordination's stock organization of antibodies for noxious purposes [14]. Bogus medication associations consider this limitation of advancement as an event to sell, and flow fake a lot antibodies to fix COVID'19 patients. A fake, bogus, or insufficientvaccine is generally created using unsatisfactory matter. Utilizing poor manufacturing operations during the progression of vaccine moreover achieve unsuitable immunizations. The infiltration of fake, counterfeit, or unsuitable vaccinations into the dull market can hurt living spirits. For instance, in view of limited operational transparency, the enemies can adequately make inoculation pass or creation information during its shipment or retail to augment profit. Blockchain advancement can perpetually store data related to various phases, stages, and events of the COVID'19 antibody, for instance, (a) improvement, (b) creation, (c) certification, and (d) portion to endorsed relationship for immunization reason. In clinics/hospitals, clinical experts can get to blockchain to distinguish, follow, and check vaccines information prior to overseeing it. It can likewise be utilized for notification the executive's purposes (constant) through lightweight keen agreements. Brilliant agreements give occasions to identify vaccines related deceits,guarantee zero personal time, and disposes of the function of third-party administrations toscreen COVID'19 immunization coordination. The permanence highlight guarantees that the insights concerning the immunization can't be modified or erased by the enemies. Smart arrangements can recognize and check the expiry date of the inoculation in an accepted manner using records, for instance, the gathering date and assurance season of the immunizer [15]. Also, splendid arrangements can use provenance data to perceive the inadmissibleand falsified immunizations created and sentthrough unapproved creators. For store network collaborations organizations, sharp arrangements can be configured to screen the state of the compartment for temperature, clamminess, pressure, and different records to guarantee the COVID'19 immunization during its shipment. The sharp arrangements can therefore tell the significant experts when the preassigned conditions for the shipment are dismissed. The sensors can furthermore assist with recognizing any unlawful undertakings that may disturb the state of the groups passing on inoculations inside the conveyance holder. Any such development can be recorded, explored to screendisobedience, and notified constantly to the material authority. The various focal points of blockchain for coordination’s of future immunization for COVID'19 consolidate (a) exchange settlement, (b) review transparency, (c) precise costing data, (d) automation, (e) decreasing human blunders, and (f) implementing tariff and exchange approaches.

4. TRACING CONTACTS:

Blockchain empowers data to be gatheredfrom people without distinguishing them by utilizing an arrangement of public and privatekeys. For instance, the BeepTrace frameworkutilizes blockchain to give scrambled andanonymized individual ID while permitting controllers and medical care suppliers to contact individuals in danger of contamination because of contact with a tainted individual. The framework utilizestwo chains and a public key created by the public authority or a public element to produce area information yet additionally creates a diagnostician key to confirm test results[16].The contaminated individual offers agree to the diagnosing substance, which takes an interest in the blockchain to confirm results; notwithstanding, the public authority can't recognize the person. Noticescan be shipped off the individual utilizing a different chain.

Source: BeepTrace: Blockchain Enabled Privacy-Privacy Contact Tracing for Covid19. 3025953 (2020)

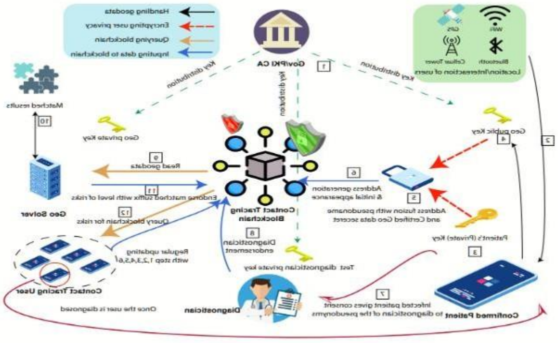

Handling geodata
Encrypting user privacy
Querying blockchain
Inputing data to blockchain
Gov/PKI CA
1
Key distribution
Key distribution
Key distribution
GPS
WiFi
Cellular Tower
Bluetooth
Location/Interaction of users
Matched results
10
Geo private Key
Geo public Key
4
2
9
Read geodata
6
Address generation & initial appearance
5
Address fusion with pseudoname and Certified Geo data secret
Patient's (Private) Key
3
11
Endorse matched suffix with level of risks
Geo Solver
12
Query blockchain for risks
Contact Tracing Blockchain
8
Diagnostician endorsement
Regular updating with step 1,2,3,4,5,6
Test diagnostician private key
Contact Tracing User
Once the user is diagnosed
Diagnostician
7
Infected patient gives consent to diagnostician of the pseudonyms
Confirmed Patient

Figure 2. Beep-Trace Framework Demonstrating the functionality ofContact tracing Techniques during Covid-19 Pandemic

Already, a similar protection and information sharing plan was likewise proposed in other blockchain-based applications. The key is that through anonymizing and cryptography, a blockchain-based contact following application guarantees singular protection while permitting general healthcare offices tocontact individuals who may have been presented to SARS-CoV-2, the infection that causes COVID19, through a tainted individual. These highlights of security, protection, trust, and effectiveness are incorporated into the engineering of blockchain and have been hard to reproduce or grow dependably in different applications. Nations, for example, Taiwan and South Korea have indicated that a robust arrangement of contact tracking can control the spread of disease while permitting ordinary life to proceed for healthyindividuals who are generally safe for contamination. However, worries about protection and security may restrict the usageof such techniques in various pieces of the world, especially in the United States, whichhas the most noteworthy quantities of cases and death rates [17]. Blockchain advances that empower people to share their own data in a safe way with general wellbeing organizations without uncovering their character or contributing that data to a unified government or corporate information basemay help distinguish individuals who come into contact with a patient who has triedpositive for COVID-19. This can be accomplished through general health offices or through distributed warnings, where just the positive status can be shared without sharing other clinical or individualinformation. The ability to follow people whoare positive for COVID-19 and to check theirseropositive status for contamination might be utilized as a vital apparatus to empower more capable resuming of the economy without causing a flood in cases. As we createimmunizations or create group insusceptibility for the disease, blockchaininnovation may likewise be utilized to give wellbeing accreditations that can be checked effectively by managers and general wellbeing organizations to approve the statusof a person.

5. CONVEYANCE OF REMOTE HEALTHCARE AND MEDICAL SUPPLIES

Utilizing progressed distant health operations, for example, telehealth and telemedicine administrations to limit the transmission danger of infectious infections can empower indicative patients to distantly speak with healthcare experts through IT framework. Remote diagnosis and treatment of patients can significantly limit quiet access and labor force restrictions, and along these lines the utilize capacity of distant healthcare administrations can viably control and breaking point the fast expansion in worldwide COVID19 cases [18]. Being administered and overseen by a concentratedpower, distant medical services frameworks are defenseless against a solitary purpose of failure issue, which eventually influences therespectability and dependability of the healthcare records. The characteristichighlights of progressive blockchain innovation can bring different benefits to the distant medical care industry. The essential benefits incorporate building up the provenance of electronic healthcare records, checking the authenticity of clients requesting quiet information, guaranteeing persistent anonymity, and mechanizing miniature installments for utilizing remote healthcare administrations. The tractability component helps effectively set up theprovenance of self-testing clinical packs for COVID19 testing. Following the testing result, people whose test outcomes areantagonistic are typically obliged to follow self-isolate approaches to relieve the spread of the infection to society. The necessities ofsecure track and hint of clinical supplies for self-isolated people achieve open doors for blockchain innovation to straightforwardly store time-stepped area information of clinical supplies on the ledger. Ensuring social separating and wearing face coversduring performing business exercises (e.g., pertinent medical services members) can helpwith controlling the spread of COVID19. The worldwide expanding COVID19 confirmed cases request con-tactless conveyance of medicines to the patients particularly in zones of high infection transmission rate to additionally forestall COVID19 fromspreading. For this reason, airborne vehicles can be utilized to ship medicines and clinical supplies to distant patients. Flying vehicles can likewise help with shipping clinical supplies among medical clinics that are housed at removed areas. For example, China tested (in 2020) utilizing flying vehicles to supply medications starting with one city then onto the next during the COVID19 pandemic [19]. Blockchain innovation canhelp to track and follow the area of the ethereal vehicles, verify provisioned administration level, and compute the standing score of an aeronautical vehicle dependent on its presentation in a trusted, responsible, and straightforward way. Through actualizing access control conventions and personality, the board, blockchain innovation limits the chance of assaults by the antagonistic vehicles. It permanently stores orders that are given to the airborne vehicles (for review purposes to confirm resistance with gave orders) by the control room alongside activities to purify the profoundly infection contaminated territoriesand distinguish human developments and collaborations. A multitude is involved numerous self-governing elevated vehicles that cooperate to accomplish a shared objective. Blockchain innovation can be utilized by the multitude of airborne vehiclesto arrive at an exceptionally solid worldwide choice by safely executing on the blockchain. For example, through a blockchain-based democratic framework, elevated vehicles of a multitude can distinguish the most thickly populated public spots to splash sanitization.

6. SYSTEM DESIGN AND IMPLEMENTATION

The execution of the stage would be a decentralized application (DApp) supporting a private blockchain network with an appropriated file framework (DFS) at the back end. Ethereum was utilized to present savvy contract structure for medical care blockchain. This is an open-source stage and right now one of the greatest public blockchain networks with a set up local area and an enormous assortment of public DApps. The stage presently utilizes an agreement evidence of-work (PoW) calculation called Ethash yet designs are attempting to transform it into a proof-of-stake (PoS) adaptability calculation in the short term. Preferably, for the plan of circulated applications, a Delegated Proof-of-Stake (DPoS) or Functional Byzantine Fault Tolerance (PBFT) agreement calculation is fit [20]. The DApp will possibly recognize inconsistencies, unapproved information additions and missing elements by coordinating DFS content with record registers. Each stage is marked with an Audit Timeline. The fundamental components of the brilliant agreements are capacities, occasions, state factors, and modifiers and are written in the robustness programming language. To pay the exchange charge, Remix and Kovan test network is utilized to send savvy contracts on the testnet and testnet ethers. Three phases are engaged with the advancement of brilliant agreements, which use Solidity programming to compose, aggregate, and report. The bytecode is made by the ongoing compiler Solidity. Ethereum Wallet has been utilized to unveil savvy Blockchain contracts. Since brilliant agreement programming started with Ethereum and Solidity, it is yet a controlunder development. Ethereum utilizes aspecific elliptic bend and set of numerical constants as characterized by the US NationalInstitute of Standards and Technology (NIST) standard called secp256k1 [20].Elliptic bend cryptography, or ECC, is a solid procedure to cryptography from a very notable RSA, and a developmental technique.By u sing the arithmetic behind elliptic bendsto set up protection between key sets is a strategy utilized for public key encryption. All through the previous few years, ECC has consistently expanded in notoriety because ofits capability to give a similar degree of insurance as RSA with a much lower key size. The accompanying capacity portrays the secp256k1 bend, which produces an elliptic bend:

p2= (q3 + 7) over (Fx) --- (1)
OR
p2 mod p = (q3 + 7) mod x --- (2)

The mod p (indivisible number p modulo) shows that this bend arrives at a limited primerequest p field, likewise, composed as Fx, where x = 2^256 – 2^32 – 2^9 – 2^8 – 2^7 –2^6 – 2^4 – 1, that is an exceptionally enormous number. Since this bend is characterized over a limited prime request field instead of over the genuine numbers, a two-dimensional example of spots, making it hard to envision. Picking a gathering ' P ' witha huge gathering cardinality or number, suggested by # Q and a huge 'r ' is significant, absolutely from a cryptographic perspective. While working in the spatial plane, on any smooth cubic bend, we can characterize a gathering structure. In the typical type of Weierstrass, such a bend will have an extra point at limitlessness, O, at the homogeneousdirections that work as the gathering's personality. Since the bend is balanced about the x-pivot, we can take −X to be the contrarypoint, given any point X. We're believing −Oto be O. In the event that X and Y are two focuses on the bend, at that point in the accompanying way we can portrayparticularly a third point, X + Y First, adhereto a meaningful boundary among X and Y. Ata third point, Z, this will ordinarily convergethe cubic. So, we take X + Y as −Z, the contrary incentive to Z as demonstrated in below.

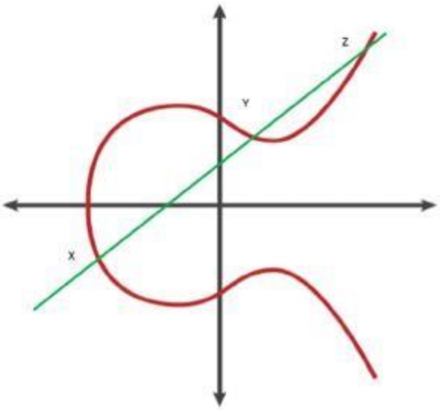


Figure 3. Graph of curves shows the expansion of X+Y+Z=0 where two values of X and Y adding third value Z for cubic bend coverage.

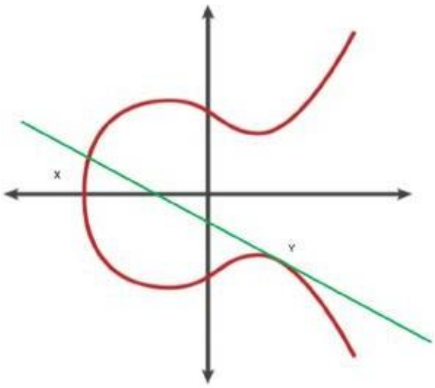


Figure 4. Graph of curves expansion focus for X+Y+Y=0.

This definition for expansion works aside from assortment of endlessness and convergence in a couple of uncommon cases identified with the point. The main point is when O is one of the focuses. Here we characterize the gathering's way of life as X + O= X = O + X. To begin with, we characterize X + Y = O if X and Y are inverseto one another. At last, on the off chance thatwe have just one point in X = Y, we can't depict the distance between them. For this situation, now we are utilizing the digressionline to the bend as our line. The digression crosses a second point Z by and large and wecan take the inverse. We can in any case portray a gathering structure for a cubic bend that isn't typical in Weierstrass by assigning one of its nine affectation focuses as character O. In the projective plane, when representing assortment, the line can converge a cubic at three focuses. −X is characterized as the remarkable third point onthe line that goes through O and X for a point X. Subsequently, X + Z is characterized as −R for any X and Y, where Z is the one of a kind third point on the X and Y segment. LetL alone a field that decides the bend (for example the coefficients of the characterizingcondition or bend conditions are in L) and mean the bend by M. So, M's L-levelheaded focuses are the focuses on M, the directions of which are all in L, including the limitlessness point M(L) indicates the arrangement of L-objective focuses. It likewise frames a gathering, since polynomial condition properties show that onthe off chance that X is in M(L), at that point −X is additionally in M(L), and on the off chance that two of X, Y, and Z are in M(L), at that point the third is the equivalent. Hence,in the event that L is a K subfield, at that pointM(L) is a M(K) subgroup. Chart of bends areoutlined in Figure 5 and Figure 6.

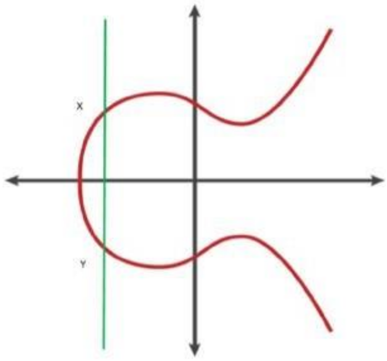


Figure 5. Graphical Expansion for providing curves identification for X+Y+0=0 where values of X+Y additionwith 0 giving a result of Summation.

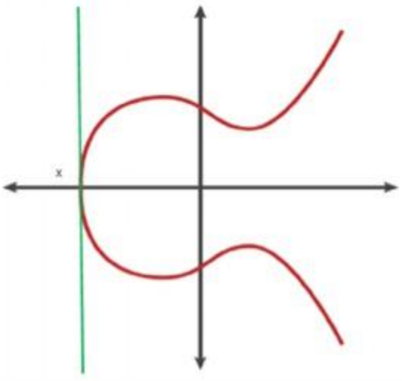


Figure 6. Graph of curves demonstrating the values ofgiven expression Y+Y+0=0.

The smart contact is converted into machine-level byte code where every byte characterizes each cycle, and afterward added as an EVM-1 exchange to the blockchain. A digger gets it, and affirms Block-1. At the point when a client sends theinquiry through the web interface, the EVM-2 inquiries and installs the online informationinto Transaction tx and conveys it to theblockchain. In Block-2 the exchange tx statusis changed. At the point when hub 3 choosesto test the states that are put away in the agreement, at that point it should synchronizethe progressions up to in any event Block - 2to see the progressions that exchange made.

Blockchain Based Smart Contract framework to control the transmission of COVID-19:

A smart contact, in light of blockchain innovation, might be fabricated and would have all the conditions from taking care of different consents to getting to information asfound in Figure 7 and it very well may be seen that an assortment of partners are partaking in this plan performing various exercises. It would assist with making more grounded doctor tolerant encounters. The standards controlling information authorization are coordinated into savvy contracts. It can likewise help screen all activities from their root to theiracquiescence, with interesting Id. Close by all the jobs and strategies, a savvy contract with different associations has been created and explained very much coordinated in the keen agreements. Figure 7 shows the capacityof savvy contracts with Ethereum, where for disentanglement the mining interaction is forgotten about. There will be no requirementfor an incorporated body to supervise and approve the interaction as it tends to be dealt with straightforwardly by means of the keen agreement that incredibly diminishes the administration cycle organization costs. The PC exchanges are enlisted with private key (patient or doctor) of the proprietor. The framework's square substance reflects information possession and access authorizations traded by individuals from a private distributed organization. Blockchain innovation underpins the utilization of savvy contracts which permit us to computerize and screen certain state advances, (for example, a difference in access rights or the ascent of another interaction record). We record persistent specialist connections on an Ethereum blockchain through savvy contracts which join a clinical record with survey rights and information recovery bearings (successfully information pointers) for outside worker activity to guarantee against control, we remember a cryptographic hash of the record for the blockchain to guarantee information security [21]. This organization will be associated with the nearby and worldwide information bases to guarantee satisfactory observing and regulation of the contamination.

Source: Use of Blockchain Technology to Curb Novel Coronavirus Disease (COVID-19) Transmission) 584226(2020)

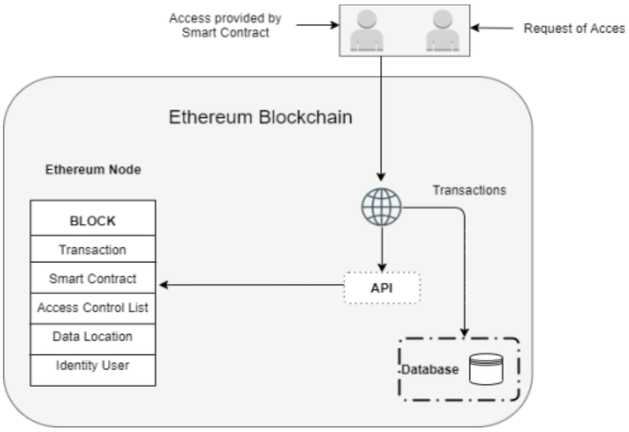


Figure 7. System Design and Workflow of Smart Contact

Wellbeing analysts need broad informational indexes to propel illness mindfulness, accelerate logical disclosure, quickly screen drug creation, and plan tolerant treatmentsystems dependent on science, lifecycle and climate. Through having patients of assorted ethnic and financial foundations and fromdifferent geographic districts, Blockchain'sshared information organization will incorporate a huge assortment of informational index [22]. It gives ideal information to longitudinal investigations on the grounds that blockchain accumulates information about an individual's wellbeing over a long period. A medical care blockchain can extend wellbeing information handling to incorporate information from gatherings of people by and by under-servedor not generally occupied with science. Blockchain's open information environment makes it simpler for "difficult to-reach" crowds to be included, and more sagacious for the overall population to convey results. The will likewise encourage the defeating of regulatory failures among patients, specialists and the medical services association. This framework will aid the recovery, audit and the executives of complex information and practices in the medical care area. The key goal is to share the data through keen blockchain decreases by empowering emergency clinics, doctors, crisis facilities and different accomplices to successfully access and trade the remedial data of a patient among various partners.

Existing Applications Based on Blockchain:

I. Blockchain to monitor data flow:

Getting quality data is more essential than ever. HashLog is a platform built on a blockchain that envisions the spreading of the COVID-19 pandemic continuously. To accomplish it, it works with information from various nations and relevant authority.Anybody can check the quantity of contaminated, expired, and recuperated patients across the world. MiPasa is a similar system, while supported by the World Health Organization and software giants such as Oracle or Microsoft. Considering IBM's HyperLedger Fabric blockchain arrangement, this groundbreaking project aims to enhance the insights available on theevolution of the COVID-19 [23]. As indicated by its supporters, it will permit the early recognition of COVID-19 patients and the basic virus focuses [24]. This will be accomplished by incorporating private and authority information, along with data frommedical clinics and medical servicesestablishments. The security of protection will be a basic component.

Source: Utilize blockchain-backed Covid-19 data with Mipasa , IBM, 2020.

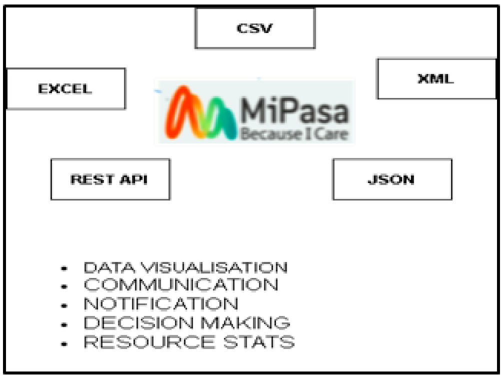


Figure 8. Utilized Blockchain COVID-19 data with MiPasa

II. Blockchain to oversee solutions andIdentities:

Another application is Prescrypto that deals with the stockpile of medication. This modern wave of clinical solutions ensures the character of patients while securing their individuality. Undoubtedly, imaginative advancements that diminish the weight on medical care frameworks will be a very welcome instrument in the battle against COVID-19.

Concluding Remarks and Further Recommendations

In this paper, we examined in detail how the arising blockchain innovation highlights andbenefits can be utilized for battling the COVID'19 pandemic. We investigated the potential blockchain applications from fundamentally the medical services crisis viewpoint to talk about the key job that blockchain can play during the COVID'19 pandemic. We identified the critical prerequisites of the partaking associations to create blockchain-based frameworks for medical care crisis administrations to battle the COVID19 pandemic. We examined existing blockchain-based frameworks thatare grown as of late to actualize assortedservices.

Our key findings and proposals include:

- The focal points of blockchain innovation regarding the significant trust, security, detectability, and reliability can incredibly help the specialists to devise answers for fight against the COVID'19 pandemic. For instance, immutable information identified with the flare-up of COVID19 in a city can be utilized by the specialists to effectively distinguish disease hotspots. Admittance to such critical data can help the specialists to define approaches for keeping theinfection from additional spreading.

- Execution of tracing contact arrangements incredibly relies upon the sum and speed of gathered data identified with area, travel history,and COVID19 test consequences of people. It is energetically suggested that the protection of the client's information should be safeguarded bythe contact tracing initiatives.
- Blockchain innovation is planned to give a helpful, responsible, and cooperative climate for members thatare engaged with the inventory network coordinations of vaccine. The selection pace of blockchain innovation by members incredibly relies upon the working transparency and affirmation of consistence with astandard to secure information against its abuse.

REFERENCES

1. Chang, M. C., Hsiao, M. Y., & Boudier-Revéret, M. (2020). Blockchain technology: efficiently managing medical information in thepain management field. *PainMedicine*, *21*(7), 1512-1513.
2. Ballini, A., Cantore, S., Scacco, S., Coletti, D., & Tatullo, M. (2018). Mesenchymal stem cells as promoters, enhancers, and playmakers of the translational regenerative medicine 2018. *Stem Cells International*, *2018*.
3. Angeles, R. (2019). Blockchain- based healthcare: Three successful proof-of-Concept pilots worth considering. *Journal of InternationalTechnology and Information Management*, *27*(3), 47-83.
4. Siyal, A. A., Junejo, A. Z., Zawish, M., Ahmed, K., Khalil, A., & Soursou, G. (2019). Applications of blockchain technology in medicine and healthcare: Challenges and future perspectives. *Cryptography*, *3*(1), 3.
5. Drescher, D. (2017). *Blockchain Basics–A Non-Technical Introduction in 25 Steps. California: Apress*. DOI 10.1007/978-1-4842- 2604-9.
6. Velavan, T. P., & Meyer, C. G. (2020). The COVID-19 epidemic. *Tropical medicine & international health*, *25*(3), 278.
7. McGhin, T., Choo, K. K. R., Liu, C. Z., & He, D. (2019). Blockchain in healthcare applications: Research challenges and opportunities. *Journal of Networkand Computer Applications*, *135*, 62-75.
8. Kalla, A., Hewa, T., Mishra, R. A., Ylianttila, M., & Liyanage, M. (2020). The role of blockchain to fight against COVID-19. *IEEE Engineering ManagementReview*, *48*(3), 85-96.
9. Zarour, M., Ansari, M. T. J., Alenezi, M., Sarkar, A. K., Faizan, M., Agrawal, A., ... & Khan, R. A. (2020). Evaluating the impact of blockchain models for secure and trustworthy electronic healthcare records. *IEEE Access*, *8*, 157959-157973.
10. Gajendran, N. (2020). Blockchain- Based secure framework for elearning during COVID-19. *Indian journal of science and technology*, *13*(12), 1328-1341.
11. Fusco, A., Dicuonzo, G., Dell'Atti, V., & Tatullo, M. (2020). Blockchain in Healthcare: Insights on COVID-19. *International Journal ofEnvironmental Research and PublicHealth*, *17*(19), 7167.
12. Bouncken, R. B., Kraus, S., & Roig- Tierno, N. (2019). Knowledge-and innovation-based business modelsfor future growth: digitalizedbusiness models and portfolio considerations. *Review of Managerial Science*, 1-14.
13. Taquet, Maxime, et al. "Bidirectional associations between COVID-19 and psychiatric disorder: retrospectivecohort studies of 62 354 COVID-19 cases in the USA." *The Lancet Psychiatry* 8.2 (2021): 130-140.
14. Torky, Mohamed, and Aboul EllaHassanien. "COVID-19 blockchain framework: innovative approach." *arXiv preprintarXiv:2004.06081* (2020).
15. Fischman Afori, Orit, Miriam Marcowitz-Bitton, and Emily Michiko Morris. "A Global Pandemic Remedy to Vaccine Nationalism." *Available at SSRN3829419* (2021).
16. Darwish, Saad, Anjali Mary Gomes, and Umair Ahmed. "RiskManagement Strategies And Impact On Sustainability: The Disruptive Effect Of Covid 19." *Academy of Strategic Management Journal* 20 (2021): 1-19.

17. Xu, H., Zhang, L., Onireti, O., Fang, Y., Buchanan, W. J., & Imran, M. A. (2020). Beeptrace: Blockchain- enabled privacy-preserving contact tracing for covid-19 pandemic and beyond. *IEEE Internet of ThingsJournal*, *8*(5), 3915-3929.
18. Malik, Y. S., Kumar, N., Sircar, S., Kaushik, R., Bhat, S., Dhama, K., ... & Singh, R. K. (2020). Coronavirus disease pandemic (COVID-19): challenges and a global perspective. *Pathogens*, *9*(7), 519.
19. Jnr, Bokolo Anthony. "Use of telemedicine and virtual care for remote treatment in response to COVID-19 pandemic." *Journal ofMedical Systems* 44, no. 7 (2020): 1-9.
20. Khatoon, A. (2020). Use of Blockchain Technology to Curb Novel Coronavirus Disease (COVID-19) Transmission. *SSRNElectronic Journal*.
21. Kalla A, Hewa T, Mishra RA, Ylianttila M, Liyanage M. The role of blockchain to fight against COVID-19. IEEE Engineering Management Review. 2020 Aug4;48(3):85-96.
22. Kassen M. Understanding decentralized civic engagement: Focus on peer-to-peer and blockchain-driven perspectives on e-participation. Technology in Society.2021 Aug 1;66:101650.
23. Ahmad RW, Salah K, Jayaraman R, Yaqoob I, Ellahham S, Omar M. Blockchain and COVID-19 pandemic: Applications and challenges. IEEE TechRxiv. 2020 Sep 17.
24. Kumar A, Gupta PK, Srivastava A. Areview of modern technologies for tackling COVID-19 pandemic. Diabetes & Metabolic Syndrome: Clinical Research & Reviews. 2020 Jul 1;14(4):569-73

AUTHOR INDEX