International Journal of Innovation Engineering and Science Research

Editorial Board

International Journal of Innovation Engineering and Science Research

Volume 5 ISSUE 3 July-August 2021

International Journal of Innovation Engineering and Science Research
(ISSN: 2581-4591)
www.ijiesr.com

Study of Plant Biodiversity and Density mapping using Geoinformatics

[1]Narueset Prasertsri [1]Satith Sangpradid [1]Thinnakon Angkahad [2]Ilada Aroonsri

[1]*Department of Geoinformatics, Faculty of Informatics, Mahasarakham University, Kantarawichai District, MahaSarakham 44150, Thailand;*

[2]*Department of Business Digital, Faculty of Management Sceinces, ValayaAlongkorn Rajabhat University Under the Royal Patronage, 1, Moo 10, Klong 1, Klong Luang, Pathum Thani 13180, Thailand*

ABSTRACT

The important of plant species diversity with the rich ecosystem to maintain the balance of the species. The objectives of this research 1)to classification the forest type using the satellite imagery. 2) to study the plant biodiversity of tree with measurements from the sampling plot using the importance value index (IVI) in the KhokKaow national forest. 3) to determine the plant density map of relative density (RD) using the geographic information system. The results of the forest type classified found that the closed deciduous forest showed the most of high density or high vegetation was the central region of the study area. The results of the IVI value found that the rank of three highest value was *Xyliaxylocarpa* (34.02), *EllipanthustomentosusKurz* (28.41), and *Dipterocarpus tuberculatusRoxb*. (23.9), respectively. The three lowest value of IVI was *CareyaarboreaRoxb.* (0.21), *Phyllanthus emblica L.* (0.22), and *Diospyros ebenum* (0.22), respectively. The results of plant density map found that high plant density found that the west and south region of the KhokKaow national forest.

Keywords—*plant species diversity; vegetation index; spatial interpolate;*

I. INTRODUCTION

In Thailand, the area of forest cover at 163,974.51 sq.km. in 2018-2019 or roughly 31.68% of total land area divided into forest area in the central region as 19,574.10 sq.km., northeastern region as 25,148.87 sq.km., eastern region as 7,559.27 sq.km., western region as 32,196.76 sq.km., southern region as 17,955.16 sq.km., and the northern region as 61,331.38 sq.km., respectively [1]. The northeastern Thailand located on the Khorat Plateau, bordered by the Mekong River (along the border with Laos) to the north and east, by Cambodia to the southeast and the Sankamphaeng Range south of Nakhon Ratchasima [2]. MahaSarakham province is located in the northeastern region of Thailand. The province is mostly a plain covered with rice fields. The total forest area is 213.76 sq.km. or 3.8 percent of provincial area [1]. KhokKaow national forest is the habitat and nurture of many kinds of animals both inside and outside the forest. KhokKaow national forest is the source of water, a rich source of food, source for learning and researching of natural, an eco-tourist attraction and a source of the local herbs [3]. An area of the KhokKaow national forest, MahaSarakham Province, with a total area of 8.62 sq.km.

The definition of species diversity is the number of species and the abundance of each species that live in a specific place. The number of species living in a place is called species richness. The important of species diversity in a complete ecosystem, a diverse and balanced number of species exist to maintain the balance of ecosystem. The relationship of an ecosystem, all the species depend on directly or indirectly. More productive and sustainable ecosystem, it is important to keeping high species diversity. The species diversity varies according to different geographical location with tropics having highest and declines as we move towards poles. Most species of rich environments include tropical rainforests, coral reefs and ocean bottom zone. Each species plays an important role in the ecosystem. Conservation of diversity is important because once it's extinct, we can't get it back.

Presently, the technique for instrumental in the assessment biodiversity to analyze landscape patterns and spatial relationships between the biodiversity called the geoinformatics technology. Satellite imagery is on of most technique for the classification of land used or land cover (LU/LC) to classify for represent the LU/LC in the study area. Usually, the classify form satellite image uses the multi-spectral images to classification of LU/LC. The normalized difference vegetation index (NDVI) is one of vegetation index with a calculated for the multi-spectral image between the Red band and the Near infrared band. The NDVI index can be representing parameter for identifying vegetation [4] and the NDVI values used for classify the different of land cover type in the study area [5]. In this paper, presented the classify of land cover of the forests type using NDVI index from the satellite image and using field surveying for data collection of the plant biodiversity. Furthermore, the estimating of density mapping of plant biodiversity using the geographic information systems (GIS) in the KhokKaow national forest, MahaSarakham province, Thailand.

II. Material and methods

A. Study area

The KhokKaow national forest the located cover ChuenChom district, MahaSarakham province, Thailand. The study are lies between 16.31º to 16.32º N Latitude and 103.06º to 103.10º E Longitude shown in Fig 1.

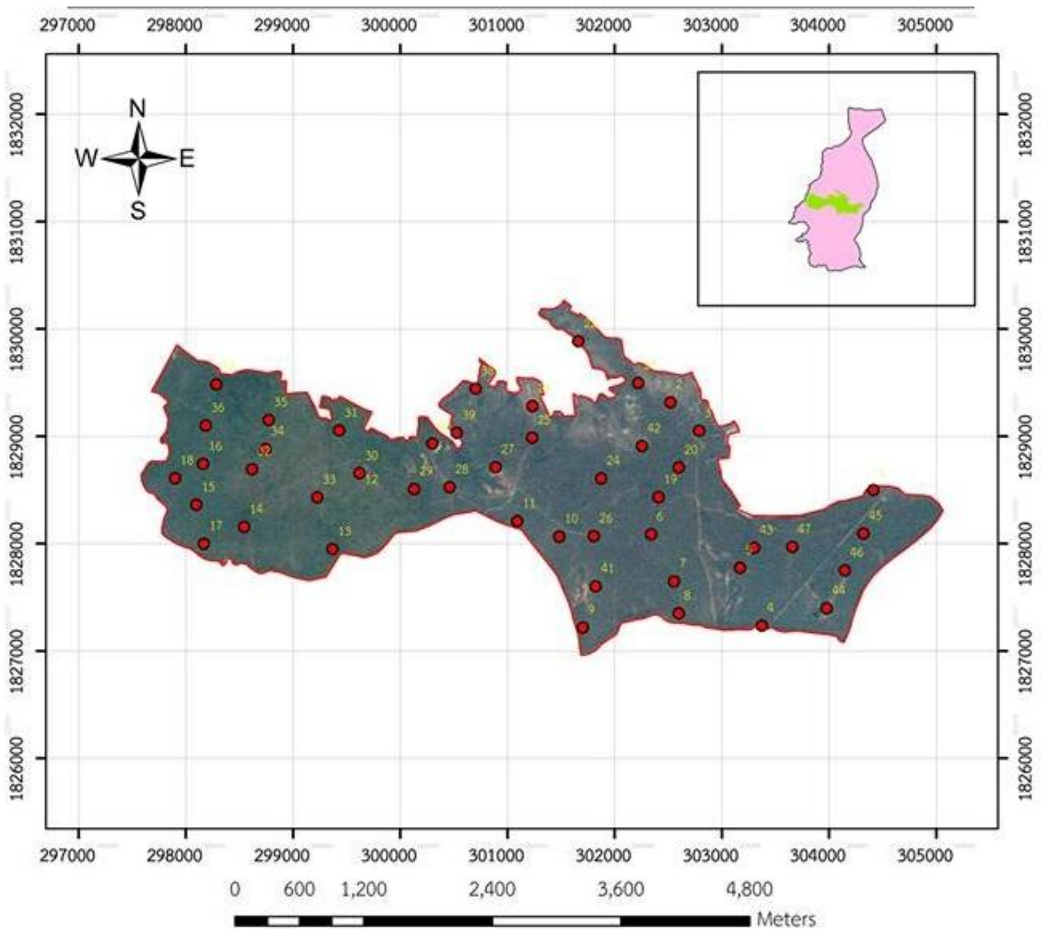


Fig 1. The study area, KhokKaow national forest,MahaSarakham, Thailand.

B. Image Classification from satellite imagery using NDVI

The normalized difference vegetation index (NDVI) was used to classification of forest type in the study area. NDVI was used in several application for classified land cover [6-8] and used for estimated the carbon sequestration of the tree [9-12]. The ability of NDVI can be separate healthy vegetation from other land cover types. In their original equations, they provide normalized values in the interval from −1 to 1. They show higher values for vegetation, positive low values for water and bare soils and negative index values for clouds. The NDVI is a normalized ratio of NIR (near infrared) and Red (red band) defined as [13]:

$$NDVI = \frac{NIR - Red}{NIR + Red}$$

C. Data Collection and Analysis the biodiversity of plants

1)Data collection and analysis the biodiversity of plants

Planning for observation of plant biodiversity in the study area used the sampling plot was the enumeration of randomly sample plots using the stratified random method [12]. The total of sampling plot was 47 plots by each plot has a dimension of 40 m. × 40 m. Data collection was conducted to all tree plant with the height of trees greater than 120 cm found in the observation plots. The attribute of data collection was record include: name species, name of trees, number, height, and diameter. The parameter for definition of the biodiversity indexes using the Shannon-Weiner index methods [13]. The parameter for analysis of important of the plant species by using the relative density (RD), relative domination (RDo), relative frequency (RF), and importance value index (IVI) [14-17].

2)Spatial interpolation using Inverse Distance Weighting (IDW)

Interpolation predicts values for cells in a raster from a limited number of sample data points. It can be used to predict unknown values for any geographic point data, such as elevation, rainfall, chemical concentrations, noise levels, and so on. Inverse Distance Weighted (IDW) is a method of interpolation that estimates cell values by averaging the values of sample data points in the neighborhood of each processing cell. The closer a point is to the center of the cell being estimated, the more influence, or weight, it has in the averaging process. This method assumes that the variable being mapped decreases in influence with distance from its sampled location [18].

D. Proposes Methods

This paper proposed the methods to evaluated the plant biodiversity using the technique of geoinformatics, the step following:

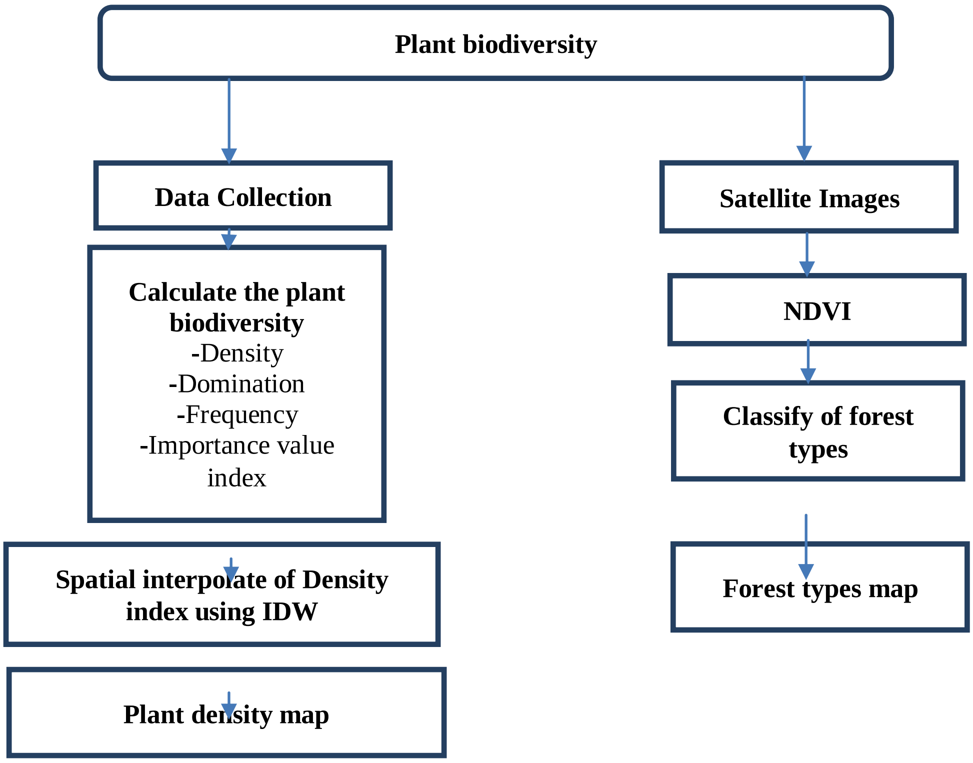


Fig 2. Proposes methods

III. Results

A. The result of land cover classification from satellite image using NDVI index

The results of forest type classified using the vegetation index namely NDVI from the satellite imagery was classified into 4 classes are the water bodies, open deciduous forest, dry deciduous forest, and closed deciduous forest the results shown in Fig 3:

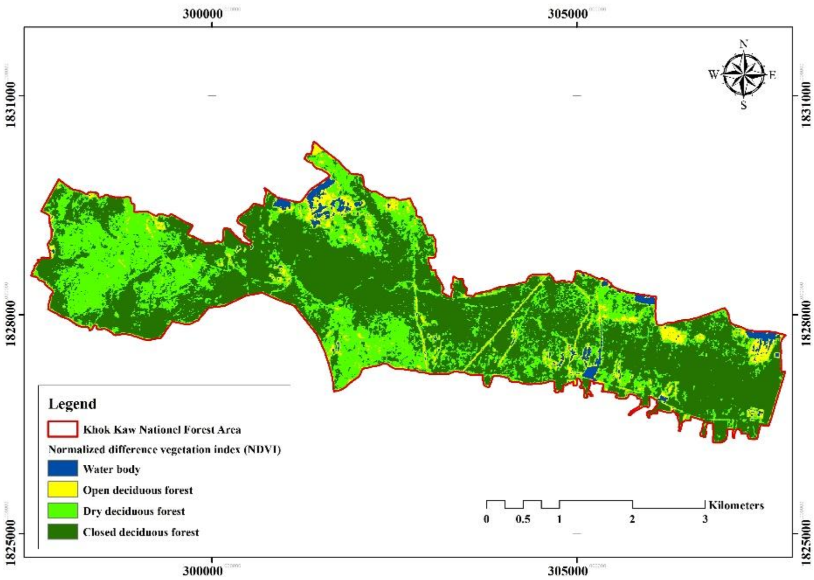


Fig 3. Forest type classification

B. The results of plant biodiversity and analysis of important index

The results of the data collected from the field surveyed sample plot with 47 plots found that the total number of 2,421 trees. The results revealed that the diversity of tree were 61 species. The results of the plant species by used the RD, RDo, RF, and IVI values show that the rank of highest value to lowest value by IVI shown in Table 1.

Table 1. The results of importance value index (IVI) value

ID	Name	Number of Tree	RD	RF	RDo	IVI
1	*Xyliaxylocarpa*	372	15.36	6.21	12.45	34.02
2	*Ellipanthustomentosus*Kurz.	270	11.15	6.35	10.91	28.41
3	*Dipterocarpus tuberculatus*Roxb.	220	9.09	3.32	11.49	23.9
4	*horeasiamensis*	149	6.15	4.47	6.81	17.44
5	*Buchananialatifolia*Roxb.	108	4.46	5.63	4.75	14.84
6	*Glutausitata* (Wall.) Ding Hou.	109	4.5	3.46	4.8	12.77
7	*Canariumsubulatum*Guillaumin	82	3.39	4.62	4.64	12.65
8	*Dipterocarpus Obtusifolius*Teijsm.exMiq.	114	4.71	2.45	4.15	11.31
9	*Shoreaobtusa* Wall.	90	3.72	3.61	3.61	10.93
10	*Pterocarpus macrocarpus*	67	2.77	3.75	3.64	10.16
…	…	…	…	…	…	…
55	*Diospyros filipendula* Pierre ex Lecomte	1	0.04	0.14	0.04	0.22
56	*Hymenopyramisparvifolia*Moldenke	1	0.04	0.14	0.04	0.22
57	*Terminalia alata*Heyne ex Roth	1	0.04	0.14	0.03	0.22
58	*Peltophorumpterocarpum* (DC.) K.Heyne	1	0.04	0.14	0.03	0.22
59	*Diospyros ebenum*	1	0.04	0.14	0.03	0.22
60	*Phyllanthus emblica* L.	1	0.04	0.14	0.03	0.22
61	*Careyaarborea*Roxb.	1	0.04	0.14	0.03	0.21
Tital		2,421	100	100	100	300

In the Table 1. found that the most of trees from the highest of the IVI value was *Xyliaxylocarpa* (34.02), *EllipanthustomentosusKurz* (28.41), and *Dipterocarpus tuberculatusRoxb.* (23.9), respectively. The lowest value of IVI was *CareyaarboreaRoxb.* (0.21), *Phyllanthus emblica L.* (0.22), and *Diospyrosebenum* (0.22), respectively.

C. The results Mapping of plant density using IDW

The result of forest density map from the relative density (RD) using the interpolated of IDW technique shown in Fig 4.

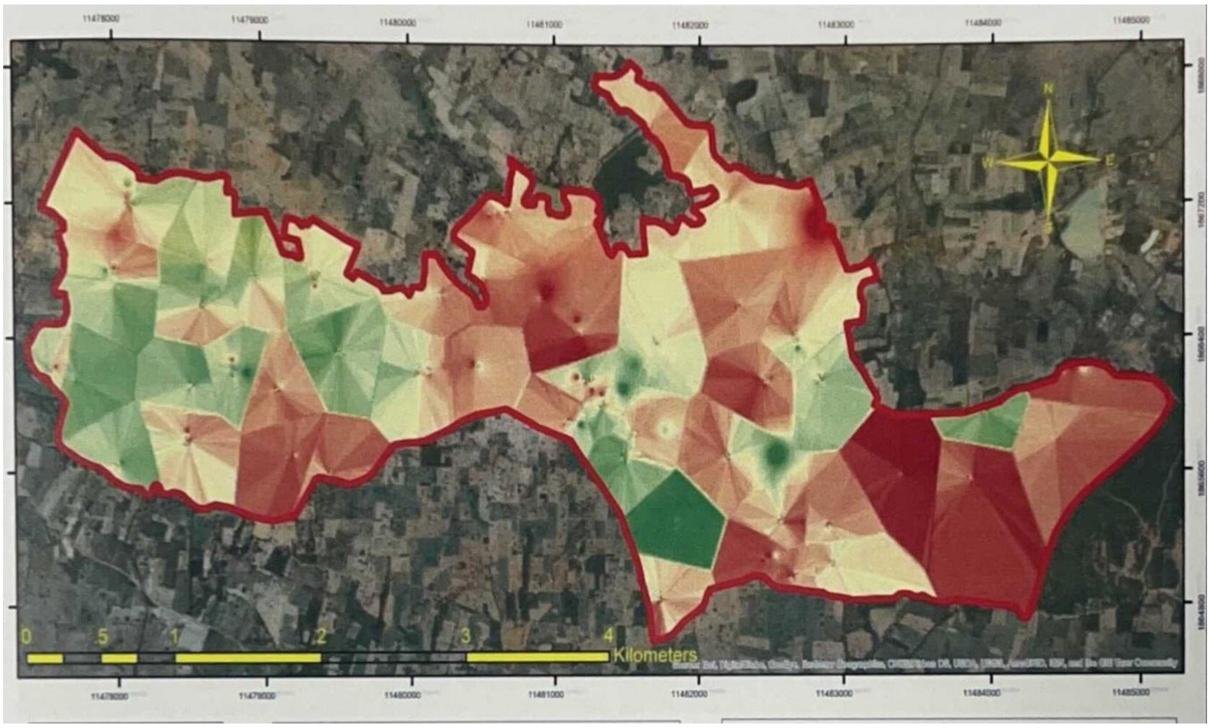


Fig 4. Plant density map from relative density using IDW

IV. Discussion and Conclusion

The experiment results of forest type classification by using the NDVI index shown in Figure 3 and percentage of an area for the forest type in the KhokKaownational forest show that the highest vegetation of forest (Closed Deciduous Forest) found that the central region of the KhokKaow national forest, and lowest vegetation forest (Open Deciduous Forest) found that the a little region of KhokKaow national forest.

The KhokKaow national forest found that the plants from 61 species and 31 families. The relative of plant biodiversity for rank of three highest value of IVI was *Xyliaxylocarpa*,*Ellipanthustomentosus*Kurz, and *Dipterocarpus tuberculatus*Roxb., respectively. The rank of three lowest value of IVI was *Careyaarborea*Roxb., *Phyllanthus emblica* L., and *Diospyrosebenum, respectively.* The relative density (RD) and relative domination (RDo) show rank of three highest value was *Xyliaxylocarpa*,*Ellipanthustomentosus*Kurz, and *Dipterocarpus tuberculatus*Roxb., respectively. The rank of three lowest value of RD and RDo was *Careyaarborea*Roxb., *Phyllanthus emblica* L., and *Diospyrosebenum,* respectively. The relative frequency (RF) show rank of three highest value was *Ellipanthustomentosus*Kurz*, Xyliaxylocarpa*, and *Buchananialatifolia*Roxb., respectively. The rank of three lowest value of RF was *Careyaarborea*Roxb., *Phyllanthus emblica* L., and *Diospyrosebenum,* respectively.

The results of the spatially interpolated plant density map shown in figure 4 found that the green color is high plant density found that the west and south region of the KhokKaow national forest. The high density of the map has a relationship with the sampling plot at plot numbers 3,11, 15, 17, 19, and 45.

Acknowledgment

This research project is financially supported by Faculty of Informatics, Mahasarakham University and Research unit of Geo-informatics for Local Development.

References

[1] Forest Land Management office. The implement information of forest area, Thailand. Final report in the Forest Land Management office. 2020.

[2] wiki pedia. Maha Sarakham Province. Available online: https://en.wikipedia.org/wiki/Maha_Sarakham_Province (accessed on 20 June 2021).

[3] Srimongkara, T. The report of natural attractions in Chuen Chom District, Maha Sarakham, Thailand. Available online: https://www.gotoknow.org/posts/630340 (accessed on 20 June 2021).

[4] Jeevalakshmi. D, S. Narayana Reddy, B. Manikiam. Land Cover Classification based on NDVI using LANDSAT8 Time Series: A Case Study Tirupati Region. International Conference on Communication and Signal Processing, April 6-8, 2016, India.

[5] Sangpradid, S. Change Vector Analysis using Integrated Vegetation Indices for Land Cover Change Detection. International Journal of Geoinformatics. 2018,14 (4), p.71-77.

[6] Amano, H., and Iwasaki, Y. Land Cover Classification by Integrating NDVI Time Series and GIS Data to Evaluate Water Circulation in Aso Caldera, Japan. International Journal of Environmental Research and Public Health. 2020, 17, p.6605.

[7] Sangpradid,S. Evaluate of THEOS Pan-sharpened Images using the Vegetation Indices for Orchard Classification. International Journal of Advance Research in Engineering, Science & Technology. 2020, 7(12), p.1-9.

[8] Taufik, A., Sakinah, S.A.S., Ahmad, A. Classification of Landsat 8 Satellite Data Using NDVI Thresholds. Journal of Telecommunication, Electronic and Computer Engineering. 2016, 8(4), p.37-40.

[9] Zhao, L., Zhang, P., Ma, X., and Pan, Z. Land Cover Information Extraction Based on Daily NDVI Time Series and Multiclassifier Combination. Mathematical Problems in Engineering. 2017.

[10] Uttha, T., Buasri, N., Prasertsri, N., and Sangpradid, S. Assessment of Above Ground Biomass in Phu Pha Wua Forest Park Kalasin Province, Thailand. International Journal of Innovation Engineering and Science Research. 2020, 4(6), 74-80.

[11] Laosuwan, T., and Uttaruk, Y. Estimating Above Ground Carbon Capture using Remote Sensing Technology in Small Scale Agroforestry Areas. Agriculture & Forestry. 2016, 62(2), p.253-262.

[12] Vicharnakorn, P., Shrestha, R.P., Nagai, M., Salam, A.P., and Kiratiprayoon, S. Carbon Stock Assessment using Remote Sensing and Forest Inventory Data in Savannakhet, Lao PDR. Remote Sensing. 2014, 6, p.5452-5479.

[13] Marod, D. Sampling technique and plant Community analysis. Available online: http://bioff.forest.ku.ac.th/main/?p=603 (accessed on 20 June 2021).

[14] Spellerberg, I.F., and Peter J. F. A tribute to Claude Shannon (1916–2001) and a plea for more rigorous use of species richness, species diversity and the 'Shannon–Wiener'Index. Global Ecology and Biogeography. 2003, 12(3), p.177-179.

[15] Prasertsri, N., Sangpradid, S., Buasri, N., Utta, T., Chusrithong, D., and Angkahad, T. Application of Geo-informatics for Survey the Ephiphytic plants in Tad Sung Waterfall Forest Park, Kalasin Province. Journal of Applied Informatics and Technology. 2021, 3(1), p.14-29.

[16] Prasertsri, N., Sangpradid, S., Buasri, N., Utta, T., Chusrithong, D., Angkahad, T., and Butaka, P. Application of the Geo-informatics to Survey GUTTIFERAE family plants in Khok Phak Kut and Pong Daeng National Forest, Maha Sarakham Province. Journal of Applied Informatics and Technology. 2021, 3(1), p.30-41.

[17] Tolangara, A., Ahmad, H., and Liline, S. The Composition and Important Value Index of Trees for Wildlife Feed in Bacan Island, South Halmahera. International Conference on Life Sciences and Technology : Earth and Environmental Science. 2019, p.1-7

[18] Ismail, M.H., AHMAD FUAD, M.F., HASSAN, Z.P., and NAIM JEMALI, N.j. Analysis of importance value index of unlogged and logged peat swamp forest in Nenasi Forest Reserve, Peninsular Malaysia. 2017, 7(2), p.74-78

[19] ESRI. How IDW works. Available online: https://desktop.arcgis.com/en/arcmap/10.3/tools/3d-analyst-toolbox/how-idw-works.htm (accessed on 20 June 2021).

International Journal of Innovation Engineering and Science Research
(ISSN: 2581-4591)
www.ijiesr.com

Particulate Triggers Spread of Covid-19, A Case Study of Industrial Clustered Region of Jharkhand, India

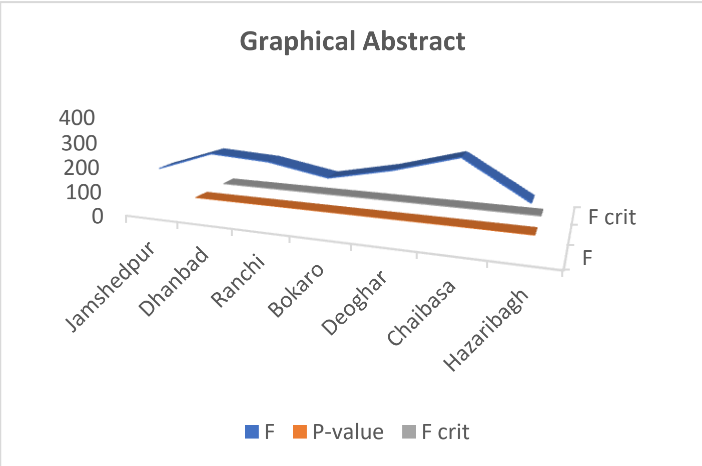
Shailendra Kumar Singh [1,2*], Ranjeet Kumar Singh [1,2], Krishna Kant Kumar Singh [1,2], Ranjeet Kumar Singh [3] & Siddharth Singh [1,2]

PARTICULATES TRIGGERS SPREAD OF COVID-19, JHARKHAND, INDIA

[1] *CSIR-Central Institute of Mining and Fuel Research, Dhanbad 826015, India*
[2] *Academy of Scientific and Industrial Research, Ghaziabad 201002, India*
[3] *BIT, Sindri, Dhanbad, 828123, India*

ABSTRACT

SARS-CoV-2 named as COVID-19 in talking terms, was first reported in the sea-food market of Wuhan township in China in Dec., 2019. Since then, SARS-CoV-2 has caused panic around the world and has been declared pandemic by WHO on March 11, 2020. The disease took the form of an epidemic in beginning 2020 causing widespread loss of lives and economic slowdown in the form of lockdowns with sudden burst still being reported. The root cause of the disease transmission was assigned to human-to-human contact. Later on, as investigations progressed around the world, the aerosol transmission was reported by many authors. Similar investigation was studied in seven most densely populated district of Jharkhand, India for nearly one month from November 4 to December 4, 2020. The statistical analysis of the data of PM_{10}, $PM_{2.5}$ and percent Covid cases by state authorities clearly established positive relationships in five of the seven districts with negative relationship in two districts at insignificant levels. Dhanbad, Ranchi, Bokaro, Deoghar, Chaibasa showed positive while Jamshedpur and Hazaribagh showed negative Pearson correlation. One-way ANOVA tests indicated that the P-value is less than the significance level of 0.05. In all the cases, the F value crossed the critical value (F crit) rejecting null hypothesis. Thus, the sample data provide strong enough evidence to conclude that there is relationship between the dust concentration to percent Covid cases occurring at a site with particulate characteristics supporting the cause.



Keywords: Corona virus; Dhanbad; Jharia Coal Field; Particulate matter; SARS-CoV-2

I. Introduction

Coronaviruses is a newly discovered strain of SARS-CoV-2 commonly referred as COVID-19. It causes respiratory disorder and has taken a pandemic dimension with reported deaths crossing 2,850,521 and infected more than 131,020,967 around the world. [1] The pandemic is likely to cause more deaths with increased intensity as being seen now. Moreover, Scientists across the world have formulated vaccine to cure the disease. Some vaccine has recently been permitted and is progressing around the globe. Partial success has been observed in many countries. Still, there is sharp upsurge of covid cases in some parts of the world and even in some states of

India. One of the major identified causes are the atmospheric air pollutants which is recently been endorsed by WHO. The present study refers to identify such precursors and their impacts on COVID-19 cases.

II. Study site

For identifying the impacts for the contribution of air pollutants, we have selected seven densely populated and Industrial agglomerated cities of Jharkhand dominated by mining industry particularly coal. Out of these, Jharia coalfields (JCF) located in the Dhanbad district is top-most polluted city in India (2019). The newly identified coronavirus, has caused a worldwide pandemic called COVID-19.

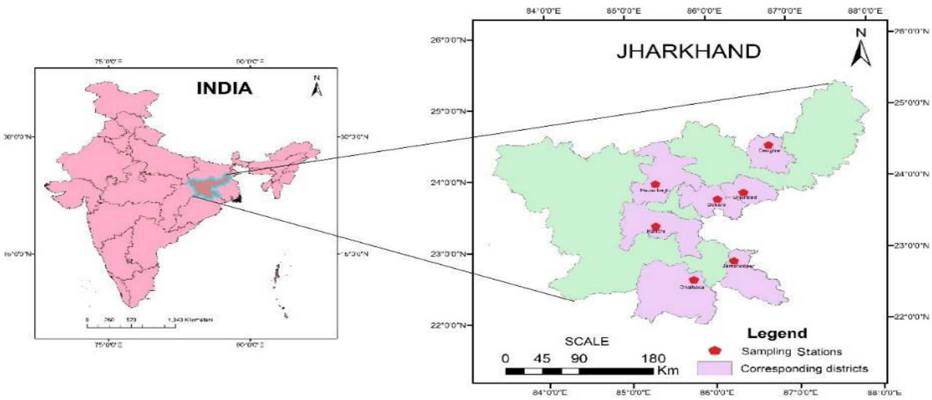


Fig. 1. Location of the study sites

Corona virus

Coronaviruses belong to a large group of viruses encircled with positive, single stranded RNA. Its classification is as follows:

Order- Nidovirales
Family- Coronaviridae
Subfamily- Coronavirinae

Twenty-six species are known in this group and have been divided into four genera- Alpha, Beta, Gamma and Delta. Till now alpha and beta strains are pathogenic to humans. [2]

Morphology, structure and replication

Coronaviruses is given this name because of crown (solar like) appearance. This appearance is attributed to spike peplomers of the glycoprotein emerging from lipid envelope. [3] There are two major proteins envelopes with a third glycoprotein called haemagglutinin-esterase (HE). The antigen that helps in receptor binding and cell fusion is S-glycoprotein [4] and transmembrane glycoprotein which is responsible for budding, crown formation and virion assembly. [5]

Coronaviruses are highly mutable to new hosts by recombinant technology of its single stranded genetic material. RNA-dependent RNA polymerase (RdRp) facilitates replication of the viral gene. Point to point mutation with other Coronavirus strain is very much reported with inherent error of 1,000,000 per mutation/site/replication of RNA-dependent RNA polymerase.

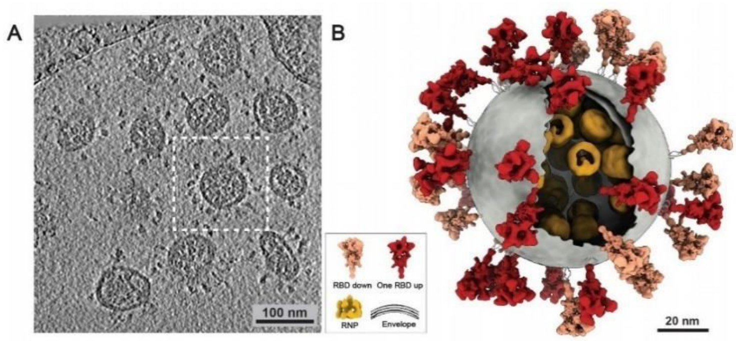


Fig. 2. Molecular architecture of the SARS-CoV-2 Virus[6]

Epidemiology, morbidity and mortality

Corona virus causes inflammation and burning in the upper region of respiratory tract along with similar viruses such as respiratory syncytial virus (RSV), parainfluenza, influenza and rhinoviruses. [3] The symptoms resemble those of common cold and Pneumonia. More than 15-30% all the common cold cases are caused by Coronaviruses. Gastroenteritis, body ache, dizziness, anxiety, fever, common ailments of cold and cough such as sneezing, oozing nose, head ache are the other symptoms Coronaviruses. They can also cause complex diseases in other animals too. [7,8]

Facts about 2019-nCoV

At the end of 2019, a new coronavirus was detected in Wuhan sea-food market of China commonly referred as novel corona virus. It was identified as a strain of beta coronavirus with sub-genous sarbecovirus having genomic similarity with SARS-CoVupto 80%.[9-11] Many more strains are being reported with more lethal affect being highly mutable such as British, Brazil and South African strains. Zoonotic transmission to bat and animals were initially known but later human to human transmission was also traced with an incubation period of 2-14 days. The observed symptoms are respiratory illness including cold and cough, mild fever, dyspnoea. Proof have been reported for the imperviousness in the two-pronged ground glass in the CT scan of the Chest. [12] Acute respiratory distress was seen in patients with severe illnesses requiring oxygen in ICU care.
The virus may pose lower threat to the individuals, but have a significant risk at the population level if transmitted easily. Asymptomatic patients are a global alarm. Given its pandemic potential, careful surveillance of nCoV is critical to monitor its future host adaption, viral evolution, infectivity, transmissibility and pathogenicity. [11]

Prevention

The Countries around the world are doing vaccination on a war footing to check the menace of Covid 19. But looking at the sudden upsurge of Covid 19 in some parts of the world, some Governments are enforcing stricter norms against nCoV with monetary charge and others. Therefore, physical means like isolation and quarantine are still enforced at the community levels to control the epidemics and Covid infections. [3,13] Asymptomatic patients with 2-14 (mean 4 days) days of incubation are the most potential risks in the case of SARS [14] and from 2-15 days in the case of MERS. [14] Centre for Disease Control and Prevention (CDC) recommends use of airborne infection isolation procedures in the care of all confirmed MERS infections in that country. [16]

Materials and methods

The values (Avg. of 24 hr) of $PM_{2.5}$ and PM_{10} were collected through the online beta system of www.accuweather.com [17]. The no. of covid cases during the period was collected through beta system of https://www.bing.com/covid/local/jharkhand_india. [18] Sample size of the period was the average of the estimated daily tests conducted available from the state agencies. The data marked as * is the average of the available data approximated to the nearest decimal as no data was available on that date. No data was attributed on 17th and 22nd Nov.,2020 due to festive holidays in India (Ref. suppl. file, table 1-5).

Results and discussion

PM_{10}

The values of PM_{10} (Fig 3) showed increasing trend from 4th Nov. to 4th Dec. except between 12-14th Nov. This can be attributed to the seasonal fluctuation with temperature and wind flow contributing the most. The top most polluting city contributing to the PM_{10} particulates is Hazaribagh and the lowest being Jamshedpur. The descending order is Hazaribagh> Chaibasa> Deoghar>Bokaro>Ranchi >Dhanbad >Jamshedpur. This pattern indicates the observed impact of wind flow and temperature.

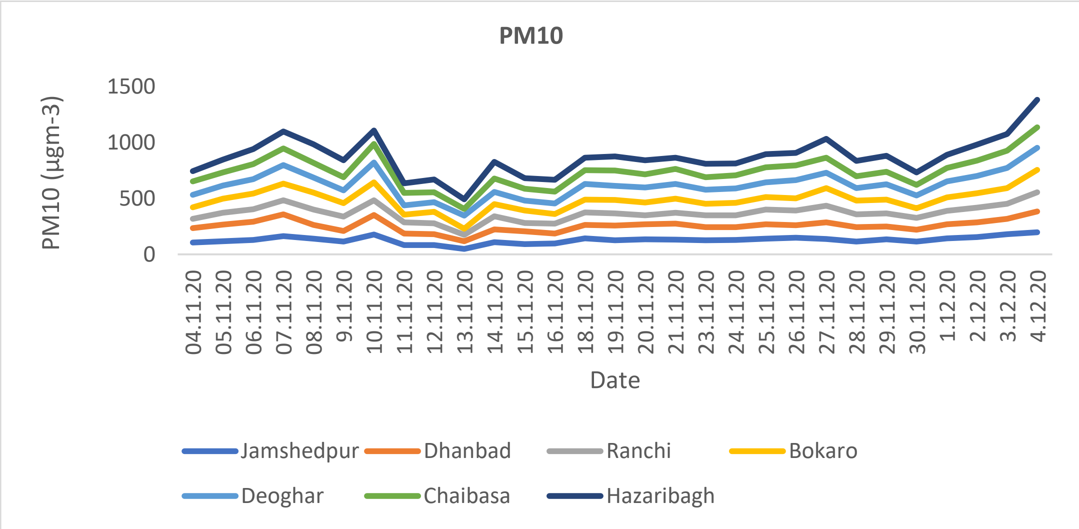


Fig. 3. Values of PM_{10} (μgm^{-3}) across the selected sites [17]

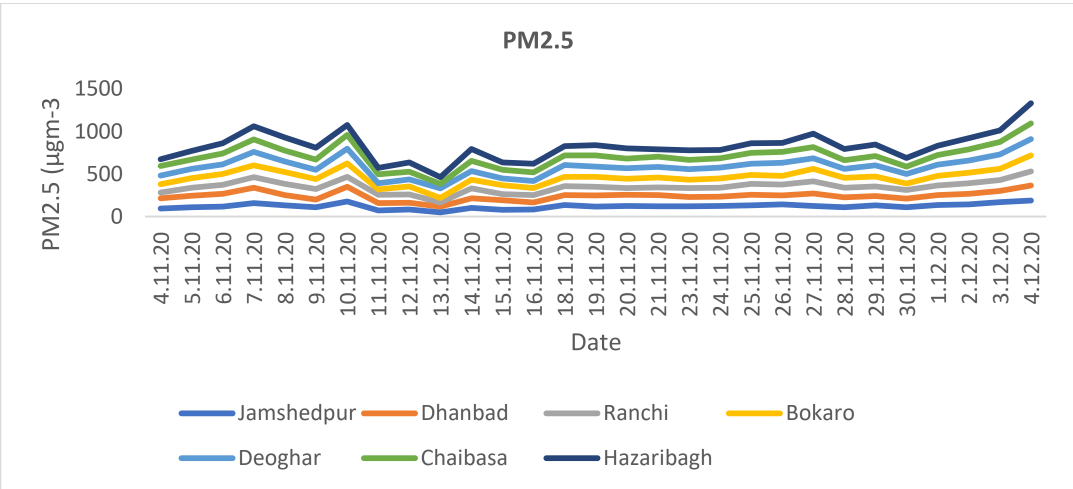


Fig. 4. Values of $PM_{2.5}$ ($\mu g\ m^{-3}$) across the selected sites [17]

$PM_{2.5}$

The values of $PM_{2.5}$ (Fig 4) showed increasing trend from 4th Nov. to 4th Dec., 2020. There was decreasing trend between between 12-14th Nov., 2020. Temperature differential along with moisture and wind direction attributed to this cause. The top most polluting city contributing to the $PM_{2.5}$ is Hazaribagh and the lowest being Jamshedpur. The observed descending order during the study period was Hazaribagh>Chaibasa>Deoghar>Bokaro>Ranchi>Dhanbad>Jamshedpur.Thescenario indicates seasonal gust with temperature, moisture and wind pattern playing the important role. It can be seen from the pattern of both the particulate sizes that the normal distribution trend amongst the selected sites remained almost the same during the monitoring period. There might be some fluctuation due to location of the sampling stations.

Covid cases (%) and particulate concentration

Except to metrological factors such as temperature, wind speed and direction, relative humidity whose affect is largely noticed on the emission and transport of dust particulates, the positive relationship is clearly observed between the dust concentration and covid cases (%) from 6th November onwards of the monitoring schedule (Fig 5). The relationship between particulate concentration and covid cases (%) is more explained by the statistical analysis. Lately many authors have advocated the role of air pollution in spread of Covid cases. Aerosol and surface contamination in quarantine centres was reported by Santarpia et al. (2020)[20] in Nebraska Medical Centre. Out of 13 samples collected, all the samples reported positive results of nCoV infection supporting the evidence that air particulates are a potential source of transmission of nCoV. Similar findings were reported by Liu et al. (2020)[21] in two Wuhan hospitals for SARS-CoV-2. The aerodynamic size analysis of RNA indicated that there was more concentrated occurrence on bathroom aerosol rather than isolation wards and patient rooms.

Setti et al. (2020) [22] reported similar findings on SARS-CoV-2. The studies reported by Harvard School of Public Health [23] also supported strong association between particulate concentration and the mortality rate due to Covid 19. Italian Society of Environmental Medicine [24] came to similar observation with Covid cases and particulate matter (PM_{10} and $PM_{2.5}$) in Northern Italy with specific climatic conditions.

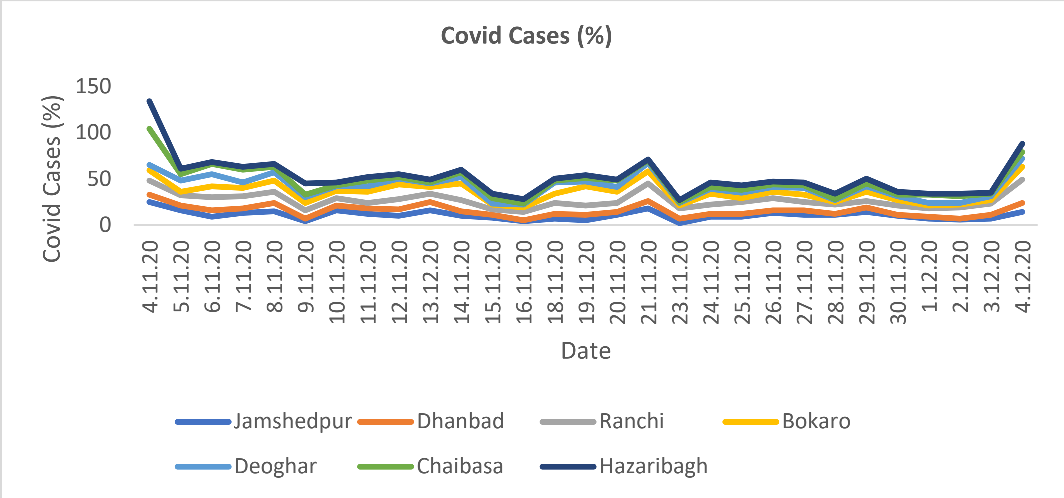


Fig. 5. Covid cases (%) (calculation based on no. of covid cases and sample size)

Anthropogenic activity has led to the increase in particulate concentration preferably <2.5μ called fine particulates in the ambient atmosphere. The nano particulates (1 to 100 nanometres in size) are still more dangerous because of their surface area and penetrating ability deep into the lungs. Ample proofs indicate that the size of the most of the viruses including nCoV possesses great affinity to this size range inflaming the alveoli of the lungs disrupting O_2 exchange in the haemoglobin. The blood has 100 times more affinity to CO_2 rather than to O_2. Thus, the concentration of carboxy haemoglobin increases in the blood causing fatigue and breathlessness and ultimately death. A sudden collapse in most of the Covid patients can be attributed to this cause. Interaction of the primary aerosol with other pollutants such as volatile organic compounds (VOC's), SO_2, O_3, NOx, CO, NH_3 along with toxic heavy metals such Hg, Cr, Cd, As, Pb etc. [25] and viruses forms secondary and tertiary compounds in the presence of moisture. These reactions are mostly temperature dependent as has been observed in most of the Covid epidemic centres around the world. An environmental genomics study [26] of the smog in Beijing, China indicated presence (0.1%) of numerous pathogens including viruses in PM_{10} and $PM_{2.5}$ particulates (Fig 6). Their concentration in the particulates is directly proportional to the ambient pollution load and rise in the Covid cases of the region. Increased viral concentration causes severe inflammation in the lungs [26] leading to sudden collapse.

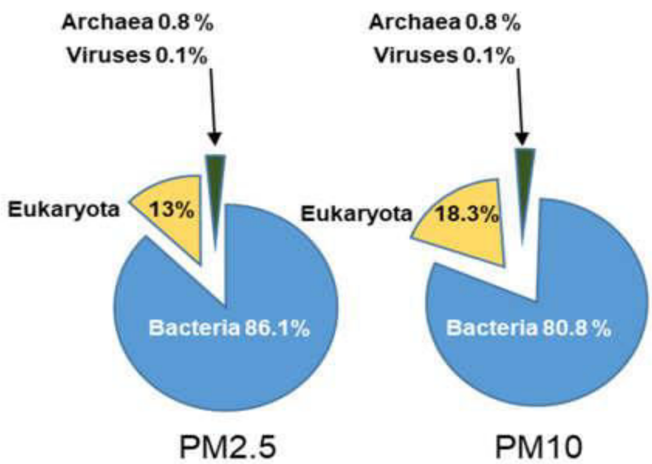


Fig 6. Percent representation of various organisms in PM_{10} and $PM_{2.5}$ in the Bejing, China [26]

Particle characteristics

Absorption and adhesion

Viruses and other microbes get stick to the particles through their binding properties such as absorption and adhesion. Microbe-particle complex formation is reported by Wei et al., 2020.[27] These complexes were observed in H5N1 (avian influenza virus) [28], RSV. [26] Bianconi et al. (2020) [29] related Covid cases and deaths to past exposure to aerosol particulates (PM_{10} and $PM_{2.5}$). Regression analysis depicted strong positive relationship between particulates (PM_{10} and $PM_{2.5}$) and Covid cases and deaths reported in the Italian city of Lombardi and Veneto. [29] Air pollution links to Covid cases and deaths were also established by Frontera et al. (2020).[30] Wang et al. (2020)[31] reported SARS-CoV-2 transmission through Flügge droplets. Particles <5 μm can remain for a long time and can spread to longer distances. [32] Most important aspect is exposed surface area to weight ratio that permits larger adsorption per unit weight. As the particle size becomes more and more smaller the exposure surface per unit mass increases thereby increasing adsorption efficiency creating acute toxicity in the cell. [33]

Particle Size

The particles of solid and/or liquid nature that remain suspended in the atmosphere for at least 1 hr with minimum stability are called aerosols (Fig 7).[34] It is considered as the most important part of particle character seeing its behaviour. It affects removal process, residence time and scattering hindering visibility. Their origin is traced to natural or anthropogenic sources ranging from few nanometres to several microns. Smaller and finer particles undergo coalition and adhesion with several chemical reactions to form bigger particles of secondary and tertiary nature from $<10^{-3}$ to 100 μm. [35] The particles greater than few hundred microns are not considered aerosols as they settle very fast. [34] These aerosols, though present in lower amount, play an important role in physicochemical reactions of the atmosphere such as electrical conductivity, scattering of the solar radiation, compression of water vapor over nuclei forming fog and smog. The climatic alterations may support specific microbial population to grow. The details of the reactions of the particle size chemistry are depicted in Fig 7.[34]

The diameter in PM_{10} particulates in the industrial agglomeration of Jharkhand specially in coal mining areas varies from 0.0525-0.598 μm. [36] This particle size is even smaller than 1 μm and is carried away to longer distances by wind due to its lighter weight. They tend to coalesce with liquid and gaseous aerosol along with viruses and bacteria to form complex secondary and tertiary aerosols. Recent studies citing air particulates as the precursor of bacterial transmission have been reported by many authors. [37] Reports indicate that these are highly dependent on size and concentration of the particulates. Particle size and their count are an important parameter accelerating viral and bacterial coalescence. Lighter and smaller aerosol particles (<0.1 μm) coalesce much faster than the larger ones. [33]

The aerosol and its behaviour in the environment with relative humidity and moisture was described by Hanel in 1976 (Fig 9). [38] Particle properties such as reactivity, solubility, charging capacity, polarity/or hydrophobicity, aspect ratio, binding state, and the interactive potential with other tissues generating reactive oxygen affect the level of particle toxicity. [33]

Industrial dust particulates are generated through anthropogenic means like (i) vehicular transportation and road dust (ii) industrial processes like coal and fuel burning (iii) coal based ancillary industries (iv) metallurgical processing (v) waste burning. These particulates serve as a potential precursor for virus transmission through binding forces described herein. Over the last decade, particulates <2.5 μm are more emphasised due to larger surface area and its penetration ability. Elemental carbon released during coal burning are very tiny (<2 μm) coalesce rapidly even during burning. The processes of physicochemical nature occurring during burning of pulverized coal (70 μm) was depicted by Okazaki, 1993[39] (Fig 8). Two pathways (Fig 8) are broadly followed - (i) swelling and (ii) shrinking. In the first case, numerous particles from 0.5 to 30 μm are generated through fragmentation and swelling. In the second case, numerous particles ranging from 0.02 to 0.2 μm are generated through vaporisation-condensation-nucleation-coagulation-coalescence processes of shrinking. The process may also generate particulates of 10-90 μm size through clustering, expansion and slaking. Therefore, very high gas to particle formation occurs. This gives a wide scope of microbial association which has been proved by many authors.

Particle properties such as reactivity, solubility, charging capacity, polarity/or hydrophobicity, aspect ratio, binding state, and the interactive potential with other tissues generating reactive oxygen affect the level of particle toxicity. [33]

Industrial dust particulates are generated through anthropogenic means like (i) vehicular transportation and road dust (ii) industrial processes like coal and fuel burning (iii) coal based ancillary industries (iv) metallurgical processing (v) waste burning. These particulates serve as a potential precursor for virus transmission through binding forces described herein. Over the last decade, particulates <2.5 μm are more emphasised due to larger surface area and its penetration ability. Elemental carbon released during coal burning are very tiny (<2 μm) coalesce rapidly even during burning. The processes of physicochemical nature occurring during burning of pulverized coal (70 μm) was depicted by Okazaki, 1993 [39] (Fig 9). Two pathways (Fig 8) are broadly followed - (i) swelling and (ii) shrinking.

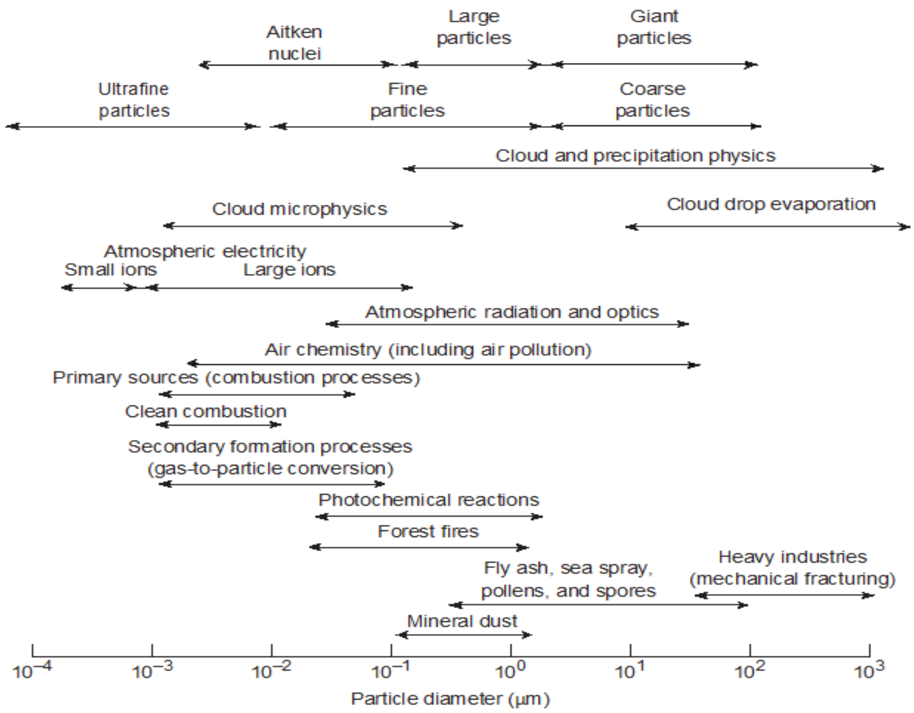


Fig 7. Size range of aerosols in the atmosphere & their role in atmospheric physics & chemistry [34]

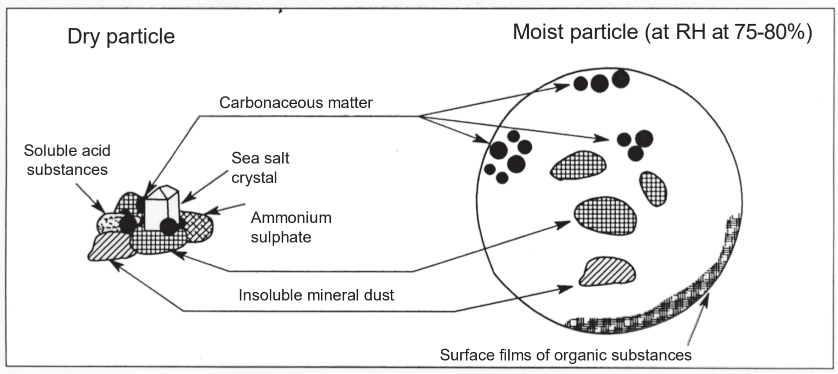


Fig 8. Sketch of an aerosol particle for dry air (left) and humid air (RH=75-80%) (right), with soluble (acid, sea-salt, ammonium sulphate) and insoluble (carbonaceous, mineral dust, organics) substances suspended inside the moist particle steadily increasing by condensation until the formation of a water droplet with soluble salts, acids, organic compounds etc. [38]

In the first case, numerous particles from 0.5 to 30 μm are generated through fragmentation and swelling. In the second case, numerous particles ranging from 0.02 to 0.2 μm are generated through vaporisation-condensation-nucleation-coagulation-coalescence processes of shrinking. The process may also generate particulates of 10-90 μm size through clustering, expansion and slaking. Therefore, very high gas to particle formation occurs. This gives a wide scope of microbial association which has been proved by many authors.

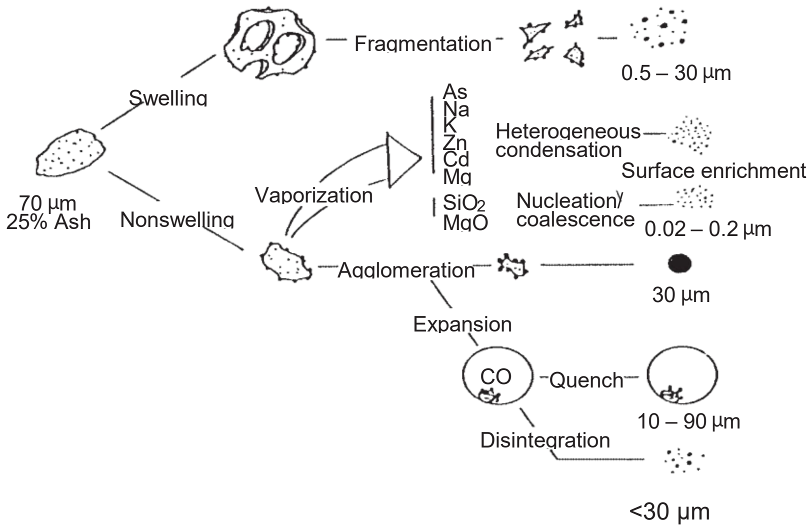


Fig 9. Particle formation during combustion of pulverized coal through different combinations of swelling, shrinking, fragmentation, vaporization, condensation, nucleation, coagulation, expansion, quenching, and disintegration processes. [39]

Particulate characteristics of JCF region

SEM images of coarse particulates (PM_{10}) of coal mining area indicates formation of complex aggregates consisting unburnt carbon and metal associates (Fig 10). They settle very easily being larger (>10 µm) and heavier. SEM images of $PM_{2.5}$ particulates reveal completeness of the properties of burning having spherical shapes. The unburnt carbon in the form of soot with embedded fly ash particles are spongy and glassy in nature. [40] It contains carbon rich soot aggregate with spherical as fly ash and elongated as mineral (metal silicates). Sources indicate the possibility of unburnt carbon from coal ignition which may harbor deadly microbes including viruses (Fig. 10b). Particles of the coal mining area of JCF possesses (i) aluminosilicates with diverse components (ii) carbonaceous matter (iii) metallics with iron oxides such as magnetite, hematite, and maghemite (iv) lime particles (v) minerals with quartz and mullite (vi) formless particles with irregular shapes, black or brownish in color. Remarkable changes from irregular fractured to compact and spherical shapes is observed (Fig 10) with decreasing size and scattering ratio in the soot. [41] The organic carbon present in the particulates with polycyclic aromatic hydrocarbons (PAH) may harbor microbes possessing mutagenic properties. [35] The non-coal mining sites ($PM_{2.5}$) contain fluffy soot aggregate containing carbon and metallics with high % of silica, a potential source of silicosis. Sources indicate construction activities. The finer particles (<2.5µm) obstruct CO_2 and O_2 exchange in alveoli causing breathlessness and death.

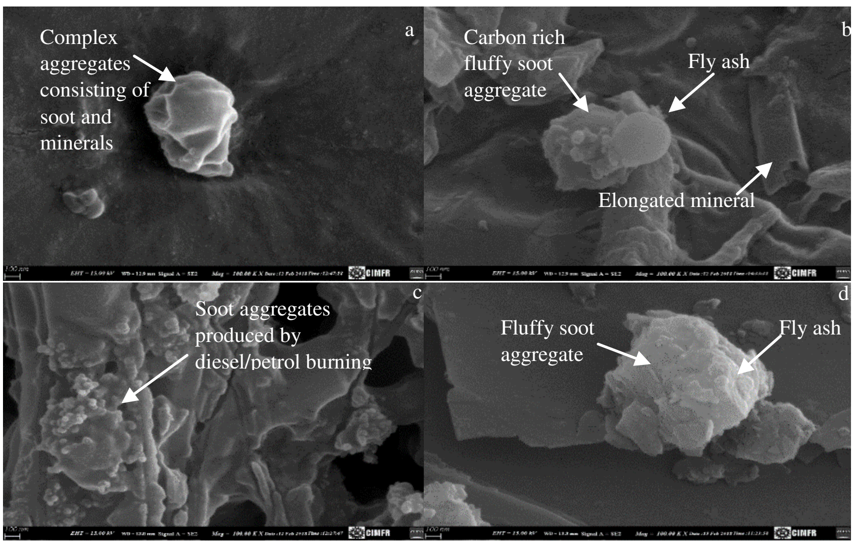


Fig 10. SEM images of coal mine area (a) PM_{10} (b) $PM_{2.5}$& non-coal mine area (c) PM_{10} (d) $PM_{2.5}$

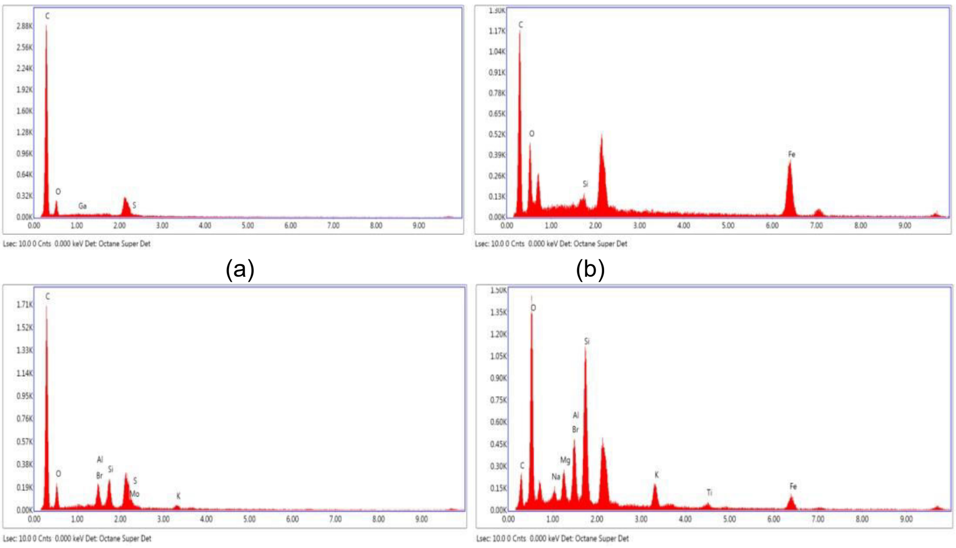


Fig 11. EDX images of coal mine area (a) PM_{10} (b) $PM_{2.5}$& non-coal mine area (c) PM_{10} (d) $PM_{2.5}$

High content of Si and Al with varying percentage of Mg, K, Fe, S and Co are present in aluminosilicates.
It consists of kaolinite, illite, montmorillonite and feldspar, which are typical terrigenous minerals. The consistent occurrence of Fe/Cu with Al/Si/O indicates particle origin from a localized material by weathering of a geological structure unique in nature. The prominent elements present in the PM_{10} particulates of the coal mining area of JCF are C, O, Na, Mg, Al, Si, S, Cl, K, Ca while the prominent elements present in non-coal mining area contains C, O, Na, Al, Ca, Si, S, Cl, K, Mg, Fe and Ti. Elemental composition reveals presence of many elements viz., Si, Al, K, Fe, Ca and C in appreciable quantity. This indicates origin from non-natural sources. Vehicular burning of fossil fuels, industrial sources and thermal power plants in the vicinity of the site are the potential sources. The origin of TiO_2 can be traced to paints, papers and plastics. [42]
EDX spectra of various elements (both PM_{10} and $PM_{2.5}$) at selected locations indicates high concentration of C, O and Si followed by Fe, Al, Ca, Mg, K and Cl (Fig 11). The sites dominated by mining operations shows high C-concentrations with trace of S and K. These particles are typical of C-domination showing spherical, amorphous and typically non aggregated shape.
Non-mining sites showed irregular shaped particles in all size variations. This is attributed to vehicular fossil-fuel combustion. Other dominant elements are- Si, Al, Fe, Mg, Ca & K. They mostly originate from soil, crustal dust and anthropogenic sources. Si associated with Al, Na, Ca, Mg, Fe and K has been thought to indicate the presence of mineral, clay and feldspar particles.
The presence of Fe and Al_2SiO_3 in dust (PM_{10} and $PM_{2.5}$) mineral constitutes mostly inorganic elements. XRD analysis indicates presence of quartz (SiO2) and dolomite [$CaMg(CO_3)_2$] in the dust samples of almost all the area of JCF. [34] This indicates the contribution of regional particulates to the overall dust of the area. Ca rich particles indicates the presence of $CaCO_3$ in calcite phase.
The reported trace elements in the dust (PM_{10} and $PM_{2.5}$) comprises of Pb, Ni, Cu, Mn, Fe. They have been observed at significant concentration levels. The level of As (Arsenic) is below detectable limit. [36]

Health effects of particle size

Particulates are emitted to the ambient air through a lot of activities which can be either natural or anthropogenic. The natural sources are biogenic sources and are non-anthropogenic such as windblown dust from soil surface or sand and wildfire etc. The anthropogenic sources are agriculture, mining, civil construction such as road, buildings etc. These particles when combine with other particulates and gases in the ambient air such as VOC, ammonia, SOx and NOx form secondary and tertiary aerosol of primary, secondary or tertiary nature. Health effects are manifested but is primarily directed towards pulmonary function. The particle size <1μ directly enters alveoli to interrupt $CO \rightleftharpoons O_2$ exchange (Fig 12) causing breathlessness and death. Other effects include:

- Increase in bronchial irritation and asthma
- Frequent heart attacks
- Increased blood pressure
- High level of sugar
- Environmental damage

The environmental damage include:

- Acid rain
- Lowering soil nutrition and crop production
- Decrease in biodiversity

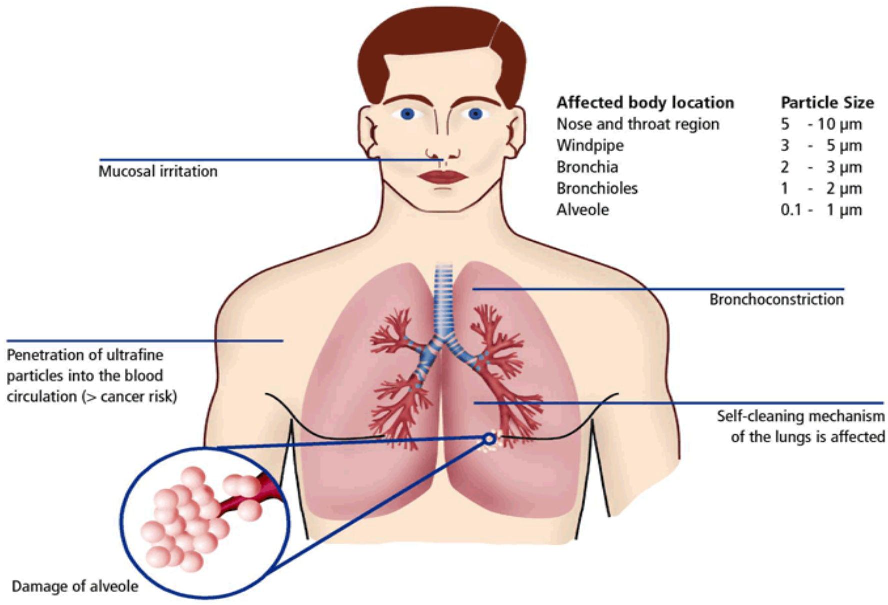


Fig 12. Effect of dust inhalation with different particle size in human lung [43]

Statistical analysis
The bivariate test was taken in the form of Pearson's correlation and ANOVA (one-way) tests to prove the effect of particulates on Covid cases. The relationship was described using linear mapping such as Pearson correlation. It was found that out of the seven cities for PM_{10}, five showed moderate to high positive correlation while two cities of Jamshedpur and Hazaribagh showed negative correlation at insignificant levels (<0.05) (Table 6, Fig 13). Ranchi showed the highest positive correlation of 0.40. The reason can be attributed to three factors- population density, dust pollution, and weather conditions with moisture and wind direction. Dhanbad, Bokaro, Deoghar and Chaibasa showed moderate correlation of 0.25, 0.11, 0.14 and 0.14.

Location	PM_{10}	$PM_{2.5}$
Jamshedpur	-0.02051	-0.02801
Dhanbad	0.247474	0.198208
Ranchi	0.395063	0.368577
Bokaro	0.109909	0.120024
Deoghar	0.135509	0.116604
Chaibasa	0.137967	0.127443
Hazaribagh	-0.04234	-0.07767

Table 6. Pearson correlation between PM_{10} and $PM_{2.5}$ with Covid Cases (%) in seven most densely populated cities of Jharkhand

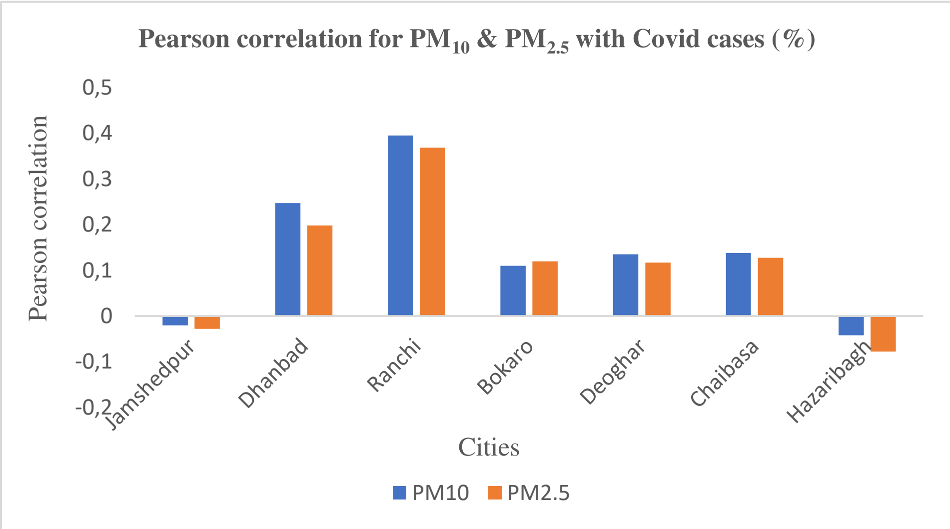


Fig 13. Pearson correlation between PM_{10} and $PM_{2.5}$ with Covid Cases (%) in seven most densely populated cities of Jharkhand

Similarly, for $PM_{2.5}$, five cities showed moderate positive correlation while two cities showed negative correlation with statistically insignificant (<0.05) for Jamshedpur and significant for Hazaribagh (>0.05) (Table 6, Fig 13). Thus, in both the particle size analysis the relationship analysis showed significant attributes showing particulate as the potential carrier of the nCoV.

One-way ANOVA test

Table 7 represents the detail analysis of the ANOVA test. The p-value varies from minimum of 1.98E-40 for Chaibasa to maximum of 8.63E-32 for Jamshedpur. The P-value in all the cases is less than the significance level of 0.05 rejecting the null hypothesis suggesting the means are different with statistically proven relationship. In all the cases, the F value crosses the critical value (F crit) rejecting null hypothesis. Thus, the sample data provide strong enough evidence to conclude that there is relationship between the dust concentration to percent of Covid cases occurring at a site.

Site	F	P-value	F crit
Jamshedpur	188.60478	8.63E-32	3.105157
Dhanbad	266.84656	4.048E-37	3.105157
Ranchi	254.72916	2.174E-36	3.105157
Bokaro	213.82931	1.103E-33	3.105157
Deoghar	262.74828	1.245E-36	3.106507
Chaibasa	328.30060	1.982E-40	3.105157
Hazaribagh	192.08525	4.6E-32	3.105157

Table 7. One-way ANOVA test

III. Conclusion

The investigation of the available data set of particulates of PM_{10} and $PM_{2.5}$ over a period of nearly one month (29 days) in post monsoon and beginning of winter season of the selected sites provide strong evidence for the particulate aerosol as the mode of transmission. Though not specific evidence is available but statistical correlation in linear and ANOVA tests proves beyond doubt that the aerosol particle can be a prime mode of transmission of the nCoV besides other investigations described world over for the spread of COVID 19.

IV. Acknowledgments

Authors are thankful to Director, CSIR-CIMFR, Barwa Road, Dhanbad, India for granting permission to publish this paper.

Funding

This work was supported by CSIR-CIMFR, Barwa Road, Dhanbad-826015, India.

Competing interests

The authors declare they have no competing interests

References

[1.] https://covid19.who.int/?adgroupsurvey={adgroupsurvey}&gclid=CjwKCAjw6qqDBhB-EiwACBs6x0hNwz8nSwyD9L3iyp_v2eJ4dblnoQhimdWO5AQjXBrR5BnilSyUxhoC1qwQAvD_BwE

[2.] Paules CI, Marston HD& Fauci AS, Coronavirus Infections-More than just the common cold, *JAMA network,* 2020.

[3.] Chan JF, Yao Y, Yeung ML et al., Treatment with Lopinavir/Ritonavir or interferon-β1b improves outcome of MERS-CoV infection in a nonhuman primate model of common Marmoset *J Infect Di,* **212** (2015) 1904-13. doi:10.1093/infdis/jiv392.

[4.] Song Z, Xu Y, Bao L et al., From SARS to MERS, thrusting coronaviruses into the spotlight. *Viruses,* **11** (2019) 59. doi: 10.3390/v11010059.

[5.] Tseng YT, Wang S, Huang K et. al., Self-assembly of severe acute respiratory syndrome coronavirus membrane protein, *J Biol Chem,* **285** (2010) 12862-72.

[6.] Yao H, Song Y, Chen Y et. al., *Cell.* Molecular Architecture of the SARS-CoV-2 Virus, **183** (2020) 1-9. doi:https://doi.org/10.1016/j.cell. 2020.09.018.

[7.] To-KKW, Hung IFN, Chan JFW et al., From SARS coronavirus to novel animal and human coronaviruses, *J Thorac Dis,***5** (Suppl 2) (2013) S103-8.

[8.] Berry M, Gamieldien JG, Fielding BC, Identification of New Respiratory Viruses in the new millennium, *Viruses,* **7** (2015), 996-1019. doi:10.3390/v7030996.

[9.] Hui DS, Azhar EI, Madani TA et al., The continuing 2019-nCoV epidemic threat of novel coronaviruses to global health - The latest 2019 novel coronavirus outbreak in Wuhan, China, *Int J Infect Dis*, **91** (2020), 264-266.

[10.] Zhu N, Zhang D, Wang W et. al., A Novel Coronavirus from Patients with Pneumonia in China, 2019, *N Engl J. Med.* **382** (2020), 727-733.

[11.] Perlman S, Coronaviruses post-SARS: Update on replication and pathogenesis, *Nature Reviews (Microbiology),***7** (2009), 439-450.

[12.] Huang C, Wang Y, Li X et. al., Clinical features of patients infected with 2019 novel coronavirus in Wuhan, China, *The Lancet,* (2020) 497-506. doi:https://doi.org/10.1016/S0140-6736(20)30183-5.

[13.] Jefferson T, Pietrantonj CD, Al-Ansary L Aet. al., Vaccines for preventing influenza in the elderly, *Cochrane Database of Systematic Reviews,* **7**(2020). doi: 10.1002/14651858. CD004876.pub3.

[14.] Cleri DJ, Ricketti AJ, Varnaleo JR, Severe acute respiratory syndrome (SARS). *Infect Dis Clin North Am,* 2010. doi: 10.1016/j.idc.2009.10.005.

[15.] Banik G R, Khandaker G, Rashid H, Middle East respiratory syndrome coronavirus "MERS-CoV": current knowledge gaps, *PaediatrRespir Rev,* **16 (**2015) , 197-202. doi: 10.1016/j.prrv.2015.04.002.

[16.] Al-Tawfiq JA, Memish ZA, Middle East respiratory syndrome coronavirus: Epidemiology and disease control measures,*Infection and Drug resistance,* **7 (**2014), 281-287.

[17.] www.accuweather.com (15th Dec., 2020).

[18.] https://www.bing.com/covid/local/jharkhand_india (15th Dec., 2020).

[19.] District Adm, Govt. of Jharkhand (21st Dec, 2020)

[20.] Santarpia JL, Rivera DN, Herrera VL et al., Aerosol and surface contamination of SARS-CoV-2 observed in quarantine and isolation care, *Sci Rep-UK*. **10** (2020), 1273. doi:https://doi.org/10.1038/s41598-020-69286-3

[21.] Liu Y, Ning Z, Chen Y et al., Aerodynamic analysis of SARS-CoV-2 in two Wuhan hospitals, *Nature*. **582** (2020) 557-560. doi:10.2147/IDR.S51283

[22.] Setti L, Passarini F, De Gennaro G et al., Searching for SARS-COV-2 on Particulate Matter: A Possible Early Indicator of COVID-19 Epidemic Recurrence, *Int J Environ Res Public Health* **17** (2020), 2986. doi:10.3390/ijerph17092986.

[23.] Wu X, Nethery RC, Sabath MB et al., Exposure to air pollution and COVID-19 mortality in the US: A nationwide cross-sectional study, *Science Advances,* **45 (2020)**. doi:10.1126/sciadv.abd4049.

[24.] http://www.simaonlus.it/wpsima/wp-content/uploads/2020/03/COVID_19_position-paper_ENG.pdf, Particulate Matter and COVID-19, Italian Society of Environmental Medicine (SIMA) (20th Dec., 2020).

[25.] Kim KH, Kabir E, Kabir S, A review on the human health impact of airborne particulate matter *Environ. Int.* **74** (2015),136-43.

[26.] Cao C, Jiang W, Wang B et. al., Inhalable microorganisms in Beijing's $PM_{2.5}$ and PM_{10} pollutants during a severe smog event, *Environ Sci Technol*, **48** (2014), 1499-507. doi: 10.1021/es4048472.

[27.] Wei M, Liu H, Chen J, Xu C, Li J, Xu P& Sun Z, Effects of aerosol pollution on $PM_{2.5}$-associated bacteria in typical inland and coastal cities of northern China during the winter heating season. *Environ Pollut,* **262** (2020) 114188.

[28.] Chen PS, Tsai FT, Lin CK, Yang CY, Chan CC, Young CY& Lee CH, Ambient influenza and avian influenza virus during dust storm days and background days, *Environ Health Perspect,* **118** (2010) 1211–1216.

[29.] Bianconi V, Bronzo P, Banach M, Sahebkar A, Mannarino M R & Pirro M, Particulate matter pollution and the COVID-19 outbreak: results from Italian regions and provinces, Arch Med Sci, 5 (2020) 985-992.

[30.] Frontera A, Cianfanelli L, Vlachos K, Landoni G & Cremona G, Severe air pollution links to higher mortality in Covid-19 patients: The "double-hit" hypothesis, J of Inf, **81**(2020) 255-259.https://doi.org/10.1016/j.jinf.2020.05.031

[31.] Wang B, Chen H, Chan Y L, Oliver Brian G, Is there an association between the level of ambient air pollution and COVID-19? *Am J Physiol Lung Cell Mol Physiol,* **31** (2020) 9: L416–L421.

[32.] Contini D & Costabile F, Does Air Pollution Influence COVID-19 Outbreaks? Editorial *Atmosphere* **11** (2020), 377. doi:10.3390/atmos11040377

[33.] Schraufnagel DE, The health effects of ultrafine particles. *Experimental & Molecular Medicine,* **52** (2020):311–317.

[34.] Tomasi C, Sandro F& Kokhanovsky A, Atmospheric Aerosols: Life Cycles and Effects on Air Quality and Climate, (Wiley-VCH Verlag GmbH & Co. KGaA) 1st Edition, 2017.

[35.] Heintzenberg J, Properties of the Log-Normal Particle Size Distribution*, Aero Sci and Tech,* 21-1 (1994) 46-48.doi: 10.1080/02786829408959695

[36.] Dubey B, An investigation into air quality status of Jharia Coalfield, Eastern India, PhD Thesis, ISM, Dhanbad, 2012.http://hdl.handle.net/10603/7904.

[37.] Liu H, Zhang X, Zhang H, Yao X, Zhou M, Wang J, He Z, Zhang H, Lou L, Mao W et al., Effect of air pollution on the total bacteria, and pathogenic bacteria in different sizes of particulate matter, *Environ Pollut,* **233 (**2018), 483–493.

[38.] Hänel G, The properties of atmospheric aerosol particles as functions of the relative humidity at thermodynamic equilibrium with the surrounding moist air, *Adv Geophys*, **19** (1976) 73–188. doi:10.1016/S0065-2687(08)60142-9.

[39.] Okazaki K, Submicron particle formation in pulverized coal combustion, *J Aerosol Res Jpn*, **7** (1993) 289–291.

[40.] Sabbioni C & Zappia G Characterization of particles emitted by domestic heating units fuelled by distilled oil, *Atmos Environ*, **26** (1992)18, 3297–3304. doi: 10.1016/0960-1686(92)90346-M.

[41.] Schnaiter M, Linke C, Möhler O, Naumann KH, Saathoff H, Wagner R, Schurath U and Wehner B, Absorption amplification of black carbon internally mixed with secondary organic aerosol, *J Geophys Res*, **110** (2005) D19. doi:10.1029/2005JD006046.

[42.] Reimann C & Caritat P de, Chemical Elements in the Environment. Factsheets for the Geochemist and Environmental Scientist, (Springer-Verlag, Berlin), Vol 137(5), 1998.

[43.] Anonymous from the internet (4th Oct, 2018).

International Journal of Innovation Engineering and Science Research
(ISSN: 2581-4591)
www.ijiesr.com

Marketing Strategy through Machine Learning Techniques: A Case Study at Telecom Industry

Sanjib Kumar Routray,
Faculty of Computer Science & IT, Telecom Training Center,
Bhubaneswar, Odisha, India.

ABSTRACT

The goal of machine learning is uncovering interesting information and decision making from big data in various domains of applications, which can redefine and enhance customer relationships. In the telecom Industry, on a daily basis huge volume of customer detailed record data is being generated due to a large client base. Telecom management stressed upon the decision making about finding new customers is costlier than retaining the existing ones. Accordingly, marketing analyst and customer relationship management (CRM) professionals need to analyze on the reason for churn customers, as well as, study the interesting patterns from the existing churn customers' behaviour.This paper proposes a segmentation of the market and prediction of the customer's behaviour (Churn, Choice of PLAN) model by Clustering and Classifying the customers based on their attributes. Since customer grouping is the main part of Customer Relationship Management (CRM) on predictive analysis on CRM data of telecommunication industry by using classification, as well as, clustering techniques, identification of churn customers and provisioning of suitable affordable choice PLAN are essential factors behind the churning of customers along with making of decision in sales and marketing in the telecom sector. The best data mining strategies are proposed to be used for classifying the CRM datasets for predicting the choice of PLAN and predicting the churn for retention of customers for efficient managerial decision and making project plan to reach the related goals. Accordingly, this paper proposed "Random Forest" Algorithm for predicting the churn output and "Decision Tree" is proposed for predicting selection of choice of mobile PLAN for predicting profitable mobile customers of telecommunication industry by using Rapid-Miner-9.07-tool-kit. For creating effective retention policies under CRM, to prevent churners, categorization of churn customers is highly essential. Accordingly, after classification, "k-means" clustering algorithm is proposed for providing group-based retention offers. The proposed prediction model is evaluated using no. of metrics, like accuracy, precision, recall, f-measure, and receiving operating characteristics (ROC) area.

Keywords-*Customer Relationship Management, Churn prediction, CDR,*

I. INTRODUCTION

Telecommunication industry every time offering quality products and services with new technology for attracting new customers and retaining existing customers while competing with private sectors to defend their positions in the market. In the present scenario, a huge volume of data sets is being generated by telecom companies at an exceedingly fast rate, as there is a range of telecom service providers competing in the market promotion to increase their customer share. Also the target of telecom companies is to maximize their profit and stay on competitive market [1].But customers have multiple options in the form of quality with less expensive services. When a more no. of customers not satisfied with the services of Telecom Company, service migration or switching found from one service provider to another and churn takes place. Churning takes place due to several reasons like service dissatisfaction, binding limit of post-paid customers and saturated competitive market. On continuous churning, the reputation of a company will affect and loyal customers also get affected. Thus customer's population will be decreased with a high loss of revenue [3].For building intimate long term relationships with commercial important customers or individual customer in marketing strategy, achieving quality of service or sustained customer patronage are highly essential [1, 2]. Accordingly, customer relationship management is the efficient tool for achieving the company's goal. The challenging job of CRM administration is the churn prediction for retaining their valuable customers and enhancing the CRM mechanism [5], [6]. For acquiring new customers, requirement of advertisement, involvement of the workforce and promotional marketing is more expensive than retaining existing customers [3]. With immediate attention, identification of existing churning customers can be stopped. Accordingly a high performance model is needed for identifying churn customers and in future also. Many machine learning and data mining solutions can be used to analyze such Call Data Record (CDR) data on which it can identify reasons behind customer churning. For maximizing the profit, *CRM* can employ to design retention strategies for reducing the percentage of churning customers as well as offering of suitable plans to retain customers [2]. Data mining is now popular by different names such as knowledge discovery, machine learning, business intelligence, predictive analysis, and predictive analytics. From the extracted pattern of data, in the data mining process, the behaviour of churn customers easily identified and decision maker's analysis for taking right decisions through CRM. The development of customers depends on service quality, network coverage, load errors, high technology, billing and rewards where customers can compare the service quality and benefits between the service providers [5]. Researchers focus on global prediction rate of churn customers. In general 2% of normal churn customers, create an annual loss of 100 billion dollars and predicting churn customers is 16 times cheaper than attracting new customers [8]. Ensuring of high level churning, many companies already implemented various data mining techniques on differentiating customers due to technological improvements [9] [10]. Telecom companies also need to understand about customer demands and fulfilling their needs for ultimate escaping from other competitor [11] [12]. Accordingly CRM controls the business of customers as per need of services and promotions.Existing models adopted pre-processing methods initially for removing noise for better classifying with improving performance of the models on benchmark CDR data sets [13].But for true representation, multiple algorithms are chosen on these data sets for helping correct decision making of retention and churn prediction [14]. A various machine learning algorithm is proposed in this study and it is validated on CDR of a BSNL, as an Indian telecom company. The performance of a classifier is measured by using the TP rate, FP rate, Precision, Recall, F-measure and ROC-area to prove the best model for churn prediction with the offering of suitable plan to retain customers by achieving high accuracy. For churn and no-churn classification, number of machine learning algorithm we used, where shown better performance compared to other algorithms. Further for market segmentation and study the behaviour of customers, we used k-means clustering algorithm by using attribute selection measures.The remaining part of the paper is structured as follows that provides related work, proposed customer churn and choice of PLAN prediction model, experimental evaluation with results and finally conclusion with future scope.

II. RELATED WORK

Data mining has more importance in analyzing the data and generating interesting, useful patterns and relationships. In the research paper of Clifton (2010), Author stated that originally data being stored in data warehouses, where data mining application needed to analyze the data. As storage became cheaper during 1980s, many organizations develop their storage warehouses and store transaction data for extracting patterns or relationships by adapting Artificial Intelligence methods in the area of KDD. Through applying of data mining techniques, marketing aids of customers of Telecom Company will be fulfilled by proper decision making by applying data mining techniques. According to Author Pradnya, A Shirsath, Vidya Kumar Verma (2013), data mining technology needed for transformation in many fields such as banks, Airlines, railways and so many public/private sector service provider companies to manage and make a decision of such types of huge amount of data. V .Jaaraj, J. Lavanya, J. JagatheshAmairaj, M. Rajkumar (2013) focuses on the existing data mining techniques, i.e. Association rule, Clustering, Classification and described how these techniques were applied to improve the customer relationship with the company for generating considerable profit. Ayman Alawin, Mohmmad Al. ma aitah, Al. Balqua (2014) stated to identify the profitable customers, churn one and develop two models, i.e. one is a physical model (OLAP) (Continuous mining of data sets wherever it resides) and the other is logical model, i.e. adoption of "Classification " on research point of view. From the customer churn prediction point of view in Telecommunication Industry, supervised data mining technique, i.e. "Classification" is used to model the churn prediction with reference to Telecom Industry [25] and in the computer science, application of Telecom big data is being researched from online classification from huge collected data pool [26]. Similarly for finding the targeted high value customers of Telecom Industry, market segmentation is essential through clustering data mining techniques and for increasing the market share, company growth and profitability assessment of data mining based CRM technique, i.e. "Classification" is essential.The problems associated with churning of customers through the most commonly recognized machine learning, data mining techniques that actually supporting the telecommunication company for predicting the churn customers and ultimate retaining the customers under CRM [15]. In the decision tree, there is a limitation in nonlinear complex connections between attributes, but on purning of decision tree, the accuracy of decision tree improves [16]. Decision tree algorithms have many advantages like easy visualizations, use of nonparametric method and processing of numerical and categorical data [17]. *Naive Bayes* is also a guided learning module that predicts invisible data based on the position of Bayesian, which is used to predict churn customer [18]. The novel model presented in [13] by using the classical rule inductive technique (FOIL) shows a hybrid approach linking the adapted *k-means* clustering algorithm to predict churn customer behavior. The control of a large volume of data in today's world provides an opportunity to improve the quality of service to the users. This data includes information about customers' behavior, usage pattern and network operations. Over the past decade, the telecommunications service sector has undergone a major change due to new services, state-of-the-art upgrades [19]–[24] and intensified competition due to deregulation [4]. There is a need to secure important customers, strengthen connection management and retention of CRM. In this paper, in the first phase, customers are segmented using decision rules and in the second phase, a model is developed for every leaf of the tree. Performance of random forests algorithm is best for different types of datasets on selected attributes compared to j48-decision tree, Naïve Bayes for measuring student's performance [27].This hybrid approach is compared with decision trees, random forests, Naïve Bayes decision rule and random tree [28] for churn prediction. Finally, random forest Classification Algorithm and K-Means Clustering Algorithm is proposed in the article as reviewed in this research paper for maximizing organization's satisfaction for increasing loyalty, retaining customer business over their lifetimes [28]

III. PROBLEM DEFINITION

Telecom management seriously thinks to come out of monopolistic mode and planned for enhancing customer relations, strengthening technology and upgrade infrastructure. Accordingly the present study of CRM module of marketing management with different issues analysed through machine learning techniques. Using supervised and unsupervised learning techniques, prediction of performance of machine learning algorithms through monitoring various performance parameters on different telecom datasets will be possible. Churn prediction and profit prediction on suitable PLAN are mainly two important issue in marketing management.Dataset 1 & 2, which is acquired from Telecom Company records numerous attributes act input for a classifier models predicts appropriate churn class with profit class on suitable PLAN.Similarly, the behaviour of customer information with their relationship, clustering model is used by partitioning the complete customer data sets into groups. Through clustering technique, Low, Medium & Risky customer is determined under churn class category of datasets, which will help the management to take suitable action by offering suitable PLAN to the esteemed customers.Different DM and ML techniques are applied throughout the research for suitable accuracy prediction. Algorithms used are Naïve Bayes, Decision Tree, Random Tree and Random Forest on both the datasets.

IV. PROPOSED MODEL

This section presents the proposed customer churn and profit prediction model. Fig. 1 shows the proposed churn prediction model and profit prediction model indicates number of steps. In the first step, data pre-processing is performed which includes data filtering for noise removal, removal of imbalanced data features with normalization of the data. In the second step, different classification algorithms like Decision Tree, Random Tree, Random Forest and Naïve Bayes applied for categorizing the customers into the churn and non-churn customers of one part. In the second part same algorithms applied for categorizing profit and no profit from customers on CDR data sets. In the third step, customer segmentation according to churn and profit is performed using *k-means* clustering techniques. Based on patterns of customer transactional behaviour from data sets, cluster analysis is developed. Finally the model recommends retention strategies for each churn category and individual category on a suitable PLAN of customers.

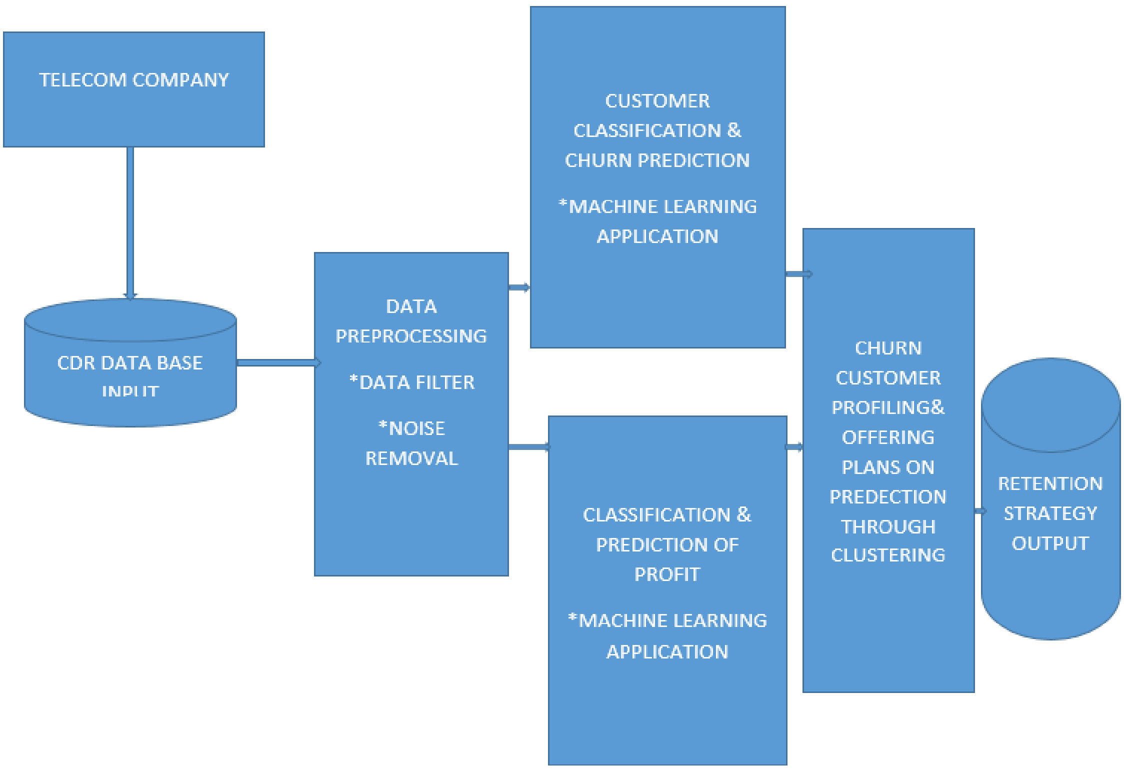


Fig.1 Proposed model for customer churn and profit prediction.

A. Data Pre-processing

1) In telecom CDR (Call Detailed Records) data set-1 as mentioned in Table-1, number of missing values, "Null" value first removed and further filtered imbalance attributes, which is the first step of noise removal under data pre-processing. In this telecom dataset, the number of attributes is 10, with useful features for using information gain and correlation attributes ranking filter techniques for feature selection on RapidMiner-9.7 tool kit to get high ranking values. Out of 10 attributes, some attributes improve performance measures and are useful for churn-decision-making process. Same procedures adopted for CDR data set-2 with 5 no. of equal important attributes for predicting profit-decision making process through performance measurement. Accordingly the performance of classification increases, if the dataset contains highly predictive and valuable variables.

.

Table-1 Number of Filtered Features of CDR-data set-1 & 2.

Sl no.	#Features	Sl no	#Features
1	State	1	GSM NO
2	Area Code	2	Category
3	Phone No.	3	Group
4	Int. Plan	4	FRC PLAN
5	Voice Mail Plan	5	PROFIT
6	Voice Mail No.		
7	Total Day Minute		
8	Total Day Call		
9	Total Day Charge		
10	Churn		

B. Customer Category and Prediction

There are three categories of customers like individual, business and others. Out of which three groups of customers divided like lower income group, middle income group and higher income group exist. According to choice of service of different customer groups, some of the customers are loyal and they are declared as non-churn customers. But some are churn customers, who is not satisfied with the services of the telecom company. The proposed model targets those churn customers by applying different machine learning classification techniques like Decision Tree, Random Tree, Random Forest & Naïve Bayes and devise retention strategies to overcome the churning problems created by different customers. Further from performance measurements of different algorithms, identification of best classification is analysed between churn and non-churn customers. Also above algorithms correctly classified the profitable customers to the company from non-profitable one between different customer groups.

V. EXPERIMENTS and RESULTS

A. Dataset Description

In this study, datasets are obtained from one GSM telecom service provider for studying the customer churn prediction problem. The data are extracted from the customer service usage pattern Call Detail Record (CDR), which consists of labeled data with two classes where 15% data is labeled as "T" (true customers) that represents churner and 85% data is labeled as "F" (false customers) that represent non-churners. It has three main types of' attributes that include call behaviour, or usage attributes, financial information attributes and marketing related attributes, selected on feature selection techniques that allows identifying the most relevant, useful and effective attributes for customer churn prediction. Another part of transactional datasets are also taken with four main attributes like lower income group (LIG), Middle Income Group (MIG) and higher income group (HIG) selecting suitable PLAN for selecting two categories of customers like profitable and no profitable (Table-2).

Table-2 Dataset Description

CRM Data sets	Instances	Selected Attributes	Models Used
1	49	5	Classification for Churning of Customers

			Clustering for Market Segmentation
2	120	4	Classification for Profitable PLAN Prediction

On the two CDR data sets, we performed a number of experiments on the proposed machine learning model of by using a Rapid Miner-9.7 toolkit for providing the factors behind customer churn. Different techniques like Random Forest, Decision Tree, Random Tree and Naïve Bayes tested. Out of which, Random forest is a useful technique for classification and shows better accuracy (85.71%) and less error in comparison to other on correctly classifying of churn and non-churn customers with 10 fold cross validation as mentioned in Table-3. Similarly for selecting the suitable PLAN, for future decision making by the management, above machine learning techniques also used. The Decision Tree technique shows excellent (100%) classification accuracy with no error with 10-fold cross validation, as mentioned in Table-4 among individual, business and other groups of people under Lower income group (LIG), Middle Income Group (MIG) and higher income group (HIG) selecting suitable PLAN. The performance vector output of Rapid-Miner tool-kit in case of above two best techniques. Further, in this study, Attribute Selected Classifier algorithm is used for identification of factors, which clearly indicates the finding of churn customers. Also in this paper, k-means clustering algorithm is used for creating retention policies by decision makers considering between of churn and no-churn customers with profit and no-profit customers of different groups.

B. Performance Evaluation Matrix

The performance of different classifiers is measured for predicting of churn and profit on the basis of some parameters i.e. (1) True positive rate (2) False positive rate(3) Precision (4) Recall (5) F-Measure(6) ROC (Receiver Operating Characteristics). Confusion matrix is used to evaluate the classifier quality for a two class problem, i.e. True Positive, True Negative, and False Positive & False Negative. Confusion matrix is a useful tool for analyzing the performance of classifiers by recognizing tuples of different classes. Similarity measurement of classifier's performance as mentioned in Table-5 includes accuracy, sensitivity (recall), and specificity, precision etc. Where further construction and evaluation requires partitioning of training set and test set. Holdout, Random Sampling and Cross validation are important steps of partitioning.

Table-3 Classification Accuracy & Errors of Various Algorithms on Own Churn-Balance Dataset for churn prediction

Method Used	Classification Accuracy (%)	Classification Error (%)
Random Forest	85.71	14.29
Random Tree	78.57	21.43
Decision Tree	57.14	42.86
Naïve Bayes	50.00	50.00

Table-4 Classification Accuracy & Errors of Various Algorithms on Own Profit Measurement Dataset for Profit prediction

Method Used	Classification Accuracy (%)	Classification Error (%)
Decision Tree	100	0.00
Random Forest	97.22	2.78
Random Tree	97.22	2.78
Naïve Bayes	91.67	8.33

Table-5: Performance measurements of various Algorithms of Classification for Churn and Profit

Method	TP	FP	Precision	Recall	F-measure

Used for Churn	Rate	Rate			
Random Forest	1	0	0.77	1	0.87
Random Tree	0.87	0.2	0.77	0.87	0.40
Decision Tree	0.8	0.2	0.44	0.8	0.56
Naïve Bayes	0.62	0.6	0.55	0.62	0.58
Method Used for Profit	TP Rate	FP Rate	Precision	Recall	F-measure
Decision Tree	1	0	1	1	1
Random Forest	0.97	1	1	0.97	0.98
Random Tree	0.97	1	1	0.97	0.98
Naïve Bayes	0.97	1	0.94	0.97	0.96

To further validate our findings, the performance of algorithms in Table-5 show that TP and F-measure is higher for Random Forest classifier and Decision Tree classifier as compared to others in case of churning and profit. The Area under Curve (AUC) is a selective performance measure which is used by many researchers in the prediction model for measuring the accuracy.

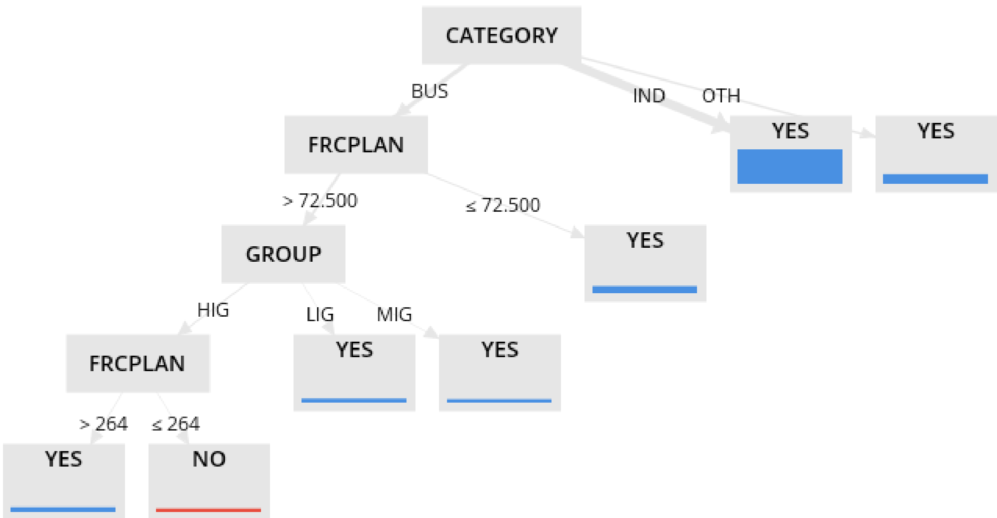


Fig. 2 Decision Tree output of Rapid-Miner tool-kit for PROFIT prediction

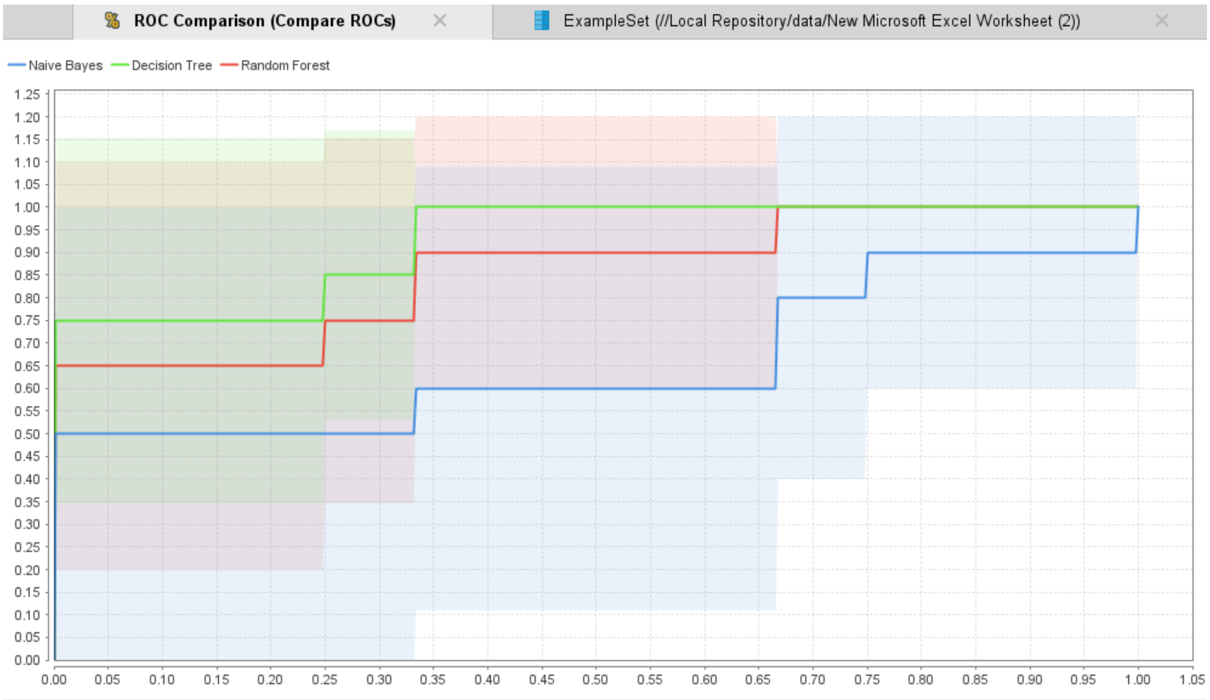


Fig. 3 ROC outputs in Rapid-Miner tool-kit for PROFIT

The complete Decision Tree output of a Rapid-Miner in Fig. 2 indicates more profit in case of "IND" head. ROC area denotes the average performance against all possible cost ratios between FP and FN. If the ROC area value is equal to 1.0, this is a perfect prediction. Similarly, the values 0.5, 0.6, 0.7, 0.8 and 0.9 represent a random prediction, bad, moderate, good and superior respectively. Accordingly ROC output shows better output in case of Decision Tree as per Fig.3.

C. Customer Segmentation and Retention

Based on the behaviour of customer information with their relationship, segmentation or clustering is used by partitioning the complete customer data sets into groups. Out of number of clustering algorithms, we used k-means clustering algorithm. K-means clustering is the iterative approach, where data can be segmented into different groups from complex heterogeneous large data sets. Arithmetic mean value of real valued data is the representative of a cluster, where we can find the hidden pattern that represents one class or cluster. In this study, the *k-means* algorithm segments the data into three groups due to the nature of the data. The three groups represent cluster-0, cluster-1 & cluster-2 among Low, Medium and Risky customers as mentioned in Fig. 4. Fig. 5, shows the number of customers in each segment, according to the value of k in *k-means* to 3 can lead to better segmentation. During n testing, it is found that, three clusters (0, 1 & 2) along with 31 numbers of customers are under no-churn category and 10 under churn category of sample data sets. Therefore, for retaining the churned customers, suitable action to be taken by the decision makers for making profit by offering suitable PLAN to the individual customers instead of business and other groups as per above classification technique.By offering suitable PLAN to a specific group of customers, Telecom Company easily understands the behaviour of customers and ultimately enhances retention and marketing performance.

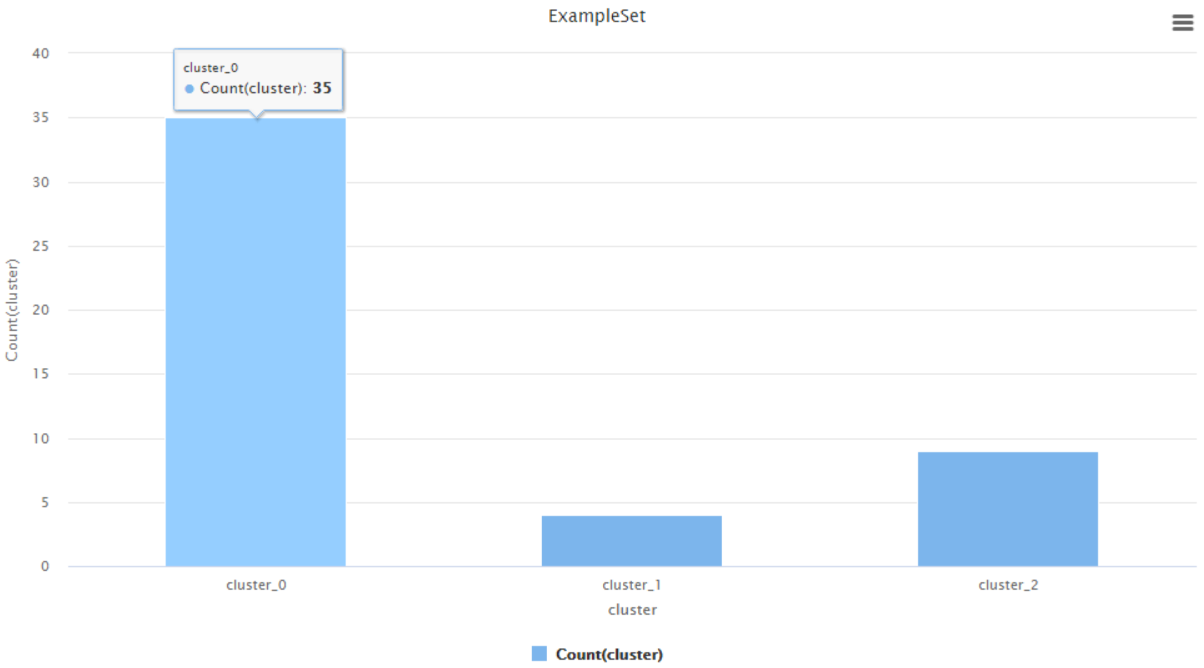


Fig. 4 Cluster count output of k-means Algorithm

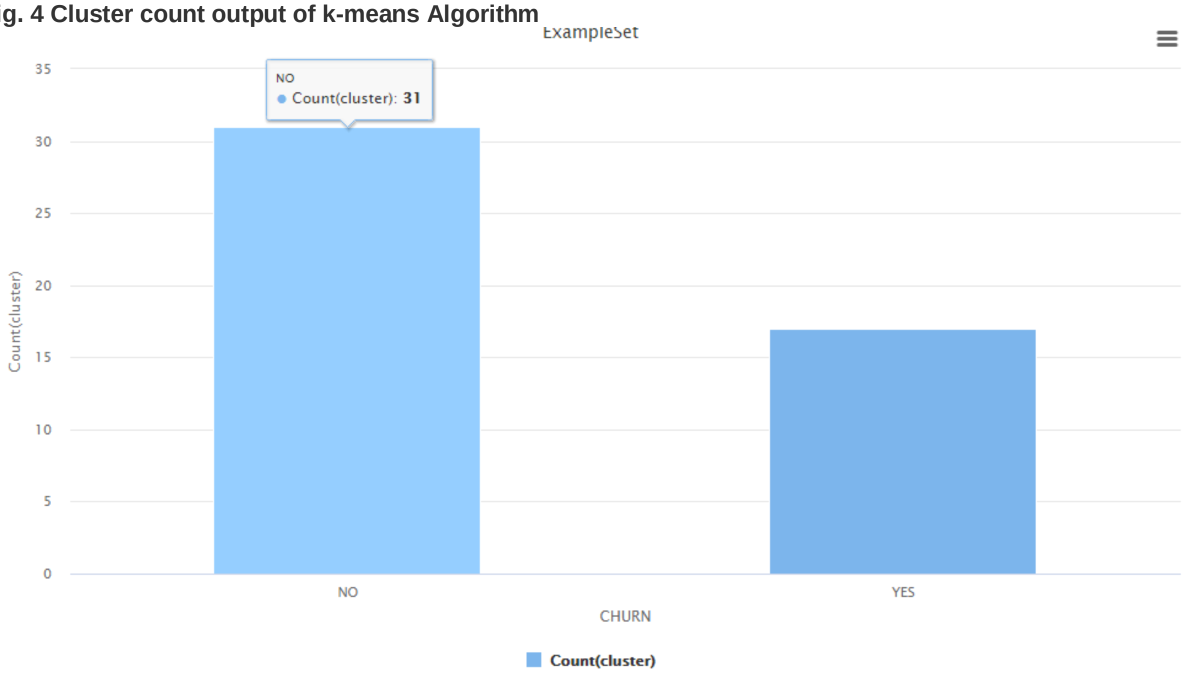


Fig. 5 Cluster count output churn and no-churn customers

VI. CONCLUSION

The challenging issue in the present competitive telecom market is churning prediction of the CRM to retain the liable customers by identifying a similar groups of customers and offering competitive PLAN (Tariff of mobile customers) to the respective groups for making more profit in organization. Accordingly, researchers finding the key factors of churning by developing certain machine learning models for retaining of customers in telecom organization. In this paper, through standard performance metrics, Random Forest, Random Tree, Decision Tree & Naïve Bayes classification techniques proposed and evaluated for the purpose of churning. Out of which Random Forest produced a better result that is 85.17% (Accuracy). Similarly for the purpose of profit calculation by decision makers, Decision Tree produced a better result that is 100% (Accuracy). Finally from the dataset, churn factors identified and performed un-supervised k-means clustering technique according to their risk of churning also recommended certain guidelines on customer retention under CRM for efficient decision making by decision makers.

In the scope of the study, applying of deep learning under the Artificial Intelligence for predictions and pattern analysis investigation is highly essential according to the frequent changing behaviour of churn customers in future.

REFERENCES

[1] S. Babu, D.N. Ananthanarayan and V. Ramesh, "A Survey on factors impacting churn in telecommunication using data mining techniques", int. j. Eng. Ras. Techmol.,vol 3, no.3, pp.1743-1748, Mar 2014.

[2] C. Geppert, "Customer Churn Management: Retaining high-margin customers with relationship management techniques, KPMG & Associates YarhandsDissou Arthur/KwakuAhenkrah/David Asamoath, 2002

[3] W.Verbeke, D.Martens, C.Mues, and B.Basens, "Building Comprehensible Customer Churn prediction models with advanced rule induction techniques", Expert system, Appl., vol.38, no. 3, pp. 2354-2364, Mar.2011

[4] Y. Huang, and M . T. Kechadi, "A rule based method for customer Churn Prediction in telecommunication services", in Proc. Pacific-Asia Conf: knowl. Discovery Data mining Berlin, Germany: springer, 2011, pp. 411-422

[5] Bhave and Ashish, "Customer Satisfaction Measurement", Quality & productive journal, pp. 562-572, 2002

[6] M. Kaur, K. Sing, and N. Sharma, "Data mining s a tool to predict the churn behaviour among Indian Bank Customers," Int. j. Recent Innov. Trends Comp. Commun., vol.1, no. 9, pp. 720-725, sept. 2013

[7] V.L. Migueis, D. VanmdanPoel, A. S Camanho and j.F.ecunha, "Modelling Partial Customer Churn: On the value of first product-category purchase sequences," Expert syst. Appl., vol.12, no.12, pp. 11250-11256, Sept. 2012.

[8] D .Manzano-Machob, "The architecture of a churn prediction system based on stream mining ", in Proc. ArtifIntell. Res. Develop, 16th int. conf. catalan Assoc. Artif. Intell., vol.256, oct 2013, p.157.

[9] P.T. Kotler, Marketing Management Analysis, Planning, Implementation and control, London, U.K: Prentice-Hall, 1994.

[10] F.F. Reichheld and W.E. Sasser, jr., "Zero defections:, Quality comes to services", Harvard Bus Rev., vol 68, no. 5, pp. 105-111,1990

[11] NematSheereen, S, "A Study on Customer Satisfaction of BSNL Services in Kerala", Intercontinental Journal of Marketing Research Review, ISSN: 2321-0346 – online, ISSN: 2347-1670,Vol. 2, Issue 9 pp.137- 140, 2014

[12] H.-S. Kim and C.-H. Yoon, "Determinants of subscriber churn and customer loyalty in the Korean mobile telephony market", Telecommunication. Policy, vol.28, nos. 9-10, pp.751-765, Nov.2004.

[13] Y. Huang and T. Kechadi, "An effective hybrid learning system for telecommunication churn prediction," Expert System. Appl., vol.40, no.14, pp.5635-5647, Oct.2013

[14] O. G. Ali and U. Ariturk, "Dynamic churn prediction framework with more effective use of rare event data: The case of private banking", Expert syst. Appl., vol.41, no.17, pp. 7889-7903, Dec. 2014.

[15] V. Lazarov and M. Capota, "Churn prediction", Bus. Anal. Course, TUM Compt. Sci, Technische Univ. Munchen, Tech. Rep., 2007. [online]. Available :http://citeseerx.ist.psu.edu/viewdoc/download? doi=10.1.1.462.7201 & rep=rep1& type=pdf

[16] R .Vadakattu, B.Panda, S. Narayan, and H. Godhia, "Enterprise subscription - churn prediction", Proc. IEEE int. conf. Big Data pp.1317-1321, Nov-2015

[17] M. Hassouna, A. Tarhini, T. Elyas and M.S. AbouTrab, "Customer churn in a mobile markets a comparision of techniques, jan. 2016 [Online] Available http://arxiv.org/abs/1607.07792.

[18] V. Umayaparvathi and K. Lyakutti, "Applications of data mining techniques in telecom churn prediction", int. j. Compt. Appl. Vol. 42, pp. 5-9, Mar. 2012

[19] A.Amin et al., "Cross company customer churn prediction in telecommunications: A comparision of data transformation methods", Int. j. inf. Manage., vol. 46, pp. 304-319, jun.2019.

[20] A. Amin, B. Shah, A. M. Khattak, T. Baker, H. Ur. Rahman Durani, and S. Anwar, "just-in-time customer churn prediction: with and without data transformation", in Proc. IEEE Congr. Evol. Comput. Jul.2018, pp.1-6.

[21] A. Amin et al., "customer churn prediction in the telecommunication sector using a rough set approach", Neurocomputing, vol. 237, pp. 242-254. May-2017.

[22] A. Amin et al., "Comparing oversampling frequencies to handle the class imbalance problem: A customer churn prediction case study",IEEE Access ,vol.4, pp.7940-7957, 2016.

[23] M. Ahmed, H. Afzal, A. Majeed, and B. Khan, "A Survey of evolution in predictive models and impacting factors in customer churn", Adv. Data Sci. Adapt. Anal., vol. 9, no. 3, jul. 2017, Art. No. 1750007.

[24] R. Rajamohamed and j. Monokaran, "Improved credit card churn prediction based on rough clustering and supervised learning techniques", Cluster comput., vol. 21, no. 1, pp. 65-77, Mar.2018

[25] Kiran Dahiya, KanikaTalwar, "Customer Churn Prediction in Telecommunication Industry using Data mining technique-A Review", International Journal of Advanced Research in Computer Science and Software Engineering, Vol-5, Issue-4, 2015.

[26] Jiang Heling, An Yang, Fengyun Yan, Hang Miao, "Research on pattern Analysis and data Classification methodology of Data mining and knowledge Recovery", International journal of Hybrid Information Technology, IEEE, Vol-9, pp.179-186, 2016.

[27] Madhuri T. Sathe, Amol C. Adamuthe, "Comparitive study of supervised algorithms for prediction of student's performance", I. J .Modern Education and Computer Science, Published Online in MECS (http://www.mecs-press.org/) vol-1,pp.1-21, Feb – 2021.DOI: 10.5815/ijmecs.2021.01.01

[28] Irfan Ullah, Basit Raza, Ahmad Kamran Malik, Muhammad Imran, SaifUl Islam & Sung Won Kim, "A churn prediction model using Random Forest Analysis of Machine Learning Techniques for churn prediction & factor Identification in Telecom sector", Open Access Journal, IEEE Access, volume-7, pp.60134-60149,2019.

International Journal of Innovation Engineering and Science Research
(ISSN: 2581-4591)
www.ijiesr.com

Decentralised voting with Ethereum blockchain

Kushal Chaganlal Choudhary
IT department, 3rd year
Pimpri Chinchwad College of Engineering, Akurdi
Pune, India

Saurabh Achal Agrawal
IT department, 3rd year
Pimpri Chinchwad College of Engineering, Akurdi
Pune, India

Mihir Manohar Gadhe
IT department, 3rd year
Pimpri Chinchwad College of Engineering, Akurdi
Pune, India

Mrs. Rohini Pise
IT department, Asst. Professor
Pimpri Chinchwad College of Engineering, Akurdi
Pune, India

ABSTRACT

When contrasted to the old methodology of pen and paper voting, e-voting decreased election costs and provided some convenience, but it was deemed unreliable since anyone with physical access to the system might impede the mechanism and alter the votes. A central framework is additionally necessary to control the whole strategy, from electronic voting through constituent comes about and following the results. Voters are not completely secure since their votes can be promptly focused on. It too postures a noteworthy threat to voting rights and openness. The purpose of this consider is to form a decentralized instead of centralised e-voting framework utilizing blockchain innovation, which guarantees voter personality security, information exchange security, and unquestionable status through an open and straightforward voting prepare.

Keywords—Blockchain technology, Decentralised System, Electronic voting, Ethereum, Secure.

I. Introduction

Elections are one of the basic cornerstones of any democratic society, since citizens vote for the most qualified candidate in order to build a healthy democracy. With technological advancements, the mechanical voting mechanism proved to be significantly more fluid, serviceable, and cost-effective, resulting in increased dependability and accuracy. Blockchain is a data structure that consists of blocks, each of which is linked to every other block via a chain. Each block contains information, a hash, and the past block's hash. In the event that the information in a block is adjusted, the hash of the block is additionally adjusted, be that as it may the taking after block will have the same unmodified hash as the going before piece, nullifying this block and all consequent blocks. This is to avoid tempering because changing one block requires calculating hashes for all subsequent blocks, but hackers can currently compute hundreds of thousands of hashes in a matter of seconds. To avoid this issue, it employs the proof-of-work idea, which slows down the formation of new blocks

II. LITERATURE SURVEY

A variety of approaches have been developed to introduce differences in electronic and online voting systems, using various strategies and procedures. Despite the system's security to some level, voting continues to take place. Information and processes must be monitored and controlled that safeguards and protects the safety and privacy of voters and their information. Block verification using the Proof of Stake protocol does not require unnecessary computations. It's been implemented for Ethereum as well as a few other altcoins. Proof-of-stake methods partition stake blocks proportionally to the present wealth of miners, rather than proportionally to the relative hash rates of miners (i.e. their mining power).

Sr. no.	Name of paper	Authors	Technique used
1.	Survey on Blockchain Based E-Voting Recording System Design	Linh Vo-Cao-Thuy, Khoi Cao-Minh, Chuong Dang-Le-Bao and Tuan A. Nguyen	AES algorithm
2.	Online Voting System	Vaibhav Anasune, Pradeep Choudhari,Madhura Kelapure and Pranali Shirke Prasad Halgaonkar	Homomorphic Encryption Technique
3.	Blockchain-Based E-Voting System	David Khoury,Elie F. Kfoury, Ali Kassem and Hamza Harb	Geth: Go-Ethereum
4.	Blockchain Based E-Voting Recording System DesignSr	G Bhavan	ECDSA(Elliptic Curve Digital Signature Algorithm)
5.	Decentralized Voting Platform Based on Blockchain	Friðrik Þ. Hjálmarsson , Gunnlaugur K . Hreiðarsson	HTML5 web-app compiled using Apache Cordova
6.	Votereum : An Ethereum-based E-voting system	Rifa Hanifatunnisa and Budi Rahardjo	External Personal Account(EOA)

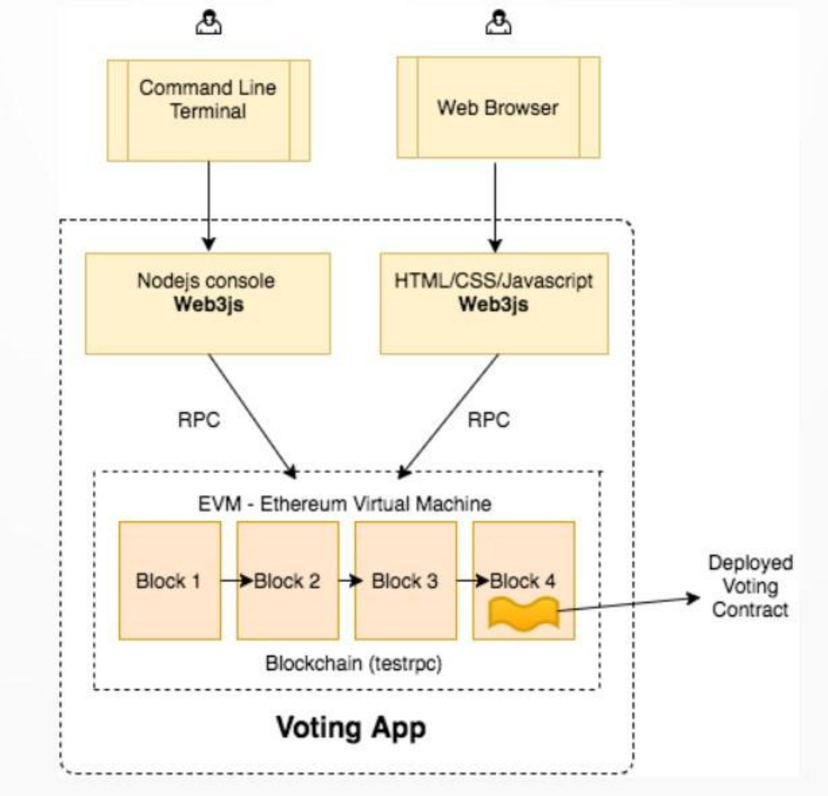
The theory behind Confirmation of Vote is that acquiring a sufficient amount of digital currency may be more difficult for miners than acquiring enough powerful computing equipment. The described polling method includes physical and logical verification of the voter and the voter's information. The tangible record of the computerised voting process can be verified (that is the national identification and biometric authentication) [Sravani C., Murali G, 2019].

Sr. no.	Factors	Ballot based systems	Electronic voting machines	Online voting systems	Our Blockckahin Model
1.	Fraud prevention	Average	Low	Low	Good
2.	Validity of ballot	Low	Good	Good	Good
3.	Vote tallying time factor	Extremely slow	Slow	Fast	Fast
4.	Cost factor	Expensive	Extremely expensive	Expensive	Less expensive on the long run
5.	Accessibility	Low	Low	Average	High
6.	Scalability	Low	Low	Average	High

III. PROPOSED SYSTEM

A few apparatuses are utilized within the proposed framework, counting ganache, truffle system, npm, and metamask. Truffle imports shrewd contracts onto the blockchain, while ganache runs the inner blockchain, which can be available by metamask. A client must have a few Ether, or

Ethereum's cryptocurrency, in arrange to make an account with a wallet address. To compose a exchange to the blockchain, the client must pay a exchange charge known as gas.After votes are cast, the process is finished by minners, a group of nodes in the network. To finish the transaction, these miners compete with one another.The miners that succeed in this transaction are rewarded with ether, which is paid by users in exchange for their votes. For mining purposes, we shall use ganache software instead of nodes.



3.1 Preliminaries

Our proposed approach may be implemented with 64-bit hardware/machines, Windows 7 and later, NMP dependencies, Truffle framework, Metamask, solidity toolkit, and Ganache..

a) Dependency NPM(Node Package Manager)
b) Truffle framework
c) Ganache
d) Metamask
e) Coding language; solidity, HTML, JavaScript, CSS

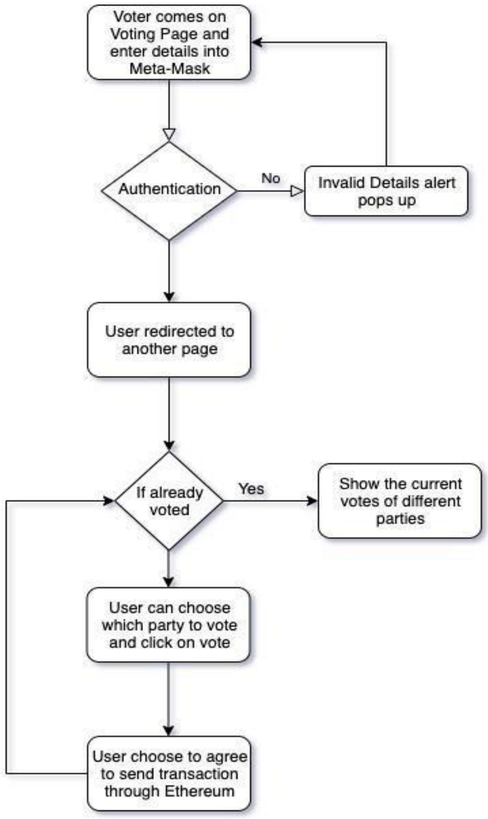
- **NPM (Node Package Manager) :**NPM is a package manager that allows you to manage, instal, update, and uninstall node.js packages in your project. It's a command-line programme. It has two modes of operation: local and global. All node.js applications are affected in global mode, but just a specific directory of an application is affected in local mode.
- **Truffle framework :**Truffle is a robust tool for interacting with Ethereum smart contracts. It is used for smart contract compilation, deployment, and linking, as well as providing a testing platform for automated contracts and managing networks and packages.
- **Ganache :**It was previously called as Testrpc and is available in both command line and graphical user interface versions. A fake blockchain creates ten regular Ethereum addresses,

each with its own private key and a simulated hundred ether. There is no mining with ganache; instead, it confirms each transaction automatically. It works with operating systems such as Windows, Linux, and Mac.

- **Metamask :**Metamask is an open source, user-friendly solution for ethereum transactions with a graphical user interface. Ethereal Dapps can work in your framework browser without requiring a full ethereum hub. Metamask is basically a connect between a browser and the Ethereum blockchain.
- **Solidity**: Solidity is a high-level language for contracts that uses JavaScript syntax. It's a way for converting EVM machinecode into basic instructions. It has the same operators as JavaScript, but it has four value types: Boolean, Integer, Address, and String.

3.2 Working of the system

After logging in to the voting website, the voter must use the Metamask Chrome Extension to connect to the local blockchain. The page is reloaded once the user is connected, and the user may see the candidates and current votes. Below that is the option to vote for a candidate; the voter selects the candidate and clicks on vote; a metamask pop-up appears, informing the user of the Ethereum transaction that must be completed; once the user clicks on Vote, the vote is given to the selected candidate, assuming the voter has not voted previously. A failed transaction will occur if the user has already voted and attempts to vote again. The vote will not be counted.



Flow model of the system

Ganache is utilized to form a local blockchain, and metamask is utilized to associate to it. The Truffle system empowers the movement of solidity-based smart contracts to a neighborhood blockchain.

Metamask permits clients to move Ether from one account to another when they vote. Each client is doled out a one of a kind ID, which is an Ethereum Address, a private key, and some Ethers are designated to each voter's account. When a client votes, Ether is moved from the voter's account to the Candidate's account, and all exchanges are handled through blocks. Once the extend is launched, all transactions will be available to everybody.

Metamask is used to connect to a local blockchain that was created with Ganache. The Truffle framework allows solidity-based smart contracts to be moved to a local blockchain. When voting, Metamask allows users to move Ether from one account to another. Every user is given a unique ID, which is an Ethereum Address, as well as a private key, and each voter's account is given an exact quantity of Ether. When a user votes, Ether is taken from the voter's account and credited to the candidate's account, and all transactions are carried out using blocks. All transactions will be open to the public once the initiative is launched [Hardwick, Freya Sheer, 2018].

3.3 Implementation and results

3.3.1 Setting up

The first thing we need to do is start up Ganache and run local blockchain. There will be no transaction after setting up ganache because we haven't done any yet (refer fig 1). By executing a command on the command line, we can now move the smart contract to the blockchain using the truffle framework. We've also used cmd to access the NPM directory. We start the project using the NPM directory using cmd after moving the smart contract (refer fig 2).

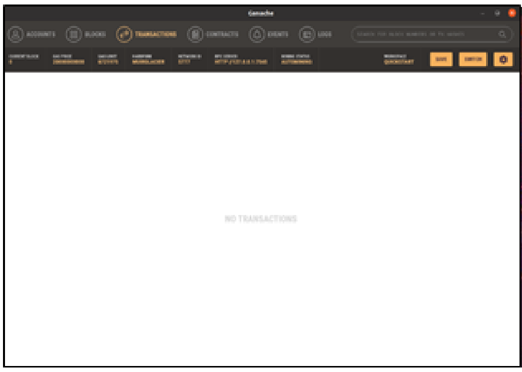


Fig. 1

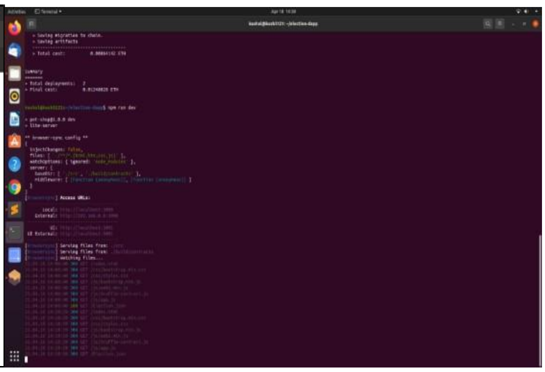
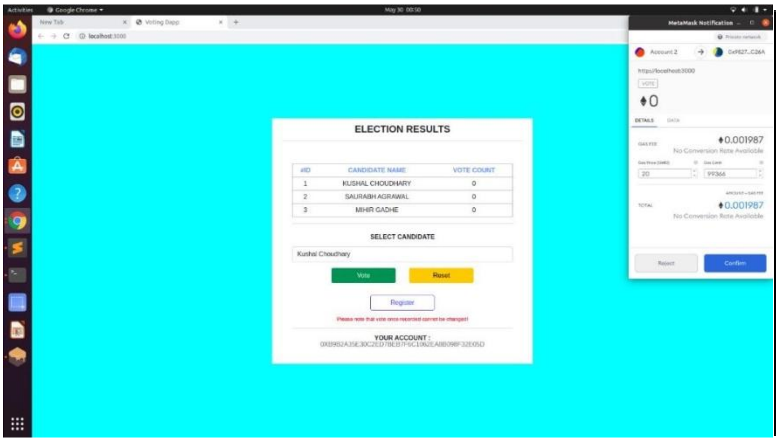
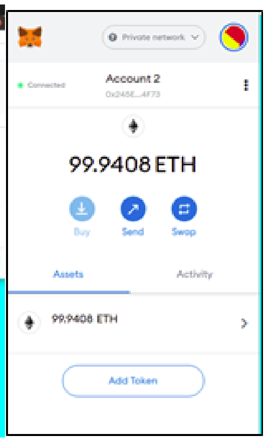


Fig. 2

3.3.2 User interface

Users engage with the e-voting system through the user interface. After logging in, the main screen appears with zero votes; the user is unable to vote until they import their account by inputting their private key.

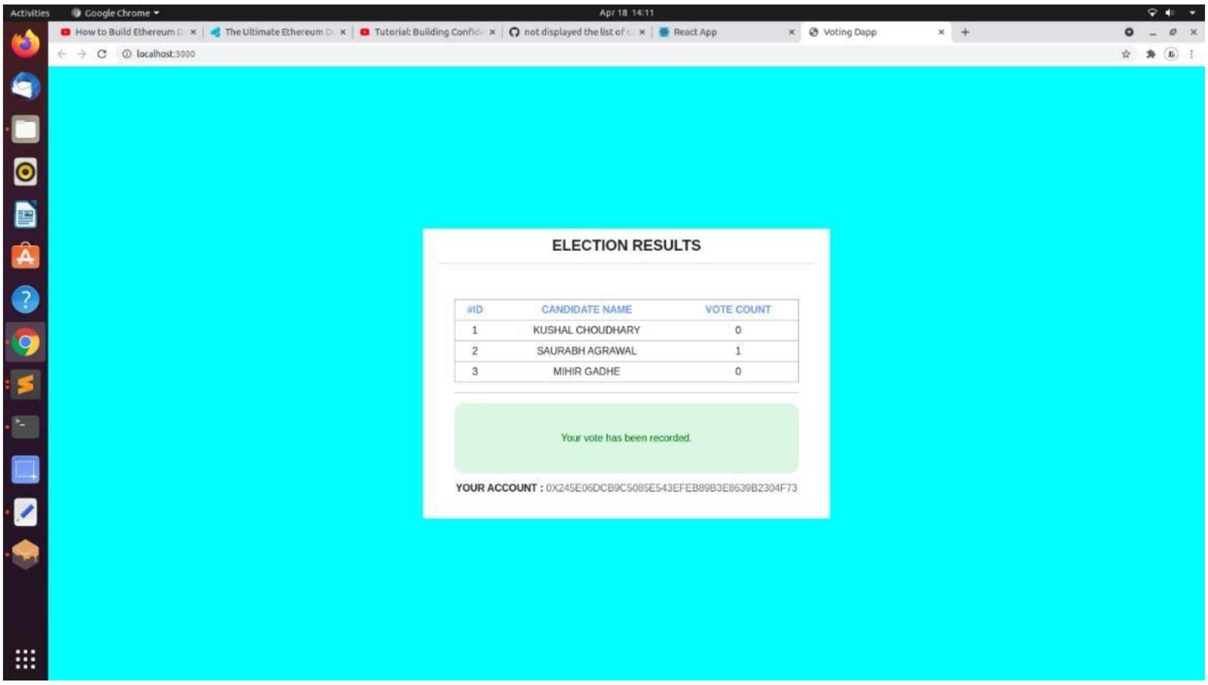




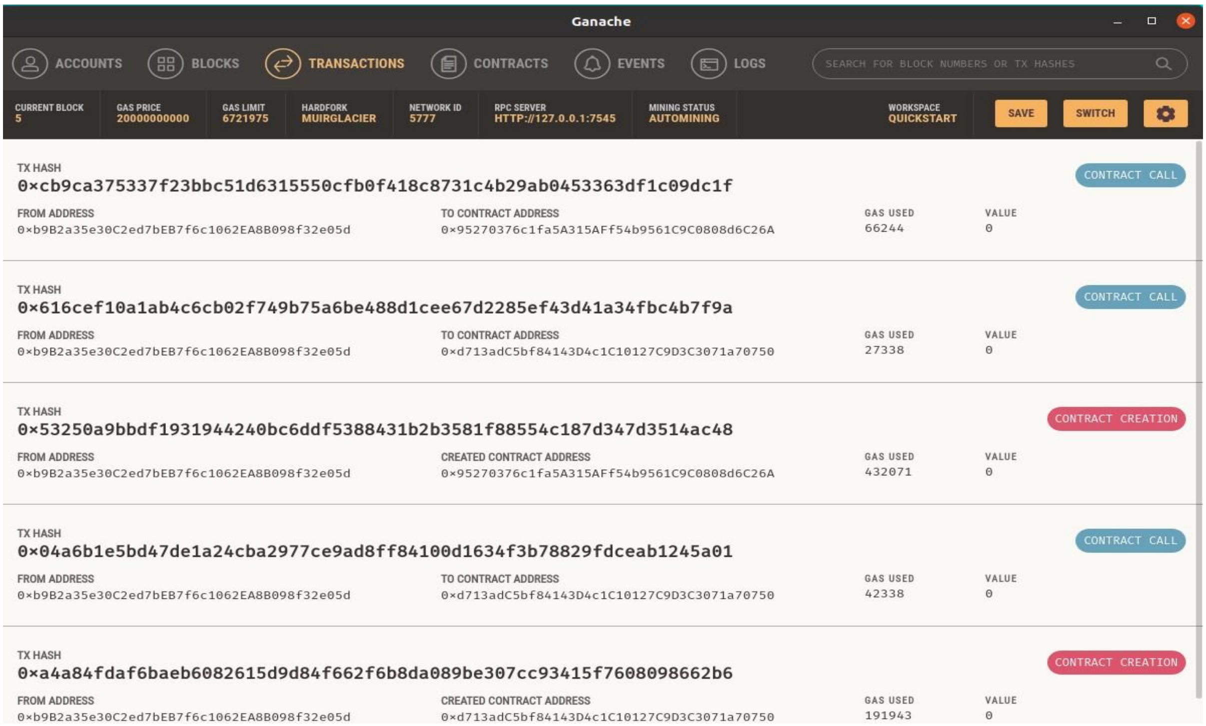
By entering the private key above, the voter imports their account. The electorate selects a candidate, and the metamask pop-up appears when the vote button is selected to finalise the transaction. After confirmation, the voter is taken to the main page, where just the results are displayed, but you can no longer vote. Others can vote in the same way by importing their accounts.

3.3.3 Checking the transactions

The transaction list will be made public to allow users to easily tally their votes.



By glancing at the transaction list, people can check the votes they've cast.



IV. APPLICATIONS

The blockchain records all exchanges (counting votes) in a disseminated record permitting open review of votes cast for a candidate, but the personality isn't uncovered. This permits freely irrefutable voting, whereas keeping up namelessness and avoids false votes. Investigate was in planning a framework is to permit voting beneath duress and uncovering a whole district's votes at the same time.

- Can be used in National Elections.
- Can be used in Television shows.
- Can be used in taking mass opinions.

V. CONCLUSION

Blockchain technology, a recent invention in the area of voting systems, has proven to be not only time and cost effective, but also safe and secure, making it more dependable and exact than previous techniques. We employed blockchain-based e-voting with smart contracts in this work, which feature a set of rules guiding communication and contract decision-making between participants. For implementation, many tools like as Ganache, Truffle framework, NPM, and metamask were utilised.

Seeing as blockchain technology is decentralised, it is very easy to temper and change such a system. Our proposed solution gives voters convenience by allowing them to connect to a system with an easy-to-use user interface, which allows them to cast their vote by importing their account and easily evaluate their vote. It instils confidence in voters by ensuring that their votes are counted and stored securely.

REFERENCES

[1.] Zhang,S., Wang, L. &Xiong, H. Int. J. Inf. Secur. (2019) Chaintegrity: blockchainenabled large-scal e-voting system with robustness and universal verifiability. International Journal of Information Security.

[2.] Gjøsteen K, Lund AS (2018) An experiment on the security of the norwegian electronic voting protocol. Annals of Telecommunications:1–9. doi:10.1007/s12243-016-0509-8

[3.] Rashid Hafeez Khokhar, Md Asri Ngadi& Satria Mandala," A Review of Current Routing Attacks in Mobile Ad Hoc Networks", International Journal of Computer Science and Security, vol (2) issue 3.

[4.] Venkata Naga Rani B, Akshay S, Arun kumar M , Ishwar Kumar M A , (2019) , Decentralized E-Voting System,International Research Journal of Engineering and Technology

[5.] Vaibhav Anasune, Pradeep Choudhari , Madhura Kelapure and PranaliShirke Prasad Halgaonkar,"Online Voting: Voting System Using B-chain",(2020), Online Voting: Voting System Using Blockchain

[6.] Sravani C., Murali G. Secure electronic voting using blockchainand homomorphic encryption. International Journal of Recent Technology and Engineering (IJRTE), vol. 8, 2019, p. 1002-1007

[7.] Shaheen S. H., Yousaf M., Jalil M. Tamper proof data distribution for universal verifiability and accuracy in electoral process using blockchain. 13th International Conference on Emerging Technologies (ICET)

[8.] Hardwick, Freya Sheer, et al. "E-voting with blockchain: An e-votingprotocol with decentralisation and voter privacy." 2018 IEEE International Conference onInternet of Things (iThings) and IEEE Green Computing and Communications (GreenCom) andIEEE Cyber, Physical and Social Computing (CPSCom) and IEEE Smart Data (SmartData). IEEE, 2018.

[9.] Alharby, Maher, and Aad van Moorsel. "Blockchain Based Smart Contracts : A Systematic Mapping Study." Computer Science & Information Technology (CS & IT)